Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society

Wayne D. Gray and Christian Schunn
Editors

August 7-10, 2002
George Mason University
Fairfax, Virginia
USA

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
18 DEC 2003	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
DARPA Support for 24th Annual Conference on the Cognitive Science Society (COG SCI 2002)	**N00014-02-1-0805**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Melinda N. Barnhart, Office of Sponsored Programs, George Mason University, Fairfax, VA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**1094**	

Copyright © 2002 by the Cognitive Science Society

All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or by any other means, without the prior written permission of the publisher.

Distributed by
Lawrence Erlbaum Associates, Inc.
10 Industrial Avenue
Mahwah, New Jersey 07430

ISBN 0-8058-4581-X
ISBN 0-8058-4583-6 (CD-ROM)
ISSN 1047-1316

Printed in the United States of America

FOREWORD

This Proceedings documents the talks, posters, tutorials, and symposia presented at the 24th Annual Meeting of the Cognitive Science Society. The meeting took place at George Mason University in Fairfax Virginia, USA from August 7 through August 10, 2002. Hundreds of submissions were received from around the world. Following last year's first European conference, it seems as if the annual meeting has become a truly international event.

The theme of this year's conference was Applied Cognition; that is, cognitive science that is either inspired by or applied to real world problems. This theme was reinforced by our plenary speakers. Stuart Card (of PARC) reminded us that successful applications of other sciences did not take a direct route from the science to application, but required the development of engineering disciplines that entailed their own research, paradigms, and formalisms. David Woods (of The Ohio State University) proposed an outline for the form that one such cognitive science-inspired engineering discipline might take. As co-Chairs, we limited our direct manipulation of the Program to our selection of these two plenary speakers. However, we are pleased to note that a large number of symposia, talks, and posters seem to have been inspired by or contributed to applied issues.

This year saw the introduction of a new submission category: publication-based submissions. Publication-based submissions allow established researchers to present talks in their area of expertise without submitting full 6-page papers. Instead, the researcher submits proof of a publication record in a given area along with a 500-word abstract of the work they would like to present at the conference. A one-page camera-ready abstract, similar to the member abstracts, is submitted later and is included in the proceedings. The goal of this new submission category is to encourage greater participation in the conference by senior cognitive scientists. Judging by the success of this category at CogSci2002, we expect it to grow over the coming years.

Like the Chairs who have gone before us, we too felt almost overwhelmed by the task of selecting and organizing our multitude of papers and symposium into coherent and non-conflicting sessions. Our task was made both easier and harder this year by the use of Simon™, conference management software commissioned by the Cognitive Science Society. Future Chairs will greatly benefit from our alpha testing of this potentially marvelous software.

There are many people to whom we owe thanks and we hope that we have listed each of their names on the pages that follow this Foreword. We would like to provide our special thanks to the following:

The Governing Board of the Cognitive Science Society for inviting us to host the meeting.

The Program Committee for acting in the capacity of Associate Editors in managing the review process.

The more than 200 reviewers for providing professional reviews and, in most cases, copious comments.

Frank Ritter and Chris Kello for organizing and coordinating the tutorial program.

Mike Byrne and Kevin Gluck for organizing the student volunteers.

Stellan Ohlsson for chairing the Marr Prize committee.

Debbie Kranz for being ready, willing, and able to coordinate the local arrangements.

Art Markman and Deborah Gruber for introducing a central Cognitive Science Society presence to the conference.

Financial support: Air Force Office of Scientific Research, DARPA, Air Force Research Labs, CHI Systems, Aptima and The Robert J. Glushko and Pamela Samuelson Foundation. George Mason University donated the use of the facilities and backed the conference with substantial financial contributions from the College of Arts & Sciences, School of Education, School of Information Technology & Engineering, Provost Office, and Psychology Department.

The plenary speakers: Stu Card and David Woods.

And our authors, symposium participants, and attendees for making the conference a true intellectual feast.

Wayne D. Gray & Christian Schunn
Conference Chairs, CogSci2002

Conference Co-Chairs

Wayne D. Gray, *George Mason University*

Christian Schunn, *University of Pittsburgh*

Conference Program Committee

Richard Alterman
Erik Altmann
Giorgio Ascoli
Larry Barsalou
William Bechtel
Andy Brook
Peter Cheng
Ron Chong
Axel Cleeremans
Rick Cooper
Gary Dell
Erik Dietrich
Dietrich Doerner
Shimon Edelman
Susan Epstein
Gilles Fauconnier
Ken Forbus
Robert French
Dedre Gentner
Gerd Gigerenzer
Adele Goldberg
Rob Goldstone
Art Graesser
Andrew Howes
Bonnie John
Ken Koedinger
Art Markman
Craig McKenzie
Doug Medin
Riichiro Mizoguchi
Nancy Nersessian
Mike Oaksford
Stellan Ohlsson
Randal O'Reilly
Daniel Osherson
Stephen Payne
David Plaut
Stephen Reed
James Reggia
Lance Rips
Brian Ross
Alan Schultz
Steve Sloman
Keith Stenning
Ron Sun
Niels Taatgen
Jim Tanaka
Josh Tenenbaum
Paul Thagard
Greg Trafton
Shimon Ulmann
Dieter Wallach
Jeremy Wolfe
Richard Young
Wayne Zachary

Local arrangements: Debbie Kranz
Submission coordinator: Frank Ritter
Conference software maintenance: Anthony Harrison
Program coordinators: Wayne D. Gray & Christian Schunn
Registration website: Arthur Markman
Website maintenance: Wayne D. Gray and Arthur Markman
Proceedings: Deborah Gruber

Cognitive Science Society
Governing Board

Susan L. Epstein (Chair) *Computer Science, Hunter College, City Univ. of New York*1997-2001

Lawrence W. Barsalou (Past Chair) *Department of Psychology, Emory University* 1997-2002

Keith Stenning (Chair-Elect) *Centre for Human Communication & Informatics, Edinburgh University*2000-2005

Arthur B. Markman (Executive Officer) *Department of Psychology, University of Texas,* 2001-2003

William Bechtel *Department of Philosophy, Washington University* 2002-2007

Kenneth D. Forbus *Department of Computer Science, Northwestern University* 1998-2003

Dedre Gentner *Department of Psychology, Northwestern University* 1999-2004

Robert L. Goldstone (Journal Editor) *Department of Psychology, Indiana University* 2001-2005

Edwin Hutchins *Department of Cognitive Science, University of California, San Diego* 2001-2006

Alan Lesgold *School of Education, Psychology, & Intelligent Systems,Univ. of Pittsburgh*1998-2003

James L. McClelland *Center for the Neural Basis of Cognition, Carnagie Mellon University* Ex-Officio member of board as chair of the Rumelhart Prize Committee

Douglas L. Medin *Department of Psychology, Northwestern University* 1999-2004

Johanna Moore *Human Communication Research Centre Edinburgh University* 2002-2007

Michael Mozer *Computer Science Department, University of Colorado* 2000-2005

Nancy Nersessian *Professor of Cognitive Sciences, Georgia Institute of Technology* 2001-2006

Paul Thagard *Philosophy Department, University of Waterloo* 1997-2002

Richard M. Young *Psychology Department, University of Hertfordshire* 2002-2007

CogSci2002 Sponsors

Air Force Office of Scientific Research

Air Force Research Laboratory/Human Effectiveness Directorate

Aptima

CHI Systems

GMU College of Arts and Sciences

GMU Psychology Department

GMU Human Factors and Applied Cognition

GMU School of Education

GMU School of Information Technology and Engineering

The Robert J. Glushko and Pamela Samuelson Foundation

Folio sponsored by Nature Publishing Group

About the Society

The Society is a non-profit professional organization, and its main activities are sponsoring an annual conference, publishing the journal Cognitive Science, and promoting research interactions across traditional disciplinary boundaries. The Society was incorporated as a non-profit professional organization in Massachusetts in 1979.

The Cognitive Science Society, Inc. brings together researchers from many fields that hold a common goal: understanding the nature of the human mind. The Society promotes scientific interchange among researchers in disciplines comprising the field of Cognitive Science, including Artificial Intelligence, Linguistics, Anthropology, Psychology, Neuroscience, Philosophy, and Education.

Contact the Society at:
Cognitive Science Society, Inc.
University of Texas at Austin
Department of Psychology
Seay Psychology Building
108 E. Dean Keeton
SPB 4.212 Mail Stop A8000
Austin, TX 78712
(512) 471-2030

CogSci2002 Reviewers

Barbara Abbott
Ray Adams
Martha Alibali
Richard Anderson
Janet Andrews
Mark Andrews
Razali Arof
Robert Atkinson
Neville Austin
Roger Azevedo
Gordon Baxter
Christopher Bearman
Safia Belkada
Sieghard Beller
G. Bengu
Bettina Berendt
Benjamin Bergen
Stefano Bertolo
John Best
Jennifer Blessing
Sergey Blok
Guido Boella
Sally Bogacz
Ronald Boring
Lera Boroditsky
Heather Bortfeld
Patrick Bouge
Gary Bradshaw
Sarah Brem
Elke Brenstein
Paul Brna
Jay Brown
Raluca Budiu
John Bullinaria
Curt Burgess
Russell Burnett
Bruce Burns
Valerie Buron
Jerome Busemeyer
Michael Byrne
Paul Cairns
Christopher Campbell
Thomas Capo
Rich Carlson
William Casebeer
Richard Catrambone
Anxo Cereijo Roibas
Craig Chambers
Tony Chan
Sanjay Chandrasekharan
Suzanne Charman
Jarinee Chattratichart
Sherry Chen
Joan Chiao
Kwangsu Cho
Yoonsuck Choe
Yu-Ju Chou
Eric Chown
Morten Christiansen
James Chumbley
Caterina Cinel
Catherine Clement
Ross Clement
Charles Clifton
Eliana Colunga
Louise Connell
Fintan Costello
Seana Coulson
Amy Criss
Jennifer Cromley
Fred Cummins
Mary Czerwinski
Hugh David
Jim Davies
Fabio Del Missier
Rutvik Desai
Arnaud Destrebecqz
Mona Diab
Tony Dickinson
Stephanie Doane
Leandro dos Santos Coelho
Hakan Duman
Michael Dyer
Maximilian Eibl
Chris Eliasmith
Michelle Ellefson
Randi A. Engle
Zachary Estes
Alberto Faro
Aidan Feeney
Ronald Ferguson
Antonio Fernandez-Caballero
Leo Ferres
Armin Fiedler
Marci Flanery
Piers Fleming
Nick Flor
Reva Freedman
Eric Freedman
Daniel Freudenthal
Wai-Tat Fu
Danilo Fum
Robert Futrelle
Michael Gasser
Merideth Gattis
Silvia Gennari
Peter Gerjets
Lisa Gershkoff-Stowe
Kevin Gluck
Susan Goldin-Meadow
Evan Golub
Judith Good
Andrew ordon
Simon Grant
Harold Greene
Jacqueline Griego
Tom Griffiths
Christopher Grindrod
Martin Groen
Stephanie Guerlain
Glenn Gunzelmann
Gianchand Gupta
Graeme Halford
Brooke Hallowell
Beverly Harrison
Lisa Haverty
Shyamanta M Hazarika
Eduard Hoenkamp
David Holliway
Kasper Hornbæk
William Horton
Eric Horvitz
Harry Howard
Roland Hubscher
Eva Hudlicka
Edward Husband
Kiyoto Ishimaru
Christian Jarrett
Todd Johnson
Randolph Jones
Dan Joyce
Marie-Odile Junker
Mark Keane
Frank Keller
Christopher Kello
Babak Khazaei
Jihie Kim
Thomas King
Nicola Knight
Chris Koch
Konrad Koerding
Janet Kolodner
Josef Krems
Amy Kruse
Maria Kutar
Abdel Labbi
Christophe Labiouse
Yannick Lallement
Steven Landry
Christian Lebiere
Michael Lee
Mark Lee
Ping Li
Hualou Liang
Alexandre Linhares

Hsi-wen Liu
Ken Livingston
Max Louwerse
Helen Lowe
Will Lowe
Christopher Lueg
Johan Lundin
Elizabeth Lynch
Dermot Lynott
Paul Maglio
Lorenzo Magnani
James Magnuson
Asifa Majid
Benise Mak
Thomas Mandl
Ken Manktelow
Denis Mareschal
Dragos Margineantu
Pinango Maria
Ioana Marian
Aleix Martinez
Amy Masnick
Rui Mata
Michael Matessa
Marshall Mayberry
Andre Mayers
Jean-Bosco Mbede
Mark McGregor
David Medler
Lisa Meeden
Lise Menn
Iza Mikeleiz
Jeanne Milostan
Naomi Miyake
Padraic Monaghan
Joyce Moore
Kenneth Moorman
Bradley Morris
Tim Morris
Julie Morrison
Robert Morrison
Doug Morse
Anandi Nagarajan
Daniel Neagu
Josef Nerb
Sourabh Niyogi
David Noelle
Kent Norman
Diarmuid O'Donoghue
Eva Olsson
Claire O'Malley
Daniel Oppenheimer
Tom Ormerod
Padraig O'Seaghdha
Magda Osman
Pierre-yves Oudeyer
Vasile Palade
Christo Panchev
Anna Papafragou
Eros Pasero
Theodore Pasquale
Philip Pavlik
Neal Pearlmutter
David Peebles
Dvora Peretz
Pierre Perruchet
Alexander Petrov
Manoj Kumar Chowdary Ponugubati
Robert Port
Matthew Posey
Athanassios Protopapas
Paul Quinn
Athanassios Raftopoulos
Michael Ramscar
Eric Raufaste
Rosamelia Ribeiro
Dale Richards
Lynn Richards
Robert Rist
Frank Ritter
David Roberts
Etienne Roesch
Laurence Rognin
Douglas Rohde
Walid Saba
William Sakas
Dario Salvucci
Lelyn Saner
Brian Scassellati
Harald Schaub
Franz Schmalhofer
Ute Schmid
Michael Schoelles
Lael Schooler
Wolfgang Schoppek
Kathy Schuh
Sam Scott
Elizabeth Sheldon
Thomas Shultz
Marin Simina
Simeon Simoff
Tom Simpson
Vladimir Sloutsky
Jesse Snedeker
Myeong-Ho Sohn
Jacques Sougne
Jon Star
Suzanne Stevenson
Anna Strasser
Fay Sudweeks
Laura Symonds
Dimitri Tabary
Federico Tajariol
Roman Taraban
Roger Taylor
Virginia Teller
Atsushi Terao
Wennekers Thomas
Sabine Timpf
Peter Torma
Joe Toth
David Townsend
Susan Trickett
Lara Triona
Peter Turney
Ryan Tweney
Aimilia Tzanavari
Alexander van den Bosch
Hans van den Broek
Ian van der Linde
Leon van der Torre
Ludger van Elst
Hedderik van Rijn
Dirk Van Rooy
Maarten van Someren
Shravan Vasishth
Alfred Vella
Andre Vellino
Alonso Vera
Rineke Verbrugge
João Veríssimo
Pirashanthie Vivekananda-Schmidt
Horatiu Voicu
Kyle Wagner
Michael R. Waldmann
Hongbin Wang
Stephan Weibelzahl
Robert West
Katja Wiemer-Hastings
Peter Wiemer-Hastings
Peter Wild
Andy Wills
Gerry Wolff
Phillip Wolff
Andree Woodcock
Gitta Wörtwein
Judith Wylie
Fei Xu
Aaron Yarlas
Daniel Yarlett
Michael C. W. Yip
Samar Zebian
Jiajie Zhang
Tom Ziemke
Corinne Zimmerman
Willem Zuidema

Tutorials

August 7, 2001

Multiple Perspectives on Consciousness for Cognitive Science 2
Richard A. Carlson (Penn State University}

APEX/CPM-GOMS: Modeling Human Performance in Applied HCI Domains 3
Roger Remington (NASA Ames Research Center) and Bonnie John (Carnegie Mellon University) and Michael Matessa, Alonso Vera, Michael Freed (NASA Ames Research Center)

How to Build Intelligent Interactive Agents Using Soar.....4
Randolph M. Jones (Colby College/Soar Technology, Inc.) and Robert E. Wray, III, Amy E. Henninger, Scott Wood (Soar Technology, Inc.) and Ronald S. Chong (George Mason University)

ACT-R 5
Christian Lebiere (Carnegie-Mellon University)

Functional Imaging of the Brain -- Developing a Synergy of Cognitive Neuroscience Behavior and Modeling 6
Walter Schneider (University of Pittsburgh)

A Cognitive Approach to Designing Human Error Tolerant Interfaces 7
Scott D. Wood (Soar Technology, Inc.) and Michael Byrne (Rice University)

Tutorial Co-Chairs

Frank E. Ritter (Penn State)
Chris Kello (George Mason University)
Local arrangements Chair: Chris Kello (George Mason University)

Tutorial Committee Members

Randolph M. Jones (Colby College and Soar Technology)
Todd Johnson (University of Texas/Houston)
Kevin Korb (Monash, Aus)
Michail Lagoudakis (Duke)
Josef Nerb (Freiburg)
Gary Jones (Derby)
Padraic Monaghan (Edinburgh)
Richard Young (Hertfordshire)

Contents

Rumelhart Prize Talk

Bayesian Modeling of Memory and Perception 9
Richard M. Shiffrin (Indiana University)

Rumelhart Symposium

Rumelhart Symposium: Honoring Richard Shiffrin 11
Susan Dumais (Microsoft Corporation)
Wilson S. Geisler (University of Texas)
Jeroen Raaijmakers (University of Amsterdam)
Mark Steyvers (University of California)

Plenary

Cognitive Science as the Engine of Innovation:
Beyond Human-Computer Interaction 13
Stuart Card (Information Sciences and Technologies Laboratory, Xerox PARC)

Steering the Reverberations of Technology Change on Fields of Practice:
Laws that Govern Cognitive Work 14
David D. Woods (Institute for Ergonomics, The Ohio State University)

Symposium

The Cognition of Complex Visualizations 18
J. Gregory Trafton (Naval Research Laboratory)
Priti Shah (University of Michigan)
Eric G. Freedman (University of Michigan)
Susan Kirschenbaum (Naval Undersea Warfare Center)
Peter C-H.Cheng (University of Nottingham)
Discussant: Mary Hegart (University of California, Santa Barbara)

Nature's Turing Test 20
Organizer: Thomas R. Zentall (University of Kentucky)
Participants:
Perceptual Classes: Edward A. Wasserman (The University of Iowa)
Superordinate Classes: Thomas R. Zentall (University of Kentuck)y
Relational Classes: Roger K. R. Thompson, Mary Jo Rattermann, and Anthony P. Chemero (Franklin & Marshall College)

The AMBR Model Comparison Project: Round III —
Modeling Category Learning 21
Organizers: Kevin A. Gluck (Air Force Research Laboratory) and Richard W. Pew (BBN Technologies)
Participants:
Experiment Design and Comparison of Human and Model Data: David Diller and Yvette Tenney (BBN Technologies)
An EPIC-Soar Model of Concurrent Performance on a Category Learning and a Simplified ATC Task: Ron S. Chong (George Mason University) and Robert E. Wray (Soar Technology,Inc.)
Developing Concept Learning Capabilities in the COGNET/iGEN Integrative Architecture and Associated AMBR ATC Model: Wayne Zachary (CHI Systems,Inc.)
An Activation-based Theory of Categorization: Christian Lebiere (Carnegie Mellon University)
Concept Learning:Knowing and Reasoning in the DCOG Architecture: Robert G.Eggleston, Air Force Research Laboratory and Katherine L.McCreight, N-Space Analysis
Symposium Discussant: Bradley C. Love (University of Texas)

Inquiry, Technology, and Cognition: Theory and Practice 23
Organizer: Sarah K. Brem (Arizona State University)
Participants:
Technical and social supports for epistemic practices of scientific argumentation: William Sandoval, Kelli Millwood (UCLA) and Marie Bienkowski, Valerie Crawford (SRI International)
Promoting critical inquiry from Web sources: Jennifer Wiley, Susan R. Goldman (UIC) and Arthur C. Graesser (University of Memphis)
Tools for representational guidance during classroom scientific inquiry: Eva E. Toth (Allegheny-Singer Research Institute)
Alternate forms of inquiry and their implications for theory and practice: Sarah K. Brem (Arizona State University)

New Models of Connectionist Language Acquisition.....24
Organizers: Ping Li (University of Richmond) and Brian MacWhinney (Carnegie Mellon University)
Participants:
Going beyond the input: the problem of generalization from sparse data: Jeff Elman
The origin of categorical representation of language in the brain: Ping Li, Igor Farkas and Brian MacWhinney
Acquisition of crisp and fuzzy concepts: Thomas Shultz

Publication-based Talks

Coordination of Talk & Action 26
Richard Alterman, Alex Feinman, Seth Landsman and Josh Introne (Brandeis University)

Developing and Validating Cockpit Interventions based on Cognitive Modeling 27
Deborah A. Boehm-Davis, Robert W. Holt, Melanie Diez and Jeffrey T. Hansberger (George Mason University)

The Information-Processing Function of Conscious Intentions 28
Richard A. Carlson, Lisa M. Stevenson, Marios N. Avraamides and Daniel N. Cassenti (Penn State University)

Activity Awareness in Computer-supported Collaborations 29
John M. Carroll (Virginia Tech)

Testing the Roles of Design History and Affordances in the HIPE Theory of Function 30
Sergio E. Chaigneau (Universidad de Tarapaca) and Lawrence W. Barsalou (Emory University)

Misrepresenting Emergent Causal Processes as Non-Emergent: A Potential Schema for Overcoming Misunderstandings in Science 31
Michelene T. H. Chi (University of Pittsburgh)

Protocol Evidence On Thought Experiments Used By Experts 32
John J. Clement (University of Massachusetts)

Putting Geometry and Function Together — Towards a Psychologically-Plausible Computational Model for Spatial Language Comprehension 33
Kenny R. Coventry, Angelo Cangelosi, Dan Joyce and Lynn V. Richards (University of Plymouth)

A Basis for a Rigorous Cognitive Science: Maintaining Context for Information Exchange between Modules in a Functional Hierarchy 34
L. Andrew Coward (Murdoch University)

Dynamic Interrelations Among Processing Efficiency, Working Memory, and Problem Solving: A Longitudinal Study 35
Andreas Demetriou (University of Cyprus)

Tutoring Real-Time Dynamic Task Performance: Using ADAPT to Augment Pilot Skill Acquisition 36
Stephanie M. Doane and Daniel W. Carruth (Mississippi State University)

Implementing Latent Semantic Analysis in Learning Environments with
Conversational Agents and Tutorial Dialog 37
Arthur C. Graesser, Xiangen Hu, Brent A. Olde, Matthew Ventura, Andrew Olney, Max Louwerse and Donald R. Franceschetti (University of Memphis) and Natalie Person (Rhodes College)

Human-Automation Interaction Strategies 38
Stephanie Guerlain (University of Virginia)

Statistical learning, implicit memory, and phonology 39
Prahlad Gupta and John Lipinski (University of Iowa)

Mental Visualizations and External Visualizations 40
Mary Hegarty (University of California)

Modeling aviation crew interaction using a cognitive architecture 41
Robert W. Holt, Jeffrey T. Hansberger, Ronald S. Chong and Deborah A. Boehm-Davis (George Mason University)

Promoting Transfer through Case-Based Reasoning: Rituals and Practices
in the Learning by Design Classroom and Evidence of Transfer 42
Janet L. Kolodner (Georgia Institute of Technology)

Dynamic Adaptation to Critical Care Medical Environment:
Error Recovery as Cognitive Activity 43
Tate T. Kubose, Vimla L. Patel and Desmond Jordan (Columbia University)

Applications of Latent Semantic Analysis 44
Thomas K Landauer (University of Colorado at Boulder)

Modeling the Development of Lexicon with DevLex:
A Self-Organizing Neural Network Model of Lexical Acquisition 45
Ping Li and Igor Farkaš (University of Richmond)

Where Do Problem-Solving Strategies Come From? 46
Marsha C. Lovett (Carnegie Mellon University)

Is There a Decision Bias For Information From Internally Consistent Sources? 47
Shenghua Luan, Robert D. Sorkin and Jesse Itzkowitz (University of Florida)

Understanding and Scaffolding Constructive Collaboration 48
Naomi Miyake and Hajime Shirouzu (Chukyo University)

Learning fromWorked-Out Examples via Self-Explanations:
How it Can(not) be Fostered 49
Alexander Renkl (University of Freiburg)

Category Use: Learning and Understanding Categories 50
Brian H. Ross and Seth Chin-Parker (University of Illinois)

Relating Properties of Human Memory to Cortico-Hippocampal Architecture.....51
Lokendra Shastri (International Computer Science Institute)

What Happened to the Imagery Debate?.....52
Peter P. Slezak (University of New South Wales)

On the Origins of Perceived Sameness in Shape.....53
Linda B. Smith (Indiana University)

The Origins, Development, and Nature of Argument Understanding.....54
Nancy L Stein (University of Chicago) and Elizabeth R. Albro (Wheaton College)

Constructive Perception: An Expertise to Use Diagrams for Dynamic Interactivity.....55
Masaki Suwa (Chukyo University)

Literary Cognition and Aesthetic Computing 56
Akifumi Tokosumi (Tokyo Institute of Technology) and Norikazu Yoshimine (Shonan Kokusai Women's College)

Diagrams to Augment Cognition.....57
Barbara Tversky, Julie Heiser and Paul Lee (Stanford University) and Jeffrey M. Zacks (Washington University)

Papers

Representation Strength Influences Strategy Use and Strategy Discovery 59
Martha W. Alibali, Tara L. Booth (University of Wisconsin-Madison)

Integrating Decay and Interference: A New Look at an Old Interaction 65
Erik M. Altmann (Michigan State University) and
Christian D. Schunn (University of Pittsburgh)

Preventing Catastrophic Interference in Multiple-Sequence Learning
Using Coupled Reverberating Elman Networks 71
Bernard Ans and Stéphane Rousset (Université Pierre Mendès-France) and
Robert M. French (Université de Liège) and Serban Musca (Université Pierre Mendès-France)

A Cognitive Account of Situated Communication 77
Rita B.Ardito, Bruno G. Bara and Enrico Blanzieri (Universita' di Torino)

Ah-Ha, I Knew It All Along: Differences in Hindsight Bias Between Insight
and Algebra Problems 83
Ivan K.Ash and Jennifer Wiley (The University of Illinois at Chicago)

A Neurocognitive Model for Students and Educators 89
Michael Atherton (Department of Educational Psychology)

Do people update spatial relations described in texts? 95
Marios N. Avraamides (The Pennsylvania State University)

An Exploration of Real-World Analogical Problem Solving in Novices 101
Christopher R. Bearman, Linden J. Ball and
Thomas C. Ormerod (Lancaster University)

Neonatal Learning of Faces: Environmental and Genetic Influences 107
James A. Bednar and Risto Miikkulainen (The University of Texas at Austin)

Conditional Promises and Threats – Cognition and Emotion 113
Sieghard Beller (University of Freiburg)

Combining Simplicity and Likelihood in Language and Music 119
Rens Bod (University of Amsterdam)

Mental Models Theory and Anaphora 125
Guido Boella and Leonardo Lesmo, (Universit`a di Torino)

Comparison and the development of knowledge 131
Lera Boroditsky (MIT)

What is universal in event perception? Comparing English & Indonesian speakers 136
Lera Boroditsky and Wendy Ham (MIT) and
Michael Ramscar (University of Edinburgh)

Atomistic and Systems Approaches to Consciousness....142
Andrew Brook and Luke Jerzykiewicz (Carleton University)

Reference Resolution in the Wild: On-line circumscription of referential domains in a natural, interactive problem-solving task....148
Sarah Brown-Schmidt, Ellen Campana and Michael K. Tanenhaus (University of Rochester)

On Straight TRACS: A baseline bias from mental models....154
Kevin Burns (The MITRE Corporation)

Contradictions and Counterfactuals: Generating Belief Revisions in Conditional Inference....160
Ruth M.J. Byrne and Clare R. Walsh (University of Dublin, Trinity College)

Anthropomorphic Agents as a User Interface Paradigm: Experimental Findings and a Framework for Research....166
Richard Catrambone, John Stasko and Jun Xiao (Georgia Institute of Technology)

The Effect of Goal Constraints on Strategy Generation....172
Suzanne C. Charman and Andrew Howes (Cardiff University)

Diagnosticity in Category Learning by Classification and Inference....178
Seth Chin-Parker and Brian H. Ross (University of Illinois)

Comprehension Monitoring and Regulation in Distance Collaboration....184
Kwangsu Cho, Christian D. Schunn and Alan M. Lesgold (University of Pittsburgh)

Second Order Isomorphism: A Reinterpretation and Its Implications in Brain and Cognitive Sciences....190
Yoonsuck Choe (Texas A&M University)

Age Differences in Transitory Cognitive Performance....196
Sy Miin Chow and John R. Nesselroade (University of Virginia)

Reminiscence and Arousal: A Connectionist Model....202
Eric Chown (Bowdoin College)

How Conceptual Metaphors are Productive of Spatial-Graphical Expressions....208
Timothy C. Clausner (HRL Laboratories, LLC)

What makes a word?....214
Eliana Colunga and Linda B. Smith (Indiana University)

Sequential Learning by Touch, Vision, and Audition....220
Christopher M. Conway and Morten H. Christiansen (Cornell University)

Feedback Effects in the Acquisition of a Hierarchical Skill....226
Andrew Corrigan-Halpern and Stellan Ohlsson (UIC)

Investigating creative language: People's choice of words in the production of novel noun-noun compounds 232
Fintan Costello (Dublin City University)

Do Expression and Identity Need Separate Representations? 238
Garrison W. Cottrell, Kristin M. Branson (USCD) and Andrew J. Calder (MRC Cognition and Brain Sciences Unit)

Cognitive Precursors to Science Comprehension 244
Kimberly G. Cottrell and Danielle S. McNamara (Old Dominion University)

The Role of Diagrams and Diagrammatic Affordances in Analogy 250
David Latch Craig, Nancy J. Nersessian and Richard Catrambone (Georgia Institute of Technology)

A Classifcation of Cognitive Agents 256
Mehdi Dastani (Institute of Information and Computer Sciences) and Leendert van der Torre (Vrije Universiteit Amsterdam)

Declarative and Procedural Strategies in Problem Solving: Evidence from the Toads and Frogs Puzzle 262
Fabio Del Missier and Danilo Fum (Trieste)

Teaching with Dialectic Arguments vs. Didactic Explanations 268
Ravi Desai and Kevin D. Ashley (University of Pittsburgh)

Modeling Human Error in a Real-World Teamwork Environment 274
Stephen Deutsch and Richard Pew (BBN Technologies)

The Quality of Test Context and Contra-evidence as a Moderating Factor in the Belief Revision Process 280
Kristien Dieussaert, Walter Schaeken and Gery d'ydewalle (University of Leuven)

The Role of Analogy in Teaching Middle-School Mathematics 286
Lindsey K. Engle, Keith J. Holyoak and James W. Stigler (University of California)

Category Size and Category-Based Induction 292
Aidan Feeney and David R. Gardiner, (University of Durham)

Why Example Fading Works: A Qualitative Analysis Using Cascade 298
Eric S. Fleischman and Randolph M. Jones (Colby College)

Evolution of Gender in Indo-European Languages 304
Harry E. Foundalis (Indiana University)

Recovering Context After Interruption 310
Jerry L. Franke, Jody J. Daniels and Daniel C. McFarlane (Lockheed Martin Advanced Technology Laboratories)

Four Problems with Extracting Human Semantics from Large Text Corpora 316
Robert M. French and Christophe Labiouse (University of Liege)

The Importance of Starting Blurry: Simulating Improved Basic-Level Category Learning in Infants Due to Weak Visual Acuity 322
Robert M. French and Martial Mermillod, (University of Liège) and Alan Chauvin (University of Grenoble) and Paul C. Quinn (Washington & Jefferson College) and Denis Mareschal (Birkbeck College)

Modelling the Development of Dutch Optional Infinitives in MOSAIC 328
Daniel Freudenthal, Julian Pine and Fernand Gobet (University of Nottingham)

Subject Omission in Children's Language: The Case for Performance Limitations in Learning 334
Daniel Freudenthal, Julian Pine and Fernand Gobet (University of Nottingham)

Does Positivity Bias Explain Patterns of Performance on Wason's 2-4-6 Task? 340
Maggie Gale (University of Derby) and Linden J. Ball (Lancaster University)

A Connectionist model of Planning via Back-chaining Search....345
Max Garagnani, (The Open University) and Lokendra Shastri and Carter Wendelken (The International Computer Science Institute)

Events versus States: Empirical Correlates of Lexical Classes 351
Silvia Gennari and David Poeppel (University of Maryland)

Interactive Knowledge Acquisition Tools: A Tutoring Perspective....357
Yolanda Gil and Jihie Kim (University of Southern California)

Taking Care of the Linguistic Features of Extraversion....363
Alastair J. Gill and Jon Oberlander (University of Edinburgh)
The Role of Roles in Translating Across Conceptual Systems 369
Robert L. Goldstone and Brian J. Rogosky (Indiana University)

The Theory of Mind in Strategy Representations 375
Andrew S. Gordon (University of Southern California)

A probabilistic approach to semantic representation....381
Thomas L.Griffiths & Mark Steyvers (Stanford University)

Strategic Differences in the Coordination of Different Views of Space 387
Glenn Gunzelmann and John R. Anderson (Carnegie Mellon University)

Understanding Similarity in Choice Behavior: A Connectionist Model 393
Frank Y. Guo and Keith J. Holyoak (UCLA)

Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions....399
Todd M. Gureckis and Bradley C. Love (The University of Texas at Austin)

A Constraint Satisfaction Model of Causal Learning and Reasoning 405
York Hagmayer and Michael R. Waldmann (University of Göttingen)

How Similarity Affects the Ease of Rule Application 411
Ulrike Hahn (Cardiff University) and Mercè Prat-Sala (King Alfred's College) and Emmanuel M. Pothos (University of Edinburgh)

Modeling Grouping with Recursive Auto-Associative Memory 417
Andreas Hansson and Lars F. Niklasson (University of Skövde)

Holographic Reduced Representations for Oscillator Recall: A Model of Phonological Production 423
Harlan D. Harris (University of Illinois at Urbana-Champaign)

Similarity and Difference Judgments Under Perceptual and Non-Perceptual Conditions 429
Uri Hasson (Princeton University) and Vladimir Sloutsky (The Ohio State University)

The /s/ morpheme and the compounding phenomenon in English.....435
Jenny Hayes, Victoria Murphy, Neil Davey, Pamela Smith and Lorna Peters (University of Hertfordshire)

Interactional Context in Graphical Communication.....441
Patrick G. T. Healey, (Queen Mary University of London) and Simon Garrod, Nicholas Fay (University of Glasgow) and John Lee, Jon Oberlander (University of Edinburgh)

Diagrams and Descriptions in Acquiring Complex Systems 447
Julie Heiser and Barbara Tversky (Stanford University)

Do argumentation tasks promote conceptual change about volcanoes?.....453
Joshua A Hemmerich and Jennifer Wiley (The University of Illinois at Chicago)

Predicting Agent Spatial Information: A Comparison Between Neural Networks and Dead Reckoning Algorithms 459
Amy E. Henninger (Soar Technology, Inc.) and Avelino J. Gonzalez (University of Central Florida) and Douglas A. Reece (SAIC)

Anatomy is Symmetry's Best Friend: Reflections on Modeling Baylis and Driver 465
Jon Hicks and Jon Oberlander (Department of Informatics, Edinburgh)

Perspective-taking in Young Writer's Descriptive Writing 471
David R. Holliway (Marshall University)

An Instance-based Model of the Effect of Previous Choices on the Control of Interactive Search.....476
Andrew Howes, Stephen J. Payne and Juliet Richardson (Cardiff University)

Modeling Capabilities and Workload in Intelligent Agents for Simulating Teamwork 482
Thomas R. Ioerger, Linli He, Deborah Lord (Texas A&M University) and Pamela Tsang (Wright State University)

Self-Organizing Recognition and Classification of Relational Structures 488
Brijnesh J. Jain and Fritz Wysotzki, (Technical University Berlin)

Integrating Perceptual Organization and Attention: A New Model
For Object-Based Attention.....494
Jerzy P. Jarmasz (Carleton University)

Children's Acceptance and Use of Unexpected Category Labels to
Draw Non-Obvious Inferences.....500
Vikram K. Jaswal and Ellen M. Markman (Stanford University)

A Model of Spatio-Temporal Coding of Memory for Multidimensional Stimuli.....506
Todd R. Johnson, Hongbin Wang, Jiajie Zhang and Yue Wang (University of Texas Health Science Center at Houston)

Analysis of the Dynamics of Reasoning Using Multiple Representations.....512
Catholijn M. Jonker (Vrije Universiteit Amsterdam) and
Jan Treur (Vrije Universiteit Amsterdam and Universiteit Utrecht)

Cue Abstraction and Exemplars in Multiple-Cue Judgment 518
Peter Juslin, Henrik Olsson and Anna-Carin Olsson (Umeå University)

Predicting Noun and Verb Latencies: Influential Variables and Task Effects.....524
Natalie Kacinik and Christine Chiarello (University of California, Riverside)

Graph Structure Supports Graph Description.....530
Irvin R. Katz (Center for New Constructs, Educational Testing Service) and
Xiaoming Xi (Univ. of California) and Hyun-Joo Kim (Columbia University) and
Peter C-H. Cheng (University of Nottingham)

Sex, Myths, and Adolescents' Conceptual Understanding of HIV 536
Alla Keselman and Vimla L. Patel (Columbia University)

A Cognitive Task Analysis of Using Pictures To Support Pre-Algebraic Reasoning.....542
Kenneth R. Koedinger and Atsushi Terao (Carnegie Mellon University)

Mutual Adaptive Meaning Acquisition by Paralanguage Information: Experimental Analysis of
Communication Establishing Process 548
Takanori Komatsu, Kentaro Suzuki, Kazuhiro Ueda and Kazuo Hiraki (The
University of Tokyo) and Natsuki Oka (Matsushita Electric Industrial Co., Ltd)

Qualitative physics as a component in natural language semantics: A progress report 554
Sven E. Kuehne and Kenneth D. Forbus (Northwestern University)

Learning Causal Structure 560
David A. Lagnado and Steven Sloman
(Department of Cognitive and Linguistic Sciences)

Data Analysis of Conceptual Similarities of Finnish verbs.....566
Krista Lagus and Mathias Creutz (Helsinki University of Technology) and
Anu Airola (University of Helsinki)

Multitasking as Skill Acquisition 572
Frank J. Lee (Rensselaer Polytechnic Institute) and
Niels A. Taatgen (University of Groningen)

Using Cognitive Decision Models to Prioritize E-mails 578
Michael D. Lee, Lama H. Chandrasena and Daniel J. Navarro (University of Adelaide)

Is Concept Formation An Age-Independent Process? 584
Kenneth R. Livingston, Janet K. Andrews and Emily Kuschner (Vassar College)

Theories and Similarity: Categorization under Speeded Conditions 590
Christian C. Luhmann, Woo-kyoung Ahn and Thomas J. Palmeri (Vanderbilt Univ.)

Case, Word Order, and Language Learnability: Insights from
Connectionist Modeling 596
Gary Lupyan and Morten H. Christiansen (Cornell University)

On Understanding Discourse in Human-Computer Interaction 602
Paul P. Maglio, Teenie Matlock, Sydney J. Gould, Dave Koons and
Christopher S. Campbell (IBM Almaden Research Center)

On the Potential of Epistemic Actions for Self-Cueing: Multiple Orientations
Can Prime 2D Shape Recognition and Use 608
Paul P. Maglio (IBM Almaden Research Center) and
Michael J. Wenger(University of Notre Dame)

Immediate Integration of Syntactic and Referential Constraints on Spoken
Word Recognition 614
James S. Magnuson (Columbia University) and
Michael K. Tanenhaus and Richard N. Aslin (University of Rochester)

Three-year-old Children's Use of Category Labels and Motion in
Drawing Inferences about Animal Kinds 620
Benise S.K. Mak and Lap Yan Lo (The University of Hong Kong) and
Alonso H. Vera (NASA Ames Research Center)

Incorporating Cognitive Styles into Adaptive Multimodal Interfaces 626
Halima Habieb Mammar and Franck Tarpin Bernard (INSA de Lyon)

Metacat: A Self-Watching Cognitive Architecture for Analogy-Making 631
James B. Marshall (Pomona College)

Where do syllables come from? 637
Evelyn Martens, Walter Daelemans, Steven Gillis and Helena Taelman
(Universitaire Instelling Antwerpen)

Reasoning from Data: The Effect of Sample Size and Variability on
Children's and Adults' Conclusions 643
Amy M. Masnick (Carnegie Mellon Univ.) and Bradley J. Morris (Univ. of Pittsburgh)

Reusable Templates in Human Performance Modeling 649
Michael Matessa, Alonso Vera, Roger Remington, and Michael Freed (NASA Ames Research Center) and Bonnie John (Carnegie Mellon University)

Collaborative Interactions: The Process of Joint Production and Individual Reuse of Novel Ideas 655
Mark U. McGregor and Michelene T.H. Chi (University of Pittsburgh)

A Strong Schema Can Interfere with Learning: The Case of Children's Typical Addition Schema.....661
Nicole M. McNeil and Martha W. Alibali (Department of Psychology)

Changes in Learners' Exploratory Behavior in a Simulated Psychology Laboratory.....667
Kazuhisa Miwa, Norio Ishii, Hitomi Saito, and Ryuichi Nakaike (Nagoya University)

Learning to Solve Complex Propositions: Does knowledge of truth-values bootstrap modal operators? 673
Bradley J. Morris (University of Pittsburgh) and David Klahr (Carnegie Mellon University)

Logical Strategery 679
Bradley J. Morris and Christian Schunn (University of Pittsburgh)

Commonalities and Distinctions in Featural Stimulus Representations.....685
Daniel J. Navarro and Michael D. Lee (University of Adelaide)

Thinking by Doing? Epistemic Actions in the Tower of Hanoi.....691
Hansjörg Neth and Stephen J. Payne (Cardiff University)

Bayesian Learning at the Syntax-Semantics Interface 697
Sourabh Niyogi (Massachusetts Institute of Technology)

Objet Trouvé, Holism, and Morphogenesis in Interactive Evolution.....703
Ron W. Noel (WCSU) and Sylvia Acchione-Noel (General Electric)

The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis? 708
Brent A. Olde, Donald R. Franceschetti and Arthur C. Graesser (University of Memphis) and Ashish Karnavat (CHI Systems, Inc)

Experience and Pseudo-Experience: Exemplar Effects Without Feedback 714
Henrik Olsson and Peter Juslin (Umeå University)

Simplicity: A cure for overgeneralizations in language acquisition?.....720
Luca Onnis, Matthew Roberts and Nick Chater (University of Warwick)

What's a Science Student to Do? 726
Tenaha O'Reilly, Danielle S. McNamara and The Strategies Lab (Old Dominion University)

Is there evidence for unconscious reasoning processes? 732
Magda Osman (University College London)

A Unified Model of the Origins of Phonemically Coded Syllable Systems 738
Pierre-yves Oudeyer (Sony Computer Science Lab)

The Pragmatics of Number.....744
Anna Papafragou (Institute for Research in Cognitive Science) and
Julien Musolino (Department of Speech and Hearing Sciences)

A Computational Theory of Complex Problem Solving
Using Latent Semantic Analysis 750
José Quesada and Walter Kintsch (University of Colorado) and
Emilio Gomez (University of Granada)

A Dynamical Connectionist Account of Conceptual Change 756
Athanassios Raftopoulos and Andreas Demetriou (University of Cyprus)

Deictic Codes, Demonstratives, and Reference: A Step Toward
Solving the Grounding Problem.....762
Athanassios Raftopoulos (University of Cyprus) and
Vincent C. Müller (American College of Thessaloniki)

When the fly flied and when the fly flew: the effects of semantics on
the comprehension of past tense inflections 768
Michael Ramscar (University of Edinburgh)

Inferring Unobserved Category Features With Causal Knowledge 774
Bob Rehder (New York University) and Russell C. Burnett (Northwestern University)

Routine Problem Solving in Groups.....780
Torsten Reimer, Klaus Opwis and Anne-Louise Bornstein (University of Basel)

Search, Structure or Statistics? A Comparative Study of Memoryless
Heuristics for Syntax Acquisition 786
William Gregory Sakas and Eiji Nishimoto (CUNY)

Modeling Driver Distraction from Cognitive Tasks.....792
Dario D. Salvucci (Drexel University)

The Impact of Problem Order: Sequencing Problems as a Strategy for Improving One's
Performance.....798
Katharina Scheiter (University of Tuebingen) and
Peter Gerjets (Knowledge Media Research Center)

Stochastic Independence between Recognition and Completion of
Spatial Patterns as a Function of Causal Interpretation.....804
Wolfgang Schoppek (University of Bayreuth)

Designing Sets of Instructional Examples to Accomplish Different Goals of Instruction.....810
Tina Schorr (Virtual Ph.D. Program: Knowledge Acquisition and Knowledge
Exchange with New Media) and Peter Gerjets (Knowledge Media Research Center) and
Katharina Scheiter (Institute of Psychology) and
Yiannis Laouris (Cyprus Neuroscience and Technology Institute)

Learning by Solved Example Problems: Instructional Explanations Reduce Self-Explanation Activity 816
Silke Schworm and Alexander Renkl (Educational Psychology)

The Psychological Implausibility of Naturalized Content 822
Sam Scott (Carleton University)

Counterfactual Undoing in Deterministic Causal Reasoning 828
Steven A. Sloman and David A. Lagnado (Brown University)

Formalizing Affordance 834
Mark Steedman (University of Edinburgh)

Providing Distinctive Cues to Augment Human Memory 840
Jeanine K. Stefanucci and Dennis R. Proffitt (Department of Psychology)

Naive Strategic Thinking 845
Eugenia Steingold (Harvard University) and P. N. Johnson-Laird (Princeton University)

Implicit Learning of Serial Reaction Time Tasks: Connectionist vs. Symbolic Models 850
Ron Sun (University of Missouri) and Chris Terry (University of Alabama)

Detecting the Local Maximum: A Satisficing Heuristic 856
Yanlong Sun and Ryan D. Tweney (Bowling Green State University)

Top-Down versus Bottom-Up Learning in Skill Acquisition 861
Ron Sun and Xi Zhang (University of Missouri)

Incremental Referential Domain Circumscription during Processing of Natural and Synthesized Speech 867
Mary D. Swift, Ellen Campana, James F. Allen and Michael K. Tanenhaus (University of Rochester)

The Role of Consciousness in Second Language Acquisition 872
Edina Torlakovic and Andrew Brook (Carleton University)

The Instantiation and Use of Conceptual Simulations in Evaluating Hypotheses: Movies-in-the-Mind in Scientific Reasoning 878
Susan B. Trickett (George Mason University) and J. Gregory Trafton (Naval Research Laboratory)

Goal Specificity and the Generality of Schema Acquisition 884
David L. Trumpower, Timothy E. Goldsmith and Maureen Below (University of New Mexico)

Precipitate Replications: The Cognitive Analysis of Michael Faraday's Exploration of Gold Precipitates and Colloids 890
Ryan D. Tweney, Ryan P. Mears, Robert E. Gibby, Christiane Spitzmüller and Yanlong Sun (Bowling Green State University)

Graphically Speaking: Do Graphics Affect Perspectives in Event Conceptualization?.....896
Ichiro Umata, Yasuhiro Katagiri (ATR Media Information Science Laboratories) and Atsushi Shimojima (Japan Advanced Institute of Science and Technology)

When participants are not misled they are not so bad after all: A pragmatic analysis of a rule discovery task 902
Jean-Baptiste Van der Henst (Institut Jean Nicod) and Sandrine Rossi (Université de Caen) and Walter Schroyens (K.U. Leuven)

Deriving a conclusion from relational premises.....908
Jean-Baptiste Van der Henst and Walter Schaeken *(University of Leuven)*

Working Memory Capacity and the Nature of Generated Counterexamples.....914
Niki Verschueren, Wim De Neys, Walter Schaeken and Géry d'Ydewalle (University of Leuven)

A Study of Object-Location Memory.....920
Hongbin Wang, Todd R. Johnson, Jiajie Zhang and Yue Wang (University of Texas Health Science Center at Houston)

Combining belief and utility in a structured connectionist agent architecture 926
Carter Wendelken and Lokendra Shastri (International Computer Science Institute)

Computer Augmented Psychophysical Scaling.....932
Robert L. West, Ronald L. Boring and Stephen Moore (Carleton University)

Adapting to a Response Deadline in Categorization 938
A. J. Wills (University of Exeter)

A Vector Model of Causal Meaning 944
Phillip Wolff (Department of Psychology) and Matthew Zettergren (Department of Electrical and Computer Engineering)

A Self-Organizing Connectionist Model of Character Acquisition in Chinese 950
Hongbing Xing (Beijing Language and Culture University) and Hua Shu (Beijing Normal University) and Ping Li (University of Richmond)

Uncertainty in Causal and Counterfactual Inference 956
Daniel Yarlett and Michael Ramscar (University of Edinburgh)

Linguistic cues enhance the learning of perceptual cues.....962
Hanako Yoshida and Linda B. Smith (Indiana University)

How are speech and gesture related?.....966
Hanako Yoshida, Linda B. Smith, Raedy M. Ping and Elizabeth L. Davis (Indiana University)

Toward An Action Based Taxonomy of Human Errors in Medicine.....970
Jiajie Zhang and Todd R. Johnson (University of Texas at Houston) and Vimla L. Patel and Edward H. Shortliffe (Columbia University)

Why do metaphors seem deeper than similes? 976
Sergey S. Zharikov and Dedre Gentner (Northwestern University)

Is Competitive Learning an Adequate Account of Free Classification? 982
Jan Zwickel (Department of Psychology) and A.J. Wills (University of Exeter)

Member Abstracts

A Formal Analysis of Intelligent Agents with Mathematical Tools.....989
Zippora Arzi-Gonczarowski, (Typographics)

Distinct Errors Arising From a Single Misconception 990
Ryan S. Baker, Albert T. Corbett and Kenneth R. Koedinger (Carnegie Mellon University)

Belief in the Hot Hand Improves Performance: A Mathematical Model.....991
Bruce D. Burns (Michigan State University)

Human reasoning: an analysis of the mathematical problem-resolution strategies.....992
Manoel Caetano and Adriana Soares, (Universidade Gama Filho)

The Role of Logical Structure and Premise Believability in Belief Revision.....993
Dustin P. Calvillo and Russell Revlin (University of California)

Displacement affects duration estimation, but not the other way around.....994
Daniel J. Casasanto and Lera Boroditsky (Massachusetts Institute of Technology)

Evaluating Information Design for Notification Systems 995
C. M. Chewar and D. Scott McCrickard
(Virginia Polytechnic Institute and State University)

The Recognition of Overlapped Chinese Characters at Two Spatial Scales 996
Yu-Ju Chou and Richard Shillcock (University of Edinburgh)

Learning the Dynamics of Vowel to Vowel Phonotactics 997
Orlando Bisacchi Coelho (UMC / FEEC & IEL – UNICAMP) and
Edson Françozo, Eleonora Albano, Laudino Roces, Pablo Arantes and
Renato Basso (LAFAPE – IEL – UNICAMP)

The Roots of Plausibility: The Role of Coherence and Distributional
Knowledge in Plausibility Judgements 998
Louise Connell and Mark T. Keane (University College Dublin)

Measures of Real Time Assessment to use in Adaptive Augmentation.....999
Martha E. Crosby, Curtis Ikehara and David N. Chin (University of Hawaii)

Semantic Memory Retrieval During Conditional Reasoning:
Every Counterexample Counts 1000
Wim De Neys, Walter Schaeken and Géry d'Ydewalle (K.U.Leuven)

Categorization of Emergent Processes by Students at Different
Levels of Expertise.....1001
Randi A. Engle and Michelene T. H. Chi (University of Pittsburgh)

The Autonoetic Hypothesis On Creativity: Memory and Cognition in Pollock's Abstract Art 1002
Carlos H. Espinel (The Blood Pressure Center and Georgetown University Medical Center)

Epistemic Belief and Semantic Categorization 1003
Zachary Estes (University of Georgia)

Learning from Transformational and Derivational Worked-out Examples 1004
Peter Gerjets and Katharina Scheiter (University of Tuebingen) and Stefan Kleinbeck (University of Freiburg) and Ute Schmid (University of Osnabrueck)

The Role of Cognitive Modeling in Enhancing Dynamic Decisions 1005
Cleotilde Gonzalez (Carnegie Mellon University)

Developing a Framework for Understanding Scientific and Technological Thinking: Notes from a Workshop 1006
Michael E. Gorman and Alexandra Kincannon (University of Virginia)

Automated Detection of Strategies in Free Text Responses 1007
Anthony Harrison, Lelyn Saner, Celestine Cookson, Darcie Kunder and Christian D. Schunn (University of Pittsburgh)

ACT-R/S: A Computational and Neurologically Inspired Model of Spatial Reasoning 1008
Anthony M. Harrison and Christian D. Schunn (University of Pittsburgh)

Belief Revision and Reasoning 1009
Uri Hasson and Philip N. Johnson-Laird (Princeton University)

Recency Effects in Category Learning are Dynamic and Adaptive 1010
Matt Jones (The University of Michigan) and Winston R. Sieck (The Ohio State University)

Cognitive Barriers to the Effective Use of a Diabetes Home Telemedicine System 1011
David R. Kaufman, Vimla L. Patel and Justin Starren (Columbia University)

The Roles of Context and Working Memory in Probability Matching 1012
Alexandra P. Kincannon (University of Virginia)

Structure of Linguistic Spatial Representation: A test for psychometric structure using Japanese spatial terms 1013
Takatsugu Kojima and Takashi Kusumi (Kyoto University)

The Effect of Attentional Distraction in the Tempo-Naming Task 1014
Laura Leach and Christopher Kello (George Mason University)

Why Animated but not Static? The Spatial-Temporal 1015
Terence C. P. Lee, Albert W. L. Chau and Benise S. K. Mak (The University of Hong Kong)

Domain Knowledge and False Memory 1016
Yuh-shiow Lee and Han-yu Lin (National Chung-Cheng University)

Acquisition of Landmark Knowledge from Static and Dynamic
Presentation of Route Maps 1017
Paul U. Lee (Stanford) and Heike Tappe, Alexander Klippel (University of Hamburg)

Bongard problems and symbolic approaches: a skeptical look 1018
Alexandre Linhares (EBAPE/FGV)

Language-Like Representation in Embodied and Situated Cognition:
A Case Study of a Situated Robot's Planning 1019
Hsi-wen Liu (Providence University)

The Comprehension of Novel Noun-Noun Compounds: The Influence of
Out-of-Context Interpretations on In-Context Understanding 1020
Dermot Lynott and Mark T. Keane (University College Dublin)

Allocation of Attention in Neural Network Models of Categorization 1021
Toshihiko Matsuka and James E. Corter (Columbia University) and
Arthur B. Markman (University of Texas- Austin)

How Goals Affect Evaluations of Animation Effectiveness 1022
Julie Bauer Morrison (Bryant College)

Cognitive Principles in a Computational Engineering Design Methodology 1023
Jarrod Moss, Kenneth Kotovsky and Jonathan Cagan (Carnegie Mellon University)

The Role of Exploration and Forward Checking in Human Scheduling 1024
Stefani Nellen and Joachim Funke (University of Heidelberg)

Cognitive Functional Processing System: Reasoning about Quantitative Relationships 1025
Kent L. Norman (University of Maryland)

Strategies and Eye-movement of an Expert in a Video Game 1026
Hidemi Ogasawara (Chukyo University) and Takehiko Ohno
(Communication Science Laboratories)

Not so Fast! (And not so Frugal): Rethinking the Recognition Heuristic 1027
Daniel M. Oppenheimer (Stanford University)

Neural Correlates of Perceptual/Semantic Encoding and
Implicit/Explicit Retrieval: An fMRI Study 1028
T. Park (Chonnam National University)

Mental Rotation Transfer 1029
Philip Pavlik and John Anderson (Carnegie Mellon University)

What do you understand for X? 1030
Célia Lúcia Gomes Pessanha and Adriana Soares
(LCC/CCH/UENF Universidade Gama Filho)

Mental representation in mathematical problem resolution 1031
Maridelma Pourbaix and Adriana Soares
(LCC/CCH/UENF Universidade Gama Filho)

Browsing Multiple Texts under Time Pressure 1032
William R. Reader and Stephen J. Payne (Cardiff University)

Color Palettes for Displays: Optimization by Genetic Algorithm 1033
John Rehling (Carnegie Mellon University)

Letter Spirit: An Architecture for Creativity 1034
John Rehling (Carnegie Mellon University)

How to Make a Computer Conscious 1035
Alexei V. Samsonovich (George Mason University)

The Role of Prior Beliefs in Processing Analogical Arguments 1036
Lelyn Saner and Christian D. Schunn (University of Pittsburgh)

A Pyramid Model of the Perception of Partially Visible Figures 1037
Michael R. Scheessele (Indiana University - South Bend) and
Zygmunt Pizlo (Purdue University)

Tomorrow's Human Computer Interaction from Vision to Reality:
Building Cognitively Aware Computational Systems 1038
LCDR Dylan Schmorrow (DARPA IPTO) and Amy A. Kruse (DARPA)

Learning by Collaborating Revisited: Individualistic vs.
Convergent Understanding 1039
Hajime Shirouzu and Naomi Miyake (Chukyo University)

Retrieval Effects on Confidence in General Knowledge 1040
Winston R. Sieck (The Ohio State University) and
J. Frank Yates (The University of Michigan)

Perception matters: Effects of perceptual richness on categorization 1041
Vladimir M. Sloutsky and Anna V. Fisher (Ohio State University)

Investigating Cognitive Gain In A Logical Experiment 1042
Adriana Soares and Cabral Lima (DCC/IM/UFRJ Universidade Gama Filho)

Children's developing ability to create external representations: Separating what
information is included from how the information is represented 1044
Lara M. Triona and David Klahr (Carnegie Mellon University)

What Does it Take to Pass the False Belief Task? An ACT-R Model 1045
Lara M. Triona and Amy M. Masnick (Carnegie Mellon University) and
Bradley J. Morris (University of Pittsburgh)

The Grounding of Symbols in Affordances 1046
William H. Vidal (Franklin and Marshall College)

Flexible use of prospective and retrospective memories 1047
Horatiu Voicu (Department of Psychological and Brain Sciences)

Motivational Patterns During Hypermedia Learning 1048
Regina Vollmeyer, Falko Rheinberg (Institut für Psychologie) and
Bruce D. Burns (Michigan State University)

Preschool Children's Use of Auditory Information in Drawing
Inferences about Animal Kinds 1049
Winnie H.K. Wai and Benise S.K. Mak (The University of Hong Kong)

If Only I Had Acted Differently: Reasons and Actions in
Counterfactual Thinking 1050
Clare R. Walsh and Ruth M.J. Byrne (University of Dublin, Trinity College)

The interaction effect of medium and pedagogy on semantic knowledge structure 1051
Alex Li Wang-on and John A. Spinks (The University of Hong Kong)

The Neural Instantiation of Number 1052
John W. Whalen and Frank Morelli (University of Delaware)

Partial Analogical Transfer in Problem Solving: Roles of Centrality and Order 1053
Tsunhin J. Wong and Albert W. L. Chau (University of Hong Kong)

Mental metalogic and its initial empirical justifications: The case of
reasoning with quantifiers and monadic predicates 1054
Yingrui Yang and Selmer Bringsjord (Rensselaer Polytechnic Institute)

"If" is easier than "or" in the GRE 1055
Yingrui Yang (Rensselaer Polytechnic Institute) and
Philip N. Johnson-Laird (Princeton University)

A Computerized Lexical Database of Cantonese 1056
Michael C. W. Yip (The Open University of Hong Kong)

Tutorials

Multiple Perspectives on Consciousness for Cognitive Science

Richard A. Carlson (racarlson@psu.edu)
Department of Psychology, Penn State University
613 Moore Building, University Park, PA 16802 USA

The huge contemporary literature on consciousness spans multiple disciplines, including psychology, philosophy, and neuroscience. This tutorial will introduce participants to major proposals about consciousness, and their empirical and methodological implications. The goal is to prepare participants to explore the consciousness literature in greater depth.

Our consideration of perspectives on consciousness will be organized by considering how these perspectives address core questions about consciousness, including: (a) How can *subjectivity* and *agency* be accommodated in a scientific theory of consciousness? (b) How can *conscious* and *nonconscious* or *unconscious* processes and representations be systematically distinguished? (c) How can conscious mental states be assessed or measured? (d) How can dissociations and impairments of consciousness be understood? The literatures to be considered address these questions in analytic, functional, computational, and implementational terms.

Philosophical Perspectives

Philosophers approach the problem of consciousness from a variety of analytic perspectives, some focusing on contemporary formulations of the mind-body problem and others on analyses of subjective experience. Among the philosophical perspectives we will consider are John Searle's (1992) analysis of consciousness in terms of intentionality, David Chalmer's (1996) distinction between "easy" and "hard" problems of consciousness, David Rosenthal's (1993) "higher order thought" proposal, and Daniel Dennett's (1991) "multiple drafts" theory of consciousness.

Neuroscience Perspectives

Neuroscientists have made a wide variety of proposals concerning the neural correlates of consciousness (NCC). A starting assumption is that a subset of current neural activity is correlated with current conscious experience. There is controversy, however, concerning how that subset is to be identified. For example, the NCC might be limited to particular types of cells or anatomical structures, or comprise global patterns of synchronized neural activity. We will consider recent proposals concerning NCC by Crick and Koch (1998), Damasio (2000), and Edelman and Tononi (2000).

Psychological Perspectives

Psychological perspectives on consciousness generally focus on functionally-defined aspects of cognition. For example, psychologists have identified consciousness with working memory (Baars, 1988), attention (Schneider & Pimm-Smith, 1997), metacognition (Nelson, 1996), and with the structure of mental states (Carlson, 1997). Cognitive research often focuses on distinguishing conscious and nonconscious influences on psychological processes such as learning (Dienes & Berry, 1997) and perception (Merikle, Smilek, & Eastwood, 2001). This research has generated a rich literature on methods for assessing consciousness.

References

Baars, B. J. (1988). *A cognitive theory of consciousness*. New York: Cambridge University Press.

Carlson, R. A. (1997). *Experienced Cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Crick, F., & Koch, C. (1998). Consciousness and neuroscience. *Cerebral Cortex, 8*, 97-107.

Damasio, A. R. (2000). A neurobiology for consciousness. In T. Metzinger (Ed.), *Neural correlates of consciousness* Cambridge, MA: The MIT Press.

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin and Review, 4*, 3-23.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Company.

Edelman, G. M., & Tononi, G. (2000). Reentry and the dynamic core: Neural correlates of conscious experience. In T. Metzinger (Ed.), *Neural correlates of consciousness*. Cambridge, MA: The MIT Press.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness: perspectives from cognitive psychology. Cognition, 79, 115-134.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102-116.

Rosenthal, D. M. (1993). Thinking that one thinks. In M. Davies, & G. W. Humphreys (Eds.), *Consciousness: Psychological and philosophical essays*. Oxford: Blackwell.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: The MIT Press.

Schneider, W., & Pimm-Smith, M. (1997). Consciousness as a message aware control mechanism to modulate cognitive processing. J. Cohen, & J. Schooler (Eds.), *Scientific approaches to consciousness: The 25th Carnegie Symposium on Cognition*. Mahwah, NJ: Erlbaum.

APEX/CPM-GOMS: Modeling Human Performance in Applied HCI Domains

Half-day tutorial (morning)
Johnson Center/Enterprise

Roger Remington
NASA Ames Research Center
Moffett Field, CA 94035
mfreed@arc.nasa.gov

Bonnie John - Carnegie Mellon University

Michael Matessa - NASA Ames Research Center

Alonso Vera - NASA Ames Research Center

Michael Freed- NASA Ames Research Center

This tutorial introduces participants to CPM-GOMS modeling using APEX, a tool for applied human performance modeling. APEX-CPM is intended to be of value to both cognitive science researchers and HCI professionals. It is also valuable for teaching students about task-analysis, user-modeling, and computational cognitive modeling in general. The tutorial will teach participants how to represent GOMS, KLM, and CPM-GOMS task analyses in APEX-CPM, and to refine models based on output in the form of automatically generated PERT charts. We will also discuss recent improvements aimed at making it practical to model in more complex HCI domains. These include capabilities that (a) facilitate representation of simulation environments and (b) allow modelers to draw on a set of reusable building blocks both for cognitive/task modeling and for physical environment modeling. The tutorial will primarily consist of a guided tour through activities supplemented with presentations. Participants will work in pairs, supervised by at least four presenters, on applied modeling problems. This tutorial should be of particular value to people interested in developing, using, and/or teaching engineering models of human performance in HCI contexts. Some background in cognitive modeling and computer programming is recommended. Tutorial participants will be given a CD containing the APEX-CPM code, a world-building tool-kit, and a number worked example models.

Roger Remington is a Senior Research Psychologist at NASA Ames Research Center. He holds a Ph.D. in Psychology from the University of Oregon.

Bonnie John is an Associate Professor in the Institute of Human-Computer Interaction and Carnegie-Mellon University. She holds a Ph.D. in Psychology from Carnegie Mellon University.

Michael Matessa is a Research Psychologist at NASA Ames Research Center. He holds a Ph.D. in Cognitive Psychology from Carnegie Mellon University.

Alonso Vera is a Research Scientist at NASA Ames Research Center. He holds a Ph.D. in Experimental Psychology from Cornell University.

Michael Freed is a Research Scientist at NASA Ames Research Center. He holds a Ph.D. in Computer Science from Northwestern University.

How to Build Intelligent Interactive Agents Using Soar

Randolph M. Jones
Soar Technology, Inc., and Colby College
rjones@soartech.com

Robert E. Wray, III, Soar Technology, Inc.
Amy E. Henninger, Soar Technology, Inc.
Scott Wood, Soar Technology, Inc.
Ronald S. Chong, George Mason University

Soar has been under development for over two decades as an architecture for building intelligent systems and human behavior models. Recent research and development activity with Soar has emphasized building competent, autonomous agents that interact with realistic and complex simulated environments. This tutorial will teach some of the methods that we use to design and engineer such behavior models. Instead of focusing on strict cognitive modeling, this tutorial will discuss the complexities that autonomous behavior and real-time interaction impose on a model. It will not involve intensive programming of intelligent agents, but will concentrate on higher level issues of task analysis, knowledge representation, process-oriented modeling, and knowledge acquisition. Many of these activities are useful to learn even if one does not use Soar to implement models, but the tutorial will also demonstrate the ways that Soar approaches and informs (and sometimes exacerbates) these tasks. To this end, tutorial participants will study and tweak a variety of interactive behavior models, and learn techniques for representing knowledge and behavior in Soar. They will also gain experience with some of the new development tools that support Soar modeling. Tutorial participants do not need an extensive background in programming.

Presenters: Each of the presenters has research and industry experience building interactive intelligent agent systems. Many of these have been "believable" agents with large amounts of knowledge developed within Soar. Soar Technology, Inc., uses Soar and other software paradigms to create intelligent and usable software for a variety of defense applications. All of the presenters also have experience developing interactive models of human behavior for various purposes, such as improving intelligent agents, improving human-computer interaction, understanding learning in problem solving, and studying human error in interactive tasks.

ACT-R

Christian Lebiere
Carnegie-Mellon University

ACT-R is a cognitive theory and simulation system for developing cognitive models. It assumes cognition emerges through the interaction of a procedural memory of productions with a declarative memory of chunks and independent modules for external perception and actions. The ACT-R 4.0 version of the theory was detailed in the book "The Atomic Components of Thought" by John R. Anderson and Christian Lebiere, published in 1998 by Lawrence Erlbaum. Since its release in 1997, ACT-R 4.0 has supported the development of over 100 cognitive models published in the literature by many different researchers. These models cover topics as diverse as driving behavior, implicit memory, learning backgammon, metaphor processing, and emotion. We have recently developed a new version, ACT-R 5.0 that extends ACT-R 4.0 to be more interruptible, to achieve greater across-task parameter consistency, to have better mechanisms of production learning, and to be more in correspondence with our knowledge of brain function. While the new system extends the capabilities of ACT-R 4.0, it involves relatively few changes and is actually simpler. This short tutorial will provide an overview of ACT-R, as it is specified in the 5.0 version, and some of its applications. It will not assume a prior background in ACT-R 4.0.

Christian Lebiere is a Research Scientist in the Human-Computer Interaction Institute at Carnegie-Mellon University. He received his B.S. in Computer Science from the University of Liege (Belgium) and his M.S. and Ph.D. from the School of Computer Science at Carnegie Mellon University. During his graduate career, he worked on the development of connectionist models, including the Cascade-Correlation neural network learning algorithm. Since 1990, he has worked on the development of the ACT-R hybrid cognitive architecture and is co-author with John R. Anderson of the 1998 book "The Atomic Components of Thought". His main research interest is cognitive architectures and their applications to psychology, artificial intelligence, human-computer interaction, decision-making, game theory, and computer-generated forces.

Functional Imaging of the Brain -- Developing a Synergy of Cognitive Neuroscience Behavior and Modeling

Walter Schneider
University of Pittsburgh
Pittsburgh 152260
wws@pitt.edu

The last ten years have produced an explosive growth in brain imaging technology and findings. The combination of MRI, ERP, DTI, and PET enable non-invasive research on humans obtaining millimeter and millisecond resolution of activation, connection tracing, and mapping of transmitter systems. This large effort (1000+ papers per year) is providing detailed data of the biology of cognition and having a large impact on the conceptualization of cognitive science. There is a "grand challenge" to the field to relate the biology and mechanisms of human thought. There is a critical need for comprehensive behavioral, theoretical, and modeling efforts to interpret the findings. This tutorial will provide an introduction to the brain imaging methods stressing both the potential and limitations of the existing methods. We will describe the challenges that cognitive science methods may be particularly beneficial to resolving. We will look at efforts to relate modeling (e.g., ACTR, LSA) and activation data. We will provide guidelines on how to get into brain imaging via collaboration or direct imaging.

Walter Schneider (B.A. Psychology, U. Illinois 1971; Ph.D. Psychology Indiana U. 1975). He is a fellow of the American Psychology Association and AAAS, is know for his classic work on automaticity and skill acquisition, published some of the first papers on fMRI in humans, and has developed software systems for empirical and brain imaging systems used in 2,000 laboratories. His current research focuses on brain imaging and modeling of learning, attention, and language processing and the modeling of skill acquisition and control/automatic processing. His web site is www.pitt.edu/~schlab/People/walt.htm

A Cognitive Approach to Designing Human Error Tolerant Interfaces

Scott D. Wood (swood@soartech.com)
Soar Technology, Inc.
3600 Green Court, Suite 600, Ann Arbor, MI 48105 USA

Mike Byrne (byrne@acm.org)
Psychology Department, Rice University
6100 Main Street, MS-25, Houston, TX 77005 USA

Human errors are inevitable, but some are more inevitable than others. Human error has been blamed for countless catastrophes (cf. Casey, 1993; Perrow, 1984), yet errors are often merely symptoms of much larger underlying design problems. Designing for human error is a major challenge for developers of safety-critical and mission-critical systems. Human error is of particular concern for banking, commerce, medicine, military and other systems where tasks are performed with high frequency, or where the consequences of mistakes are grave or costly. Most approaches to error-tolerant design use either general design guidelines or treat humans as just another error-prone system component. To make errors a little less inevitable, we must take a comprehensive, psychologically-based approach to human-error tolerant design.

The goals of this tutorial are to provide researchers with a better understanding of the underlying causes of human error and present practical techniques for applying psychology to human-error tolerant designs. This tutorial approaches error tolerant design from a cognitive perspective, focusing on practical techniques for improving your system's ability to deal with inherent human limitations. Participants will learn the basics of human error, how to classify error types, a framework for error tolerant design, how to deal with multiple aspects of error in design, and how to form a multilayered defense against error.

Understanding Human Error

Reason (1990) defines and discusses many aspects of human error and its psychological underpinnings. Although there are many questions remaining about the attentional mechanisms used by Reason to explain error occurrence, his work provides a starting point from which to study human error and error tolerant design. Reason's human error taxonomy is framed around Rasmussen's Skills-Rules-Knowledge (SRK) framework (1979). In addition to specific error types within each of the SRK levels, Reason also identifies two general forms of error, "similarity matching" and "frequency gambling", that pervade each of the levels. This leads to general rules for erroneous memory retrieval, such as, "If the correct item is not retrieved, then the most similar, frequently accessed item will be retrieved." Additional insight can be gleaned by mapping Reason's taxonomy unto a standard human information processing architecture, such as ACT-R, EPIC or Soar.

Understanding and categorizing human error can be done at many levels, depending on the focus of the study. Reason (1990) also describes different levels for classifying error instances as one of behavioral (observable actions, such as omitting a procedural step), contextual (within the context of the task, such as "failed to press the button"), or conceptual (relating to internal mechanisms, such as "failed to perceive warning label").

Designing Error-Tolerant Systems

Many error-tolerant design efforts rely on general design guidelines (e.g. Smith and Mosier, 1986). However these guidelines are often not used or misapplied because they are sometimes overly general or atheoretical in nature. Mayhew (1992), provides a more theoretical mapping of guidelines into motor, cognitive, and perceptual areas, but has no framework to comprehensively address human error.

An effective error-tolerant design must address multiple aspects of human error to build a comprehensive, multi-layered defense. These aspects include error prevention, reduction, detection, identification, and correction, resumption of normal activities, and failure mitigation. In this tutorial we will describe a framework for mapping between error taxonomies, cognitive architectures and design guidelines, and present a task-analytic technique for applying the framework in a practical way.

Understanding the psychological factors that affect these areas is essential for good design. Likewise, understanding practical implications of applied psychology may lead to further theoretical advances in the field.

References

Casey, S. (1993). *Set Phasers on Stun*. Santa Barbara, CA: Aegean Publishing Co.

Mayhew, D. J. (1992). Principles and guidelines in software user interface design. Englewood Cliffs, NJ: Prentice-Hall.

Perrow, C. (1984). Normal Accidents. New York, NY: Basic Books, Inc.

Rasmussen, J. (1979). What Can be Learned from Human Error Reports? (Riso Report N- 17-79): Riso National Laboratory.

Reason, J. (1990). Human Error. New York: Cambridge University Press.

Smith, S. L., & Mosier, J. N. (1986). Guidelines for designing user interface software (Report ESD-TR-86-278). Bedford, MA: The MITRE Corporation.

Rumelhart Prize Talk

Bayesian Modeling of Memory and Perception

Richard M. Shiffrin

Abstract

I present a framework for modeling memory, retrieval, and perception, and their interactions. The models are inspired by Bayesian induction to determine optimal decisions, in the face of a memory system with inherently noisy storage and retrieval. The starting point for this work was the Retrieving Effectively from Memory (REM) model for episodic recognition (Shiffrin & Steyvers, 1997). The general framework describes: 1) the storage of episodic traces, the accumulation of these into knowledge (e.g. lexical/semantic traces in the case of words), and the changes in knowledge caused by learning; 2) the retrieval of information from episodic memory and general knowledge; 3) decisions concerning storage, retrieval and responding. I give examples of applications to episodic recognition, and cued and free recall, perceptual identification (naming, yes-no and forced choice), lexical decision, and long-term and short-term priming, and briefly consider extensions to episodic categorization and retrieval of content from general knowledge.

Rumelhart Symposium

Rumelhart Symposium: Honoring Richard Shiffrin

Susan Dumais (sdumais@microsoft.com)
Microsoft Research, Microsoft Corporation
One Microsoft Way, Redmond, WA 98033 USA

Wilson S. Geisler (geisler@psy.utexas.edu)
Department of Psychology, University of Texas
Mezes Hall., Austin, TX 78712 USA

Jeroen Raaijmakers (raaijmakers@psy.uva.nl)
Department of Psychology, University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

Mark Steyvers (msteyver@psych.stanford.edu)
Dept. of Cognitive Sciences, University of California
3151 Social Sciences Plaza, Irvine, CA 92697 USA

This symposium honors Rich Shiffrin's contributions to cognitive science. Four former students will present highlights of their current work shaped in various ways by their mentor.

Data-Driven Approaches to Information Access (Susan Dumais)

Several lines of research that are motivated by the practical problem of helping users find and manage information in external data sources, most notably computers, will be described. The application areas include: information retrieval, text categorization, and question answering. A common theme in all these efforts is the analysis of the statistical properties of words in large volumes of real world texts. Simple statistical analyses and machine learning algorithms are used to solve practical information access problems. In addition these same statistical properties of objects in the world constrain human performance. Thus solutions to practical problems can shed light on human knowledge representation and reasoning.

A Bayesian Approach to the Evolution of Perceptual Systems (W. S. Geisler and R. L. Diehl)

Perceptual and cognitive systems, including the developmental and learning mechanisms that shape them during the lifespan, are the result of evolution by natural selection. Yet historically most approaches to the study of perception and cognition acknowledge only implicitly the role of natural selection. We propose a Bayesian theoretical framework that makes explicit the relationship between the statistical properties of the environment, the evolving genome, and the design of perceptual and cognitive systems. The proposed framework grew out of recent applications of Bayesian statistical decision theory in perception and cognition and recent efforts to measure the statistical properties of natural environments; however, the Bayesian framework encompasses many of the most important insights of previous theoretical approaches in perception and cognition. We first summarize the formal Bayesian framework and show how it can be used to formulate and test specific hypotheses about the design of perceptual and cognitive systems. We then describe the connections between the Bayesian framework and other theoretical approaches.

SAM as a General Theory for Memory (Jeroen G. W. Raaijmakers)

The SAM theory for memory retrieval will be briefly reviewed. We show how the theory (including the extension proposed by Mensink & Raaijmakers, 1988, 1989) may be used to provide a new model for spacing and repetition effects. The resulting model can be seen as a mathematical formulation of the Component-Levels theory proposed by Glenberg (1979). It is assumed that on a second presentation of an item information is added to an existing trace if the episodic memory image corresponding to that item is retrieved. If it is not retrieved, a new image is stored. It is shown that the model predicts many standard findings including findings that were thought to be inconsistent with the Component-Levels theory. This application demonstrates how SAM may be used to provide quantitative formulations for verbal theories, making it easier to examine the exact predictions of such theories.

Inferring Causal Structure from Intervention (Mark Steyvers)

Information about the structure of a causal system can come in the form of observational data - random samples of the system's autonomous behavior - or interventional data - samples conditioned on the particular values of one or more variables that have been experimentally manipulated. Here we study people's ability to infer causal structure from intervention, and to choose informative interventions on the basis of purely observational data. We develop computational models of how people infer causal structure from data and how they plan intervention experiments, based on the representational framework of causal Bayesian networks and the inferential principles of optimal Bayesian decision-making and maximizing expected information gain. These analyses suggest that people can make rational causal inferences, subject to certain processing constraints and representational assumptions that may vary across participants.

Plenary

Cognitive Science as the Engine of Innovation: Beyond Human-Computer Interaction

Stuart Card (card@parc.com)
Information Sciences and Technologies Laboratory, Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304 USA

Abstract

Successful sciences usually spawn successful applications and application disciplines; in fact, one is suspicious of a science that can't claim practical results. The na ve view is that results from the science are "applied" to problems, as in an "applied cognitive psychology," for example. The truth is more complex in general and it is particularly more complex for cognitive science. New advances in technology are amplifying still further the human role of informavore and the need for cognitive engineering and invention of cognitive products and government activities. The ability to meet these is a test of a cognitive engineering discipline and of the supporting sciences themselves. I am going to suggest some principles for organizing both cognitive science and the practical innovation around it by reflecting on what we have learned about using cognitive psychology in human-computer interaction. I will use this analysis to suggest a set of initiatives now within reach of the cognitive science community.

About Stuart Card

Stuart Card is a Xerox Research Fellow and the manager of the User Interface Research group at the Xerox Palo Alto Research Center. With Allen Newell and Tom Moran from CMU, he founded an effort to develop models of human performance that could be used in information system design. His thesis at CMU was the first thesis specifically in the new specialty of human-computer interaction. His study of input devices led to the Fitt's Law characterization of the mouse and was an important factor leading to the mouse's commercial introduction.

He and his group have developed a number of theories of human-machine interaction, including tne Model Human Processor, the GOMS theory of user interaction, and information foraging theory. They have developed new paradigms of human-machine interaction, including the Rooms workspace manager and the Information Visualizer. The work has resulted in nine Xerox products and the founding of Inxight Software, Inc.

Card is a co-author of the book, "The Psychology of Human-Computer Interaction", a co-editor of the book, "Human Performance Models for Computer-Aided Engineering", and has served on many editorial boards. He received his A.B. in Physics from Oberlin College and his Ph.D. in Psychology from Carnegie Mellon, where he pursued an interdisciplinary program in psychology, artificial intelligence, and computer science. His most recent book, "Readings in Infomration Visualization", co-written and edited with Jock Mackinlay and Ben Schneiderman, was published in January 1999.

.

Steering the Reverberations of Technology Change on Fields of Practice: Laws that Govern Cognitive Work

David D. Woods (woods.2@osu.edu)
Institute for Ergonomics
The Ohio State University
1971 Neil Ave
Columbus, OH 43210 USA

> Now all scientific prediction consists in discovering in the data of the distant past and of the immediate past (which we incorrectly call the present), laws or formulae which apply also to the future, so that if we act in accordance with those laws our behavior will be appropriate to the future when it becomes the present.
>
> Craik, 1947, p. 59

Abstract

Research on cognitive work in context has abstracted a set of common patterns about cognitive work and about the relationship of people and computers. I offer four families of Laws that Govern Cognitive Work plus Norbert s Contrast as a synthesis of these findings to guide future development of human-computer cooperation. These Laws are one prong of a general strategy to avoid repeats of past "automation surprises".

1. Patterns of Reverberations

Observational studies of cognitive work in context have built a body of work that describes how technology and organizational change transforms work in systems. Points of technology change push cycles of *transformation* and *adaptation* (e.g., Carroll s task-artifact cycle; Carroll and Rosson, 1992; Winograd and Flores, 1987; Flores, Graves, Hartfield, and Winograd, 1988). The review of the impact of new technology in one operational world effectively summarizes the general pattern (Cordesman and Wagner, 1996, p.25):

> Much of the equipment deployed ... was designed to ease the burden on the operator, reduce fatigue, and simplify the tasks involved in operations. Instead, these advances were used to demand more from the operator. Almost without exception, technology did not meet the goal of unencumbering the personnel operating the equipment
> ... systems often required exceptional human expertise, commitment, and endurance.
> there is a natural synergy between tactics, technology, and human factors ... effective leaders will exploit every new advance to the limit. As a result, virtually every advance in ergonomics was exploited to ask personnel to do more, do it faster and do it in more complex ways.
> ... one very real lesson is that new tactics and technology simply result in altering the pattern of human stress to achieve a new intensity and tempo of operations. [edited to rephrase domain referents generically]

This statement could have come from studies of the impact of technological and organizational change in health care or air traffic management or many other areas undergoing change today (see Billings, 1997, and Sarter and Amalberti, 2000, for the case of cockpit automation). Overall, the studies show that when black box new technology (and accompanying organizational change) hits an ongoing field of practice the pattern of reverberation includes (Woods and Dekker, 2000):

- New capabilities, which increase demands and create new complexities such as increased coupling across parts of the system and higher tempo of operations,
- New complexities when technological possibilities are used clumsily,
- Adaptations by practitioners to exploit capabilities or workaround complexities because they are responsible to meet operational goals,
- The complexities and adaptations are surprising, unintended side effects of the design intent,
- Failures occasionally break through these adaptations because of the inherent demands or because the adaptations are incomplete, poor, or brittle,
- The adaptations by practitioners hide the complexities from designers and reviewers after-the-fact who judge failures to be due to human error.

The pattern illustrates a more general law of adaptive systems that has been noted by many researchers (e.g., Rasmussen, 1986; Hirschhorn, 1997)

The law of stretched systems:
every system is stretched to operate at its capacity; as soon as there is some improvement, for example in the form of new technology, it will be exploited to achieve a new intensity and tempo of activity.

Under pressure from performance and efficiency demands, advances are consumed to ask operational personnel to do more, do it faster or do it in more complex ways (see NASA s Mars Climate Orbiter Mishap Investigation Board report, 2000, for a example).

2. Watching People Engineer Cognitive Work: Claims and Myths

People as advocates for investment in and adoption of new technology make claims about how these changes will affect cognitive work and the processes and products of practice. Claims about the future of practice if objects-to-be-realized are deployed represent hypotheses about the dynamics of people, technology and work (Woods, 1998). Observations

at points of technology change find that these hypotheses can be and are often quite wrong a kind of second order automation surprise (Sarter, Woods, and Billings, 1997). Envisioning the future of operations, given the dynamic and adaptive nature of the process, is quite fragile.

What patterns emerge from observations of people engineering cognitive work or of people s claims about how various advances-in-process will enable the re-engineering of cognitive work? Remarkably consistently, we observe over-simplifications (Feltovich et al., 1997) that claim the introduction of new technology and systems into a field of practice substitutes one agent for another, essentially, computer capabilities as substitute for erratic human performance. Yes, the claims of opposition of human and machine come cloaked in different and often quite sophisticated forms, yet underneath inter-substitutability or Fitts List remains the core people and machines are or can be equivalent so that new technology (with the right capabilities) can be introduced as a simple substitution of machines for people preserving the system though improving the results. This oversimplification fallacy is so persistent it is best understood as a cultural myth the Substitution Myth (Woods and Tinapple, 1999).

The myth creates difficulties because it is wrong, empirically adding or expanding the machine s role changes the cooperative architecture and changes human roles, introduces capabilities and complexities that are part of multiple adaptive cycles as human actors and stakeholders jostle in the pursuit of their goals. But moreover, the myth is unproductive as it locks us into cumbersome trial and error processes of development, blocks understanding the demands of cognitive work in context and how people in various roles and groups adapt to those demands, and channels energy away from processes of innovating use from the continually expanding power of machine information processing.

How can we better calibrate and ground claims about the future of cognitive work to avoid past cycles where change exacerbated clumsy use of technology and limited adaptations from people responsible to meet system goals? One possible tactic is to develop generalizations or laws that govern cognitive work by any cognitive agent or any set of cognitive agents from the empirical base. Such Laws could serve as a guide to enhance the use information processing technology in a practice—centered R&D process (Woods and Christofferesen, in press).

3. Predicting and Steering Change in Cognitive Work

Based on patterns about cognitive work and about the relationship of people and computers abstracted from research on cognitive work in context, I offer four families of Laws that Govern Cognitive Work as a synthesis to guide future development of human-computer cooperation (the approach is a deliberate play off Conant s 1976 laws of information that govern systems). I also offer Norbert s Contrast (Wiener, 1950) as an alternative conception of the relationship between people and computers. The current draft set of Laws is available from the author.

These laws are built on a foundation of agent-environment mutuality. Agents' activities are understandable only in relationship to the properties of the environment within which they function and an environment is understood in terms of what it demands and affords to potential actors in that world. Each is mutually adapted to the other.

The Laws fall into four families plus Norbert's Contrast. First, Laws of Adaptation build on original insights of cybernetics and control (Ashby, 1957; Conant, 1976). The driving force here is how cognitive systems adapt to the potential for surprise in the worlds of work, i.e., the foundational slogan for Cognitive Systems Engineering from Jens Rasmussen *adaptations directed at coping with complexity and surprise* (Rasmussen and Lind, 1981; Woods, 1988; Woods and Christoffersen, in press).

Laws of Models are concerned with how we understand and represent the processes we control and the agents we interact with. The driving force here is the mystery of how expertise is tuned to the future, while, paradoxically, the data available is about the past.

Laws of Collaboration address how cognitive work is distributed over multiple agents and artifacts. The driving force here is the fact that cognitive work always occurs in the context of multiple parties and interests as moments of private cognition punctuate flows of interaction and coordination. The idea that cognition is fundamentally social and interactive, not private, radically shifts the basis for analyzing and designing cognitive work and reconsidering the relationship between people and computers.

Quite surprisingly, Laws of Responsibility are the fourth family, driving home the point that in cognition at work, whatever the artifacts and however autonomous that are under some conditions, people create, operate, and modify these artifacts in human systems for human purposes.

Fifth, based on these Laws, Norbert's Contrast goes behind our fascination with increasing the power of the computer to remind us of the limits of literal minded agents and the unique competences of human cognition to handle the tradeoffs and dilemmas of a changing, finite resource, uncertain world (Wiener, 1950).

> *Norbert s Contrast*
>
> Artificial agents are literal minded and disconnected from the world, while human agents are context sensitive and have a stake in outcomes.

The key is people and computers *start* from different opposite points and tend to *fall back* or default to those points without the continued investment of effort and energy from outside the system.

Each of these families of Laws and Norbert's Contrast is quite surprising even shocking given conventional beliefs about cognition, organizations, and computers. The Laws allows us to see past these conventional beliefs to re-consider relationships across people, computers, the goals of various stakeholders and the complexities and variations in the worlds of human activity as we envision and create the future of operations.

Laws that Govern Cognitive Work have an odd quality—-they appear optional. Designers of systems that perform

cognitive work do not have to follow them. In fact, we notice these laws through the consequences that have followed repeatedly when design breaks them in varying episodes of technology change. The statements are law-like in that they capture regularities of control and adaptation of cognitive work, and they determine the dynamic response, resilience, stability or instability of the distributed cognitive system in question. While developers may find following the laws optional, what is not optional is the consequences that accrue predictably from breaking these laws, consequences that block achieving the performance goals developers and paractitioners, technologists and stakeholders set.

Respect for the Laws is essential, for in the final analysis:

> in design, we either hobble or support people s natural ability to express forms of expertise.

Acknowledgments

This piece is a companion and follow up to a previous address to the Cognitive Science Society in 1994, Observations from Studying Cognitive Systems in Context.

Many thanks to the various colleagues who in one way or another helped identify how generalizations like these operate in cognitive work.

Prepared in part through participation in the Advanced Decision Architectures Collaborative Technology Alliance sponsored by the Army Research Laboratory under Cooperative Agreement DAAD 19-01-2-0009.

References

Ashby, W. R. (1957). *An Introduction to Cybernetics*. Chapman and Hall, London.

Billings, C. E. (1997). *Aviation Automation: The Search For A Human-Centered Approach*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Carroll, J.M. & Rosson, M. B. (1992). Getting around the task-artifact cycle: How to make claims and design by scenario. *ACM Transactions on Information Systems*. 10, 181-212.

Conant, R. C. (1976). Laws of information which govern systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, 240-255.

Cordesman, A. H. & Wagner, A. R. (1996). *The Lessons of Modern War, Vol.4: The Gulf War*, (Boulder, CO: Westview Press).

Craik, K. J. W. (1947) Theory of the operator in control systems: I. The operator as an engineering system. *British Journal of Psychology*, 38, 56-61.

Feltovich, P.J., Spiro, R.J., & Coulson, R.L (1997). Issues of expert flexibility in contexts characterized by complexity and change. In P.J. Feltovich, K.M. Ford, & R.R. Hoffman (eds.), *Expertise in context: Human and machine*. Menlo Park, CA. AAAI/Mit Press.

Flores, F., Graves, M., Hartfield, B. & Winograd, T. (1988). Computer systems and the design of organizational interaction. *ACM Transactions on Office Information Systems,* 6, 153-172.

Hirschhorn, L. (1997). Quoted in Cook, R. I., Woods, D. D. and Miller, C. (1998). *A Tale of Two Stories: Contrasting Views on Patient Safety.* National Patient Safety Foundation, Chicago IL, April 1998 (available at www.npsf.org).

NASA, Mars Climate Orbiter Mishap Investigation Board. (2000). Report on Project Management at NASA, March 13, 2000.

Rasmussen, J. (1986). Information processing and human-machine interaction: An approach to cognitive engineering. Amsterdam: North-Holland.

Rasmussen, J. & Lind M. (1981). *Coping with complexity* (Ris¿-M-2293). Ris¿ National Laboratory, Roskilde, Denmark: Electronics Department.

Roesler, A. Feil, M. & Woods, D.D. (2002). Design is Telling (Sharing) Stories about the Future. Draft Working MediaPaper at url: http://csel.eng.ohio-state.edu/animock

Sarter, N. & Amalberti, R., eds. (2000). *Cognitive Engineering in the Aviation Domain*, Erlbaum, Mahwah NJ.

Sarter, N., Woods, D.D. & Billings, C.E. (1997). Automation Surprises. In G. Salvendy, editor, *Handbook of Human Factors/Ergonomics*, second edition, Wiley, New York.

Wiener, N. *The Human Use of Human Beings: Cybernetics and Society*, Doubleday NY,1950.]

Winograd, T. and Flores, F. (1986). *Understanding computers and cognition.* Norwood, NJ, Ablex.

Woods, D.D. (1988). Coping with complexity: The psychology of human behavior in complex systems. In L.P. Goodstein, H.B. Andersen, and S.E. Olsen, editors, *Mental Models, Tasks and Errors*, Taylor & Francis, London, (p. 128-148).

Woods, D. D. (1998). Designs are hypotheses about how artifacts shape cognition and collaboration. *Ergonomics*, 41, 168-173.

Woods, D. D. & Christoffersen, K. (in press). Balancing Practice-Centered Research and Design. In M. McNeese and M. A. Vidulich (editors), *Cognitive Systems Engineering in Military Aviation Domains.* Wright-Patterson AFB, OH: Human Systems Information Analysis Center.

Woods, D. D. & Dekker, S. W. A. (2000). Anticipating the Effects of Technological Change: A New Era of Dynamics for Human Factors. *Theoretical Issues in Ergonomic Science*, 1(3), 2000.

Woods, D. D. & Tinapple, D. (1999). W^3: Watching Human Factors Watch People at Work. Presidential Address, 43rd Annual Meeting of the Human Factors and Ergonomics Society, September 28, 1999. Multimedia Production at http://csel.eng.ohio-state.edu/hf99/

Symposium

The Cognition of Complex Visualizations

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory

Priti Shah (priti@umich.edu) and Eric G. Freedman (freedman@umflint.edu)
Department of Psychology, University of Michigan

Susan Kirschenbaum (KirschenbaumSS@npt.nuwc.navy.mil)
Naval Undersea Warfare Center

Peter C-H. Cheng (peter.cheng@nottingham.ac.uk)
University of Nottingham

Discussant: Mary Hegarty (hegarty@psych.ucsb.edu)
Department of Psychology, University of California, Santa Barbara

Abstract

We explore current research on how complex visualizations are perceived, comprehended, used, and taught.

Introduction

How do people perceive, comprehend, and use complex visualizations, and when are they needed? Many domains (meteorology, scientific visualization, stock market analyses) deal with very complex data that must be displayed and used in novel ways. Unfortunately, very little is known about how these complex displays are used, how to best display complex graphical information, or how to design good complex visualizations for teaching purposes. This symposium will examine:

- how people understand and use complex visualizations;
- how people gain expertise in using complex visualizations;
- how to teach complex domains by using graphs and visualizations;
- how to visualize uncertainty across many variables;
- why a visualization is hard or easy to use; and
- how current models of graph comprehension scale up to more complexity.

Building Qualitative Mental Models

Greg Trafton

How do people use a complex visualization? Most current theories predict a straightforward process of reading off specific information, typically at the request of an experimenter. Many complex domains, however (many areas of scientific visualization, meteorology, etc.) need to deal with multi-dimensional data with complex interactions and anomalies.

I will present several recent studies that show that while experts mostly conform to the standard models of graph comprehension, there are some glaring holes in current theories. Specifically, experts do more than simply read off information. First, they extract primarily qualitative information from complex visualizations (e.g., "The wind is fast over San Diego") even when quantitative information is available and needed later. With this qualitative information, they build a complex mental representation (which we call a qualitative mental model, or QMM) to reason with.

I will present data that shows how experts build these complex mental structures by looking at complex visualizations. I will also present evidence from eye-tracking and protocol studies of experts and novices working in their own domain, showing how novices seem to conform to the standard graph comprehension models while experts do not.

The Role of Prior Knowledge in Complex Data Comprehension

Priti Shah & Eric G. Freedman

People are increasingly faced with the task of interpreting complex˚multivariate quantitative data sets. Unfortunately, much research on graph interpretation has focused on how novice (college undergraduate) viewers use common formats˚ (e.g., bar and line graphs) for simple tasks (e.g., read a data point or describe a trend) and sparse (2-3 variables and few data points) and meaningless (axes labeled x and y) data.˚ In our presentation, we argue that models based on this research may not scale up to account for more complex data interpretation, which differs in several key features. Complex data interpretation usually refers to tasks involving many variables, complex interactions between the variables, and a large number of data points. Complexity extends beyond simply data complexity, however. Dealing with complex data coincides with complex tasks (e.g., making decisions or explaining data) rather than fact retrieval.˚ Complex data also involves the extensive use of prior knowledge and viewers with data interpretation skills (experts use complex data, not novices).˚ Finally, complex data is often presented via

specialized displays that sometimes incorporate animation and interactivity. Models of complex data interpretation thus require considering data complexity, domain knowledge, graph reading skills, and display characteristics.

In a number of recent studies we consider how these factors play a role in viewers' interpretations. Our results suggest that domain knowledge and data interpretation skills influence viewers' comprehension and use of complex data. In addition, domain knowledge and data interpretation skills interact with data complexity and display characteristics. Specifically, prior knowledge reduces complexity in two ways:˚ experts know what to look for and also how to retrieve that information from complex displays.˚ Second, prior knowledge reduces the influence display characteristics on viewers' comprehension of data. Third, prior knowledge helps viewers integrate data and theory and understand implications of the data.˚ Finally, display characteristics such as interactivity and visual cues also reduce cognitive complexity.

We describe a model of graph comprehension based on these results. By incorporating the interaction of top down factors (e.g., domain knowledge) and the bottom up influence of display characteristics, our model builds on prior models but provides a more comprehensive description of complex data interpretation.

Visualizing Uncertain Information

Susan Kirschenbaum

There are many qualities that make visualizations complex. Perhaps the most clear-cut definition is the multi-variable visualization. The variables may be incompatible or difficult to display in a single visualization. For example, to display (in support of maneuver decisions) the course, speed, range, depth, time, and relative motion (how two moving objects relate to one another) of submarines moving through the sound field that is the ocean requires multiple complex visualizations. Weather forecasters have similar problems. The visualizations can be either graphs or, more often, geo-referenced displays.

Visualizations of these situations would be complex even without the added problem of uncertainty. However, the submarine world is characterized by the extensive uncertainty due to limited measurable data, indeterminate algorithmic solutions, and sound transmission characteristics underwater. Naturally, with multiple variables and uncertainties, there are many options for visualizing the problem. Some limit the number of variables; many ignore or discretize uncertainty. Alternatively, there are many ways to visualize uncertainty (Pang, et al., 1997). Even when uncertainty is not displayed, decision makers find ways to assess it by multiple comparisons; across variables of interest, and by comparing models with predicted, modeled, or measured data.

I will show evidence from verbal protocol and eye-tracking data of how decision makers interpret uncertainty in visualizations and of the impact of task and expertise on the effectiveness of various visualization options.

Designing representational systems to study complex visual cognition

Peter C-H. Cheng

Representational epistemology is the term I use to succinctly describe our work on the nature cognition with complex visualizations. The central theoretical claim of representational epistemology is that representational systems are fundamental to the highest forms of human cognition, such as problem solving, conceptual learning and scientific discovery. In such activities the acquisition and transformation of knowledge is essential, so understanding the nature of the representations that codify that knowledge will be critical. For instance, in the context of conceptual learning we theorize that an effective representation will substantially determine: what is learnt; how easily learning occurs; the nature of the conceptual structures that develop; the problem solving procedures that are acquired.

There are five common stages to our representational epistemological studies. First, a conceptually demanding knowledge rich domain is selected. Educational domains that we have addressed include mechanics, electricity and probability theory. The approach is also being applied to the intensive real-world problem of University examination scheduling. Second, the content and problem classes of the domain are analyzed to reveal the underlying conceptual structure of the knowledge, which includes the ontologies, perspectives, scale levels, laws, models, prototypes and extreme cases of the domain. Third, the existing domain representations are examined to uncover the conceptual problems they cause. Fourth, a new diagrammatic system is invented to encode the inherent conceptual structure of the domain. The novel representations that we have invented are Law Encoding Diagrams, LED. By directly reflecting the conceptual structure of a target domain in its representational structure a LED is a re-codification of knowledge that should support comprehension, problem solving and learning of the domain. The LEDs are of sufficient novelty and potential that papers describing them and their use have been accepted for publication in domain specific journals. Fifth, empirical evaluations of problem solving and learning with the new LED compared to the conventional representations of the domain are conducted in the laboratory or in Schools.

Studies conducted in the domains mentioned above show that LEDs improve problem solving and conceptual learning compared to the conventional domain representations. By generalizing over these different domains, contrasting the various LEDs and the existing representations, we are formulating principles for the design of effective representations for complex knowledge rich domains. Two classes are posited: semantic transparency principles that address how the underlying conceptual structure of a domain should be encoded in the inherent structure of the representation; syntactic plasticity principles that consider how a representation should be structured to support efficient problem solving.

Nature's Turing Test

Symposium Organized by:
Thomas R. Zentall (zentall@uky.edu)
Department of Psychology, Department of Psychology,
University of Kentucky, Lexington, KY 40506-0044

Introduction

What, if anything, is special about human cognition and how might we find out? This is the crux of the Turing test. In this symposium we suggest that identification of similarities and differences between humans and other species provides an opportunity to examine the Turing Test from a different perspective. Our goal is to show that the range of conceptual learning in nonhuman animals includes several of the of the major categories traditionally attributed to humans alone. Understanding concept learning in animals other than humans provides not only a more inclusive view of concept learning, but also provides a more objective perspective from which to understand the processes involved in such learning.

Perceptual Classes

Edward A. Wasserman, Department of Psychology,
The University of Iowa, Iowa City, IA 52242-1407

The most fundamental form of concept learning involves classification according to the perceptual attributes of objects (i.e., the features that they share). There is clear evidence that pigeons can sort complex stimuli into basic classes and that the basis for such sorting is similar to that used by humans (Bhatt, Wasserman, Reynolds, & Knauss, 1988).

Superordinate Classes

Thomas R. Zentall, Department of Psychology,
University of Kentucky, Lexington, KY 40506

At a more advanced level, animals have been shown to be capable of forming "superordinate" classes or functional equivalences. In a matching-to-sample task, pigeons that have learned to assign several arbitrary samples to a common comparison stimulus can be shown to develop emergent relations among those samples; later reassignment of one or more of those samples to a new comparison results in the untrained reassignment of the other members of the superordinate class (Urcuioli, Zentall, Jackson-Smith, & Steirn, 1989; Wasserman, DeVolder, & Coppage, 1992).

Relational Classes

Roger K. R. Thompson, Mary Jo Rattermann, and Anthony P. Chemero, Whitely Psychology Laboratories, Franklin & Marshall College, Lancaster PA 17604-3003

We will present a series of results concerning the cognitive abilities of children, chimpanzees, and monkeys. By combining these results with research on pigeons (discussed by Wasserman and Zentall in this Symposium) and carefully designed simulations, we will demonstrate how comparative methods can be used to identify specialized, if not unique, human cognitive abilities. Specifically, we address the role of symbolic representation and the role of social factors in shaping the expression of abstract relational and analogical cognitive abilities.

References

Bhatt, R. S., Wasserman, E. A., Reynolds, W. F., & Knauss, K. S. (1988). Conceptual behavior in pigeons: Categorization of both familiar and novel examples from fourcalsses of natural and artificial stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 219-234.

Thompson R. K. R. & Oden, D. L. (2000). Categorical perception & conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. *Cognitive Science, 24,* 363-396.

Urcuioli, P. J., Zentall, T. R., Jackson-Smith, P., & Steirn, J. N. (1989). Evidence for common coding in many-to-one matching: Retention, intertrial interference, and transfer. *Journal of Experimental Psychology: Animal Behavior Processes, 15,* 264-273.

Wasserman, E. A., DeVolder, C. L., & Coppage, D. J. (1992). Non-similarity based conceptualization in pigeons via secondary or mediated generalization. *Psychological Science, 6,* 374-379.

The AMBR Model Comparison Project: Round III — Modeling Category Learning

Session Organizers: **Kevin A. Gluck (kevin.gluck@williams.af.mil)**
Air Force Research Laboratory
6030 S. Kent St., Mesa, AZ 85212 USA

Richard W. Pew (pew@bbn.com)
BBN Technologies
10 Moulton St., Cambridge, MA 02138 USA

The goal of the Agent-based Modeling and Behavior Representation (AMBR) Model Comparison Project is to advance the state of the art in cognitive modeling. It is organized as a series of model comparisons, moderated by a team from BBN Technologies. In each comparison, a challenging behavioral phenomenon is chosen for study. Data are collected from humans performing the task. Cognitive models representing different modeling architectures are created, run on the task, and then compared to the collected data. The current effort focuses on models of category learning in a dynamic, dual-task environment. Model comparisons such as this, especially with directly comparable human data are rare. While models of category learning are commonplace, the fact that these are models of integrative performance, not just models of category learning in isolation, makes this set of presentations unique.

Experiment Design and Comparison of Human and Model Data

David Diller (ddiller@bbn.com)
Yvette Tenney (ytenney@bbn.com)
BBN Technologies

This experiment involved a classic concept learning task embedded in an air traffic control situation. Subjects had to learn to make correct decisions to accept or reject altitude change requests, based on three bi-variate properties of the aircraft (percent fuel remaining, aircraft size, and turbulence level). A novel feature of the experiment was the addition of multi-tasking to this concept learning paradigm. In addition to the altitude change requests (the concept learning task), the participant had to hand-off a number of aircraft to adjoining controllers (secondary task).

The design consisted of 9 conditions, defined by 3 category structures and 3 workload levels. The three category structures, borrowed from Shepard, Hovland, and Jenkins (1961), were: single attribute relevant (Type I), a single-attribute rule plus exceptions (Type III), and no rule (Type VI). The three workload levels consisted of 0, 12, or 16 required handoffs, in addition to the 16 altitude requests. It was expected that both category structure and workload level would affect performance. There were 8 scenarios, or trials, lasting ten minutes each. One hour of training on the mechanics of the tasks preceded the trials.

Ninety humans and four different human performance models described in subsequent abstracts were run through the scenarios. The interface, consisting of a radar screen with moving aircraft and action buttons, was designed to accommodate both humans and models. Humans were randomly assigned to one condition (ten per condition). The models were run one or more times in each condition.

All of the modelers were given the human learning data as soon as they were collected, and while the models were still under development. It was expected, therefore, that they would fit the data fairly well. However, a transfer test (for which the modelers were not given the human data in advance) provides an opportunity to test the generalizability of the models predictions.

Results for both humans and models will be presented on the effects of category structure and workload over trials. Human data and model data are available for the following measures: learning curves (probability of error) on the concept learning task, performance errors on the secondary task (missed and incorrect actions), reaction time on both the concept learning and secondary task, self rated workload ratings (collected from models too!), and self- reports on rule discovery and other strategies on the concept task (humans only). This presentation will set the stage for the modelers to describe the mechanisms and assumptions that allow their models to replicate the results.

An EPIC-Soar Model of Concurrent Performance on a Category Learning and a Simplified ATC Task

Ron S. Chong (rchong@gmu.edu)
George Mason University
Robert E. Wray (wray@soartech.com)
Soar Technology, Inc.

During the first phase of the AMBR project, we developed a model of a simplified en-route air traffic control task. That model was built using the EPIC-Soar architecture, an integration of the perceptual and motor systems of the EPIC architecture with Soar, a learning cognitive architecture. The task to be modeled for the current phase of AMBR is the combination of the same ATC task with a new concept acquisition task. Our approach to building the new model has been to reuse, in a modular fashion, previous Soar models for the subtasks. The ATC model is essentially the same as that of the previous AMBR phases. To produce the learning behavior, we have incorporated an existing process model of concept learning called SCA (symbolic concept acquisition). SCA was developed in Soar and has been

successfully used in many Soar applications and models that require concept learning. Results of the model will be presented.

Developing Concept Learning Capabilities in the COGNET/iGEN Integrative Architecture and Associated AMBR ATC Model

Wayne Zachary (wayne_zachary@chiinc.com)
CHI Systems, Inc.

A concept learning mechanism has been added to the COGNET/iGEN modeling and human performance simulation system, and the model developed in Rounds 1 and 2 of the AMBR model competition has been extended to add the learning task. The learning mechanism developed enables learning of the conditions under which a task or goal should or should not be pursued, in addition to the current mechanism of evaluating a predefined boolean expression against contents of declarative memory. Building on the premise that concept learning can be characterized as hypothesis testing, the learning mechanism incorporates a metacognitive strategy for hypothesis generation and an hypothesis selection process based on memory for previous exemplars and their feedback, as well as previous rules or rule parameters tried, moderated by attentional and forgetting processes. The integration of the learning mechanism into the COGNET/iGEN architecture and the extended ATC model results will be presented.

An Activation-based Theory of Categorization

Christian Lebiere (cl+@cmu.edu)
Carnegie Mellon University

We propose a model of category learning implemented in the ACT-R cognitive architecture (Anderson & Lebiere, 1998). ACT-R is a hybrid architecture that combines a symbolic production system with a subsymbolic activation calculus. Our model is directly grounded in the constraints provided by the architecture, especially its declarative memory retrieval mechanism. Generalization to new instances is produced by a similarity-based partial matching mechanism that operates at the subsymbolic level. A number of latency predictions that had previously been explained in terms of a random walk process arise from an aggregate retrieval mechanism called blending. These subsymbolic mechanisms provide many of the advantages of connectionist systems while preserving inspectability at the symbolic level. The category learning model was added in a modular fashion to the existing ATC model from previous AMBR rounds.

Concept Learning: Knowing and Reasoning in the DCOG Architecture

Robert G. Eggleston (robert.eggleston@wpafb.af.mil)
Air Force Research Laboratory
Katherine L. McCreight (kate@n-spaceanalysis.com)
N-Space Analysis

DCOG is an emerging architecture of cognition. It treats cognitive behavior in terms of organized state changes that occur across a set of subsystems. Concept learning may be achieved in the architecture either by emergent knowing, derived from low-level feature-based recognition, or by higher-level reasoning using more abstract feature-derived knowledge that supports hypothesis formation and testing. In this study, a DCOG model was used to perform a complex air traffic control task that contained a concept-learning component. The concept-learning subtask was patterned after the classic Shepard, Hovland, and Jenkins (1961) task but limited to types 1, 3, and 6. Given the structure of these concept types and the balanced exemplar presentation history used in the experiment, both the knowing and reasoning pathways are viable for type 1 concepts; but only the reasoning path is viable for types 3 and 6. Both the knowing and reasoning pathways support individualistic variations or strategies and thus can emulate individual subject differences. In this presentation, we describe the feature-based concept learning infrastructure of DCOG and discuss its performance on the ATC task.

Symposium Discussant

Bradley C. Love (love@psy.utexas.edu)
University of Texas

Human category learning takes many forms, yet research in category learning has focused almost exclusively on one narrowly defined task: classification learning with no secondary task. This focus has allowed researchers to understand and explore their predictions in detail, but only within a circumscribed domain. Unfortunately, theoretical progress (as well as practical application) also demands the testing of boundary conditions. In many cases, we simply do not know how well our theories of learning generalize across task situations and induction tasks. The work presented in this symposium is an important step towards developing more general theories of learning that can make contact with human performance outside the laboratory.

Acknowledgments

AMBR Round III is sponsored by the Air Force Research Laboratory and the Office of Naval Research.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Inquiry, Technology, and Cognition: Theory and Practice

Sarah K. Brem (sarah.brem@asu.edu)
Division of Psychology in Education, Arizona State University
Tempe, AZ 85287-0611 USA

William Sandoval (sandoval@ucla.edu)
Psychological Studies in Education, University of California, Los Angeles
Los Angeles, CA 90095 USA

Eva E. Toth (etoth@wpahs.org)
Center for Genomic Sciences, Allegheny-Singer Research Institute
Pittsburgh, PA 15212 USA

Jennifer Wiley (jwiley@uic.edu)
Department of Psychology, University of Illinois at Chicago
Chicago, IL 60612 USA

Overview

Inquiry is one of the oldest areas of research in cognitive science, and one of the most interdisciplinary, drawing upon social and cognitive psychology, computer science, philosophy, and educational research. It also demonstrates how cognitive science can flourish at the intersection of theory and practice, with findings from one informing, constraining, and validating the other. There are obstacles to fully realizing this integration, however. Differences in population, setting, methodology, and epistemology have resulted in a patchwork of ideas that we have not quilted together into a functional unit.

Looking at this landscape, several questions emerge that reflect the piecemeal nature of this research. We tend to be ambiguous about what it means to conduct an inquiry, and about why a good inquiry is a good inquiry, defining it primarily in terms of the particular task at hand. We do not really know which features of inquiry are specific to a certain environment, goal, or population, and which features are domain-general. Perhaps of particular importance to those of us with educational interests, we are not always in agreement regarding what effect research in inquiry has in establishing standards, curricula, testing, and assessment, influencing what it means to be "rational," "clear-thinking," and "educated."

Our goal is to get at these questions and issues by bringing together multiple threads of research and making a concerted effort to outline areas of consensus and dissent. Limiting ourselves to the subarea of computer-assisted inquiry about scientific matters, each of us will summarize within and across our own programs of research. Together, we cover a variety of methodologies and settings, from experimental psychology in laboratories, to design experiments in classrooms, to ethnography in online communities. We will attempt to synthesize answers to a set of questions inspired by the interplay of theory and practice:

1. **The Nature of Inquiry**. What is inquiry? What does effective inquiry look like, what does it require, and what does it produce?
2. **Technology, Inquiry & Situated Cognition**. How is the form and function of inquiry facilitated and/or impeded by the environment? Which processes are generalizable? Which are embedded in the particulars?
3. **Educational Implications**. What are the implications of theory for educational practice, and vice versa?

Presentations

Technical and social supports for epistemic practices of scientific argumentation

William Sandoval & Kelli Millwood, UCLA
Marie Bienkowski & Valerie Crawford, SRI International

Promoting critical inquiry from Web sources

Jennifer Wiley & Susan R. Goldman, UIC
Arthur C. Graesser, University of Memphis

Tools for representational guidance during classroom scientific inquiry

Eva E. Toth, Allegheny-Singer Research Institute

Alternate forms of inquiry and their implications for theory and practice

Sarah K. Brem, Arizona State University

Acknowledgments

NSF funds SKB (REC-0133446) and JW (REC-0126265). WS is funded by the McDonnell Foundation, and an NSF contract to SRI. EET is funded by McDonnell and the Presidential Technology Initiative. EET would like to thank Daniel Suthers, Arlene Weiner, Alan Lesgold, David Klahr and the Klahr research group.

Symposium: New Models of Connectionist Language Acquisition

Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23173 USA

Brian MacWhinney (macw@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Connectionist modeling of language acquisition has attracted strong research interests in the past decades since Rumelhart & McClelland's (1986) pioneering model of the acquisition of the English past tense. Significant progresses have been made in this domain, as reflected in Elman et al. (1996), MacWhinney (1999), and more recently in Quinlan (2002). In this symposium, we propose to integrate current connectionist developmental research related to language. We present to CogSci 2002 a variety of new models in connectionist language acquisition, including an SRN model of generalization (Elman), a self-organizing model of categorical representation (Li, Farkas, & MacWhinney), and an encoder network of concept acquisition (Shultz).

In the first talk entitled "Going beyond the input: the problem of generalization from sparse data", Elman will discuss the issue of how children form generalizations given the inputs available to them. Although in quantitative terms children hear an enormous amount, that input provides only a very sparse representation of the language. Given the presence of numerous gaps in the input, how is a child to know when a gap is accidental and when a gap is systematic? Several connectionist simulations suggest the kinds of constraints on induction that may explain the patterns of both under- and over-generalization that are observed in children.

In the second talk entitled "The origin of categorical representation of language in the brain", Li, Farkas, and MacWhinney will start by considering the "brain centers" of language, areas in the brain that respond to different linguistic categories (e.g., nouns and verbs). A working hypothesis underlying "brain centers" is that different linguistic categories are subserved by different neural substrates. This study examines the emergence of categorical representation from a developmental connectionist perspective. It argues that localized linguistic representations arise as a function of the brain's organization and reorganization in response to characteristics of the environment in learning and development. A self-organizing neural network is used to explore the high-dimensional space of various linguistic categories, analyzing realistic natural language data. The model effectively captures such differences in language use.

In the third talk entitled "Acquisition of crisp and fuzzy concepts", Shultz will discuss concept acquisition by neural networks. Crisp concepts possess such rigid boundaries that instances of one concept are rarely confused with another concept. In contrast, fuzzy concepts have vague boundaries, leading to frequent misclassifications. The feature values of fuzzy concepts are considered probabilistic in that they occur in instances with less than certain probabilities. This generates well-known typicality and prototype effects, with some instances considered better exemplars of a fuzzy concept than others are. Although neural network models provide an adequate account of fuzzy concepts, they are, according to some, incapable of accounting for the acquisition and representation of crisp concepts as in, e.g., kinship terms. Simulations with encoder networks show that this view is fundamentally incorrect: encoder networks can account for a wide range of phenomena associated with concepts along the crisp-fuzzy continuum. Representational crispness in these networks is affected by isolation of the concept in semantic feature space and dispersion of its examples around a prototype. Fuzzier concepts are characterized by residence in a relatively crowded region of feature space and by relatively widely dispersed examples; crisper concepts are characterized by residence in a relatively isolated region of feature space and by relatively limited dispersion of examples. Moreover, the presence of defining features immunizes these networks against the normal *fuzzying* effects of conceptual crowding and example dispersion. Simulations also revealed the familiar developmental shift from characteristic to defining features.

Li will give an overview at the beginning of the symposium, and MacWhinney will provide an integrative discussion at the end. Each talk is scheduled for 25 minutes, including discussion and questions from the audience.

Acknowledgments

This symposium is supported by CogSci 2002 and the NSF (#BCS-9975249 and #BCS-998009).

References

MacWhinney, B. (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum.

Quinlan, P. (2002). *Connectionist models of development.* Brighton & New York: Psychology Press.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, and the PDP research group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. II). Cambridge, MA: MIT Press.

Publication-based Talks

Coordination of Talk & Action

Richard Alterman (alterman@cs.brandeis.edu), Alex Feinman, Seth Landsman, Josh Introne
Computer Science Department, Center for Complex Systems, Brandeis University
415 South Street, Waltham, 02454 USA

The participants in a joint activity must work hard to maintain coordination. For complicated and/or novel activities, even more talk is needed to proceed. Over time, for recurrent cooperative behaviors, the participants will organize their talk as a means of organizing their actions. For recurrent activities, a sign may be introduced at the scene to fix a recurrent problem of coordination by providing some organizational structure (e.g., a stoplight). We will refer to permanent structure designed and implemented prior to a cooperative activity by a non-participant that mediates and organizes the activity as a *coordinating representation.* The main work of this talk is to explore the ramifications of, and methodology for, introducing coordinating representations into same-time / different-place computer-mediated cooperative activities.

Groupware systems are computer-based systems that support groups of people engaged in a common task (or goal) and that provide an interface to shared environments (Ellis et al, 1991). Groupware facilitates communication, coordination, and collaboration of group effort. Building a groupware system requires a detailed analysis of the work environment in which it will be deployed and extensive work on designing both the interface as it presents itself to the individual user and the mediated interaction among the users.

Within the literature on CSCW, the development of technology that supports online communication has been a core research issue. For synchronous communication, the canonical example is to convert an everyday task of several actors engaged in planning out some kind of activity in front of a whiteboard into a task that could be computer-mediated. Given the shared workspace, two issues of interest are how the participants in such an activity organize their talk, and how they organize their task.

Where a shared virtual whiteboard is an external media that can be used to support all kinds of social interaction, a coordinating representation is one kind of content realized in an external media. It is specifically designed for a particular context and it addresses a problem of coordination that emerges in the performance of a recurring cooperative activity across sessions of cooperation. For some applications, general-purpose tools that support communication and coordination, like the whiteboard, will suffice. But for tasks that occur over extended periods of time, the introduction of a coordinating representation will potentially improve the interaction among the participants during the recurrent problematic areas of joint behavior. This scheme leverages the user s participation to help make the system tailor-made.

Figure 1 shows the basic methodology we propose for tailor-making a groupware system. A groupware system is developed providing general-purpose coordination tools to support the users activity. In some cases there are difficult problems in coordination that confront the users during the normal course of their mediated behaviors and these problem areas are not easily or efficiently resolved using general tools. In these cases a pilot study is performed, and a discourse analysis is made to identify secondary structure developed by the participants to organize their talk so as to organize behavior. Based on this analysis, a second version of the system is constructed that includes coordinating representations to support user activity.

In the first part of the talk I will lay out the cognitive foundation for this approach to building tailor-made groupware systems. In the second part of the talk, I will present a methodology, focusing on issues and methods of discourse analysis. Data and evidence for this talk are drawn from both prior and existing work (and studies) at Brandeis. The examples we draw on come from the data we have collected from experiments we have performed using the VesselWorld system, demonstrated at CSCW 2000.

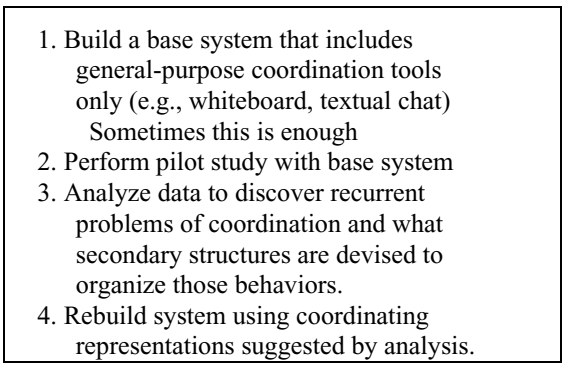


Figure 1: Basic Methodology

Acknowledgments

This work was supported by ONR Contracts N00014-96-1-0440 and N66001-00-1-8965. Additional funding came from NSF grant EIA-0082393.

References

Alterman, R., Feinman, A., Landsman, S. and Introne, J. (2001). Technical Report: CS-01-217, Computer Science, Brandeis University.

Ellis, C., Gibbs, S., and Rein, G. (1991). Groupware: some issues and experiences. *Communications of the ACM*, 34,1, 9-28.

Developing and Validating Cockpit Interventions based on Cognitive Modeling

Deborah A. Boehm-Davis (dbdavis@gmu.edu)
Robert W. Holt (bholt@gmu.edu)
Melanie Diez (mdiez@gmu.edu)
Jeffrey T. Hansberger (jhansber@gmu.edu)
Department of Psychology, George Mason University
4400 University Dr., Fairfax, VA 22030 USA

Aviation accidents are a rare event. However, when they do occur, the cause is attributed to "human error" over 60% of the time (National Transportation Safety Board, 1994). This suggests that the greatest increments in safety can be gained by improving human performance. Indeed, the typical response to an accident investigation is changes to operating procedures that pilots follow in the cockpit. However, in these situations, the changes are made in response to one specific event, which does not allow researchers to pinpoint the more general causes of errors. Further, this approach is not suited to understanding the process of pilot-system interaction that results in the errors. This makes it impossible to know how to design interventions such as training (Boehm-Davis, Holt, Hansberger, & Seamster, 1999), how to redesign instruments, displays, or software, or how to assess the effects of the intervention.

In this research project, we took an alternative approach by developing a computational model of the cognitive processes underlying pilot performance while flying a descent in an automated cockpit. The computational model was built from a cognitive task analysis coupled with empirical performance data. The cognitive task analysis of these phases was developed using NGOMSL (Natural Language GOMS, see Kieras, 1997). This information was combined with eye tracking data taken from pilots interacting with a low-fidelity desktop simulator of a 747-400 aircraft cockpit (Diez et al., 2001) to inform our design decisions about what information pilots are acquiring from the flight deck while working with automated systems. It also formed the basis of a working computational cognitive model, built using the ACT-R cognitive architecture (Anderson & Lebiere, 1998).

The computational model was used to fly the same descent that our pilots had flown on the desktop simulator. Observations of the problems encountered by the model in flying the simulator suggested a number of interventions that might mitigate error in the cockpit. Two of these interventions were selected for empirical testing. First, model runs and eye track data both suggested that the pilots/model were often unaware of changes in automation mode that were driven by the software rather than the pilot (i.e., uncommanded mode changes). A potential intervention developed for this problem is a chime that rings in the cockpit to indicate that the flight management system has autonomously changed the flight mode. We believe that this intervention will draw attention to mode changes that can then be diagnosed and understood.

Second, when the model was interrupted, it often was unable to remember the goal that it was trying to achieve; thus, the model was unable to continue flying. For this problem, new annunciations have been developed for display in the cockpit to capture the goal the automated flight system is trying to achieve. We believe that this goal-oriented display will provide guidance to the pilot about what the flight management system is doing, which can help pilots reconstruct their interrupted goal.

Empirical data collected from commercial pilots using the modified flight management system on the desktop simulator suggests that these interventions will be useful in reducing these specific errors in the cockpit. Further work remains to determine the more general benefits of these interventions.

Acknowledgments

This research was supported by grant NAG 2-1289 from the NASA, and 99-G-010 from the FAA.

References

Anderson, J.R. & Lebiere, C. (1998). *The Atomic components of thought*. Mahwah, NJ: Erlbaum.

Boehm-Davis, D. A., Holt, R. W., and Seamster, T. (2001). Airline resource management programs. In E. Salas, C. A. Bowers, and E. Edens (Eds.), *Improving Teamwork in Organizations: Applications of Resource Management Training*, NJ: Lawrence Erlbaum Associates.

Diez, M., Boehm-Davis, D. A., Holt, R. W., Pinney, M. E., Hansberger, J. T., Schoppek, W. (2001, March 5-8). *Tracking pilot interactions with flight management systems through eye movements*. Paper presented at the 11th International Symposium on Aviation Psychology, Columbus, Ohio.

Kieras, D.E. (1997). A guide to GOMS model usability evaluation using NGOMSL. In M. Helander, T. K. Landauer & P. Prabhu (Eds.), *The handbook of human-computer interaction. (Second Edition).* Amsterdam: North-Holland, 733-766.

NTSB (1994). *Safety Study: A Review of Flightcrew-Involved, Major Accidents of U.S. Air Carriers, 1978 through 1990* (PB94-917001 NTSB.SS-94/01). Washington DC: National Transportation Safety Board.

The Information-Processing Function of Conscious Intentions

Richard A. Carlson (racarlson@psu.edu), Lisa M. Stevenson (lms152@psu.edu), Marios N. Avraamides (marios@psu.edu), and Daniel N. Cassenti (dnc112@psu.edu)
Department of Psychology, Penn State University
613 Moore Building, University Park, PA 16802 USA

We argue that conscious intentions are central to the cognitive control of activity, in contrast to the view that the experience of conscious control is an illusion (Wegner & Wheatley, 1999). We suggest that instantiating a goal to form a conscious intention serves the information-processing function of establishing a procedural frame of reference that organizes mental activity. Information that specifies the origin of this frame of reference simultaneously specifies the conscious agent, the "I" who performs the action. This *cospecification hypothesis* is part of a more general theory of consciousness (Carlson, 1997). We briefly describe this hypothesis and its theoretical basis, and consider several empirical predictions and results bearing on those predictions.

Theory and Hypotheses

The cospecification hypothesis suggests that the content of a conscious intention represents the self as achieving an outcome by performing an operation on an object. For example, a conscious intention to add two digits represents the self as performing a calculation on particular tokens of those digits. Activating this representation serves to initiate a procedure to which the digit tokens are assimilated, and to establish a subjective "point of view" from which the digits are considered. The representation of an outcome that satisfies a conscious intention will thus be structurally very similar to the representation of the intention. This description parallels the representation of goals in ACT-R (Anderson & Lebiere, 1998), in which operands and results complete slots in the goal representation.

Our research has considered several implications of this hypothesis for the information-processing dynamics of goal-driven cognition. First, goal instantiation must precede effective consideration of objects to be processed (operands). Second, the availability of information specifying goals should constrain the temporal coordination of processes such as managing working memory and picking up information from the environment. Third, failures of coordination (e.g., placekeeping errors) should be reduced by activities or information that increase the spatial and temporal precision with which the acting self is specified. Fourth, the need to update the self's spatial and temporal location and orientation should constrain the strategies available for organizing sequential activities; for example, activating an intention directed toward appropriate objects may depend on updating one's perspective on the prior step. Fifth, the construction of explicitly retrievable episodic memories should be associated with goal instantiation because it involves "taking note" of the self as a spatial and temporal marker.

Empirical Results

We have examined each of these implications, using experimental paradigms that examine skilled performance of mental sequences in which the environmental availability of information is constrained. For example, in a number of studies participants solved cascaded, multiple-step arithmetic or spatial path problems in which the outcome of each step served as a starting point for the next step. These studies provide support for the first and second predictions outlined above. Under temporal constraints, individuals can effectively coordinate information pickup and cognitive processes – a process we call *temporal tuning* – only when information specifying upcoming goals is available, allowing those goals to be instantiated as intentions.

In another series of studies, we examined the use of externalizing strategies such as pointing that serve to support temporal coordination. In these experiments, participants performed simple tasks such as counting under varying temporal and strategy constraints. The results suggest that externalizing strategies can serve both to enhance the individuation of objects to be processed (coordination between steps) and to reduce intention-outcome confusions (coordination within steps).

We consider these and other results in relation to the hypotheses sketched above.

Conclusions

In general, these studies provide support for the predictions derived from the cospecification hypothesis. However, some predictions have been disconfirmed in ways that suggest further hypotheses about the constraints on explicit goal instantiation. For example, neither procedural nor explicit declarative knowledge of operator sequences allows the temporal tuning observed when operators are specified by displayed information. We consider the implications of these successes and failures for the general theory of consciousness described in Carlson (1997).

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Carlson, R. A. (1997). *Experienced Cognition*. Mahwah, NJ: Erlbaum.

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54*, 480-492.

Activity Awareness in Computer-Supported Collaborations

John M. Carroll (carroll@cs.vt.edu)
Center for Human-Computer Interaction and Department of Computer Science, Virginia Tech
660 McBryde Hall, Blacksburg, VA 24061-0106 USA

In order to collaborate effectively one needs to know many things about one's collaborators: Who are they? What do they know? What do they expect? What do they want to do? What are they doing now? What tools are they using? To what other resources do they have access? What are they thinking about? What are they planning to do in the near future? What criteria will they use to evaluate joint outcomes?

In ordinary face-to-face communication, people work to establish and maintain a shared background of understanding called common ground (Clark, 1996). Conversational interaction involves continual testing for evidence of common ground, and coordinated effort to enhance common ground. For example, if an interlocutor fails to respond to a request, one might restate presupposed information, point to a relevant object, request acknowledgement, or otherwise remediate. Common ground is unproblematic in face-to-face interactions because such a wide variety of situational elements contribute to it, and the work that people do to maintain common ground is so well integrated into habits and conventions of interaction.

When people work collaboratively, but not face-to-face, many interaction resources are disrupted (Tang, 1991): field of view is reduced, the possibility to use gesture is limited, facial expressions are eliminated or constrained, auditory cues are diminished, tools and artifacts cannot be as easily shared, exchanged information is delayed or decoupled by seconds or even minutes, and collaborators may be in different time zones or different cultures. In remote collaboration it is difficult to convey or discern successful comprehension, current focus of attention, or concomitant attitudes and affect. It is difficult to repair or remediate miscommunications. This transforms the maintenance of common ground into a significant task, which is itself problematic: People are accustomed to taking common ground for granted, as a background task. They do not want to spend attention and effort on it.

These issues have made awareness an increasingly prominent issue in the design of user interfaces for computer-supported collaboration. Investigators have explored numerous user interface tools to help collaborators establish and maintain common ground by supporting their mutual awareness of one another. Prior research has focused on *social awareness* (of the presence of one's collaborators) and *action awareness* (of what collaborators are doing or what they have recently done). Tools investigated include video tunnels (for social awareness) and radar views (for action awareness).

To effectively coordinate complex projects, collaborators need to be aware of one another beyond mere presence and individual actions. We are developing the concept of *activity awareness* (Carroll et al., 2002). Activities are longer term endeavors directed at meaningful goals like "designing the layout of a town park". Longer term activity entails top-down goal decomposition, nonlinear development of partially-ordered plan fragments, interleaving of planning, acting, and evaluation, and opportunistic plan revision. It involves coordinating and carrying out different types of task components, such as assigning roles, making decisions, negotiating, prioritizing, and so forth. These components must be understood and pursued in the context of the overall purpose of a shared activity, the goals and requirements for completing it, and how individual tasks fit into the group's overall plan.

Contemporary user interfaces support collaborative awareness through explicit notifications—requests for chat, email alerts. However, explicit messaging is problematic for supporting activity awareness: Activities are multifaceted and continuing, not simple and ephemeral like presence and action. Thus, notification messages for activity awareness must be meticulous and persistent. But creating such notifications is itself a significant task. It compels a discipline of explicitly externalizing and broadcasting one's goals and plans, something people perceive as both tedious and invasive (Grudin, 1994).

Our work is pursuing the strategy of supporting activity awareness by designing workspace views that embed activity awareness documents. One example is a timeline view of a shared file system that incorporates deadline, active task, and version documents. In such an environment, collaborators would directly, but incidentally, understand the status of a shared activity as they participate in it.

Acknowledgment

Supported by the National Science Foundation IIS 0113264.

References

Carroll, J.M., Neale, D.C., Isenhour, P.L., Rosson, M.B. & McCrickard, D.S. (2002). *Notification and awareness: Synchronizing task-oriented activity.* Center for Human Computer Interaction, Virginia Tech, Blacksburg, VA.

Clark, H. H. (1996). *Using Language*. New York: Cambridge University Press.

Grudin, J. (1994). Groupware and social dynamics: Eight challenges for developers. *Communications of the ACM, 37(1),* 92-105.

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies, 34*, 143-160.

Testing the Roles of Design History and Affordances in the HIPE Theory of Function

Sergio E. Chaigneau (schaigne@uta.cl)
Departamento de Psicologia, Universidad de Tarapaca
Av. General Velasquez 1775, Arica, CHILE

Lawrence W. Barsalou (barsalou@emory.edu)
Department of Psychology
Emory University, Atlanta, GA 30329

Two views currently dominate theories of object function. According to the affordances view, function arises from an object's structure and use the object's design history is relatively unimportant. According to the historical view, function reflects the intention of an object's creator structure and use are relatively unimportant. A new view, the HIPE theory, integrates the affordance and historical views, proposing that function cumulatively requires history, goals, structure, and use to be complete (Barsalou, Sloman, & Chaigneau, in press; also see Chaigneau & Barsalou, in press; Chaigneau, Barsalou, & Zamani, 2002).

Three experiments in Chaigneau (2002) tested the HIPE theory. In each, participants read scenarios that described an artifact's design history and physical structure, along with an agent's goal and actual use. After reading a scenario, participants either rated how appropriate a name was for the object ("mop"), how well the scenario illustrated a category's function (mop), or how likely the scenario was to cause the functional outcome (sopping up spilled liquid). In the baseline scenarios, all four components were intact. In the critical scenarios, one or more components were compromised. Design history could be accidental instead of intentional; the goal to use the object for its function could be absent; the physical structure could be insufficient; the action could be insufficient.

As predicted, Experiment 1 found that compromising each component reduced an object's functionality relative to baseline, consistent with HIPE's prediction that all four components are cumulatively necessary for a complete function. However, compromising structure and use generally produced the largest decrements, consistent with the affordances view. Furthermore, design history was more important for naming than for function and causality judgments, consistent with the causal link between history and naming in historical theories.

Experiment 2 tested the historical view's assumption that design history is causally sufficient for function. If so, then compromising any other component after compromising history should have no effect. Compromising goals, however, produced an additional decrement, consistent with HIPE's cumulative view.

Experiment 3 explored the finding in recent experiments that history is more important than structure and use in naming (e.g., Gelman & Bloom, 2000; Matan & Carey, 2001). In these studies, however, the scenarios lacked sufficient detail about structure and use to derive affordances, thereby leaving history as the most informative factor. When sufficient information was provided so that participants could derive affordances, history became much less important for naming than structure and use.

Overall, these three experiments support three conclusions. First, function is a cumulative construct. Second, affordances are more central to this construct than history, although both are cumulatively important. Third, history is particularly important for naming, and less so for understanding function conceptually and reasoning about it causally.

Acknowledgement

This work was supported by National Science Foundation Grant SBR-9905024 to Lawrence W. Barsalou

References

Barsalou, L.W., Sloman, S.A, & Chaigneau, S.E. (in press). The HIPE theory of function. In L. Carlson & E. van der Zee (Eds.), *Representing functional features for language and space: Insights from perception, categorization and development.* Oxford: Oxford University Press.

Chaigneau, S.E. (2002) *Studies in the conceptual structure of object function*. Doctoral dissertation, Department of Psychology, Emory University, Atlanta, GA.

Chaigneau, S.E., Barsalou, L.W. (in press). The role of function in categorization. *Theoria et Historia Scientiarum*.

Chaigneau, S.E., Barsalou, L.W., & Zamani, M. (2002) Function as a multimodal relational construct. Manuscript in preparation.

Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition, 76*, 91-103.

Matan, A., & Carey, S. (2001). Developmental changes within the core of artifact concepts. *Cognition, 78*, 1-26.

Misrepresenting Emergent Causal Processes as Non-Emergent: A Potential Schema for Overcoming Misunderstandings in Science

Michelene T. H. Chi (chi@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Middle and high school students encounter numerous scientific and "everyday" processes in their curriculum. Some of these processes (e.g., electricity, heat flow, natural selection) seem particularly troublesome for them to learn with deep understanding. One reason for this difficulty is that students often possess alternative conceptions (or misconceptions) that are naïve and scientifically incorrect. These misconceptions are extremely robust and resistant to instruction, therefore preventing students from acquiring the correct understanding.

This paper provides a conceptual analysis that explains why there is a barrier in understanding these processes and what can be done to overcome it. The analysis essentially suggests that these often-misunderstood concepts are bi-level processes in which the global level pattern emerges from the collective individual actions/interactions at the micro level. Thus, the explanatory mechanism that causally relates the micro and the macro levels is an emergent one. Students, however, intuitively misrepresent an *emergent* mechanism as a kind of a non-emergent (or *direct*, for lack of a better term) causal mechanism.

Two types of features of the underlying explanatory mechanisms of *emergent* causal and *direct* causal processes are identified. One type of feature, shown in Table 1, describes the nature of the behavior of the individuals at the micro level. The behavior (i.e., the actions/interactions) of the individuals of an emergent causal process suggest that their actions/interactions must be considered as a *collection*, whereas the behavior of the individuals of a direct causal process suggest that their actions/interactions can be partitioned into distinct *classes*. Thus, this set of six features can serve the purpose of helping students recognize when it is appropriate to consider a set of actions/interactions collectively rather than distinctively.

Table 1: Six features of the actions/interactions (A/I) of individuals in a collection versus classes.

Emergent (Collection)	Non-Emergent (Classes)
● Same kind of A/I	● Different kind of A/I
● Random A/I	● Fixed A/I
● Co-occur or parallel	● Sequential or linear
● Independent A/I	● Dependent A/I
● Uniform status	● Unique or central status
● Ongoing, continuous	● Bounded, terminating

The second set of five features, shown in Table 2, describes the relationship between the micro and the macro levels. These bi-level relational features are the ones that students can appeal to in explaining the causal relationship between the levels. These two sets of features, together, provide a preliminary specification of *emergent* causal and *direct* causal schemas. The claim is that students use their *direct* causal schema to interpret processes with an emergent explanatory mechanism, and therefore misunderstand them.

Table 2: Five features relating the micro individual and macro aggregate level.

Emergent	Non-Emergent
● Indirect	● Direct
● All individuals	● Some of the individuals
● Local & decentralized	● Goal-directed & intentional
● Disjoint	● Corresponding
● Collective summing within each instance of time	● Cumulative summing across time

Several reasons can be postulated for why students commit such misattributions. First, these 11 features, being mutually exclusive, suggest that *emergent* causal and *direct* causal processes may be ontologically distinct; therefore, repairing such misconceptions requires a radical conceptual shift. Second, students may not even realize that they have misrepresented an emergent kind of causal process as a direct kind. Without such awareness, they lack the motivation to seek ways of re-representing emergent processes correctly. Third, students may altogether lack an emergent schema. Without such a schema, students cannot correctly conceptualize an emergent process. Finally, people in general might have a natural predisposition to interpret all events as a direct causal kind.

The implication of this analysis is that teaching students an emergent schema of the underlying explanatory mechanism may allow them to discriminate an emergent kind of causal process from a non-emergent kind, which then may lead to improvements in their understanding of emergent concepts across various disciplines.

Protocol Evidence On Thought Experiments Used By Experts

John J. Clement (clement@srri.umass.edu)
Scientific Reasoning Research Institute
College of Natural Sciences and Mathematics
and School of Education
Lederle GRT 434
University of Massachusetts
Amherst, MA 01003 USA

Despite recent advances, the Fundamental Paradox of Thought Experiments continues to challenge us: How can findings that carry conviction result from a new experiment conducted entirely within the head? The data base for this study comes from ten professors and advanced graduate students in scientific fields who were recorded while thinking aloud about the following spring problem:

> A weight is hung on a spring. The original spring is replaced with a spring made of the same kind of wire, with the same number of coils, but with coils that are twice as wide in diameter. Will the spring stretch from its natural length more, less, or the same amount under the same weight? (Assume the mass of the spring is negligible.) Why do you think so?

Clement (1989) documented analogies, Aha! insights and cyclical model evaluation and revision processes in these protocols. Working from these transcripts, a variety of untested thought experiments (in the *broad sense*) have also been identified, characterized as the act of making a prediction for an untested, concrete, but absent situation (the experiment"). Aspects of the experiment must be new and untested in the sense that the subject is not informed about its behavior. In a case study of one subject, S2, whether the spring wire is bending or twisting eventually becomes a central issue. Textbooks tell us that it is twisting, whereas many subjects assume bending. S2 examined what the effect of twisting would be in the following *Elemental untested thought experiment* used to make a prediction for the base of an analogy to short and long rods being twisted:

> (1) If I have a longer (raises hands apart over table) rod and **I put a twist** on it (**moves right hand as if twisting something**), it seems to me--again, physical intuition--that it will twist more...I think I trust that intuition. I'm **imagining holding something** that has a certain twistyness to it, a-**and twisting it**. Now I'm confirming (**moves right hand close to left hand,**) that.. As (repeats motion) I bring my hand up closer and closer to the original place where I hold it, I realize very clearly that it will get harder and harder to twist.

Bold type above identifies examples of (both kinesthetic and other) imagery-related observation categories: personal action projections, depictive hand motions, and dynamic imagery reports, in that order. None are infallible indicators on their own, but together they are most plausibly explained using a framework that includes flexible perceptual motor schemas that generate and run imagistic simulations, via the extended application of a schema outside of its normal domain, implicit knowledge, or spatial reasoning (Clement, 1994). One can point to such sources as potential origins of conviction in TEs, to help us begin to explain the fundamental paradox. They can also explain the effectiveness of the extreme case at the end of the transcript above as an example of imagery enhancement, a phenomenon difficult to explain in other ways (ibid.).

A second concept of TE in a *narrower sense* that I have found useful is what I call an evaluative Gedanken experiment*:* This is a special kind of untested TE designed or selected by the subject to help evaluate a concept, model or theory. An example is the case of a spring made of a vertically oriented band of material (the reader might imagine the metal unwound from a coffee can, reshaped to make a spring, say, 3 wide.) The subject imagined that such a spring would still be quite stretchable even though it cannot bend in the up-down direction, challenging the necessity of bending as not particularly relevant at all
In this type of evaluatory Gedanken experiment he designs a special case where the bending model yields a prediction, (no stretch) but where he also has some other independent source of information that can evaluate that prediction.

I believe both the broad and narrow concepts of TEs as clarified here are useful, and both can be analyzed in think aloud protocols. The broad concept is appropriate for expressing the fundamental paradox. The narrower concept of an evaluatory Gedanken experiment encompasses some famous TEs in the history of science, impressive in that they can even contribute to eliminating an established theory.

Acknowledgements

The research reported in this study was supported by the National Science foundation under Grant RED-9453084.

References

Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving. Tirosh, D. (Eds.), *Implicit and explicit knowledge*. Norwood, NJ: Ablex Publishing Corp.

Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. Glover, J., Ronning, R., and Reynolds, C. (Eds.), *Handbook of creativity: Assessment, theory and research*. NY: Plenum.

Putting Geometry and Function Together —Towards a Psychologically-Plausible Computational Model for Spatial Language Comprehension

Kenny R. Coventry[1] **(kcoventry@plymouth.ac.uk), Angelo Cangelosi**[2] **(angelo@soc.plymouth.ac.uk), Dan Joyce**[2] **(danj@soc.plymouth.ac.uk) and Lynn V. Richards**[1] **(lynnr@soc.plym.ac.uk)**

[1]Centre for Thinking and Language, Department of Psychology & [2]Centre for Neural and Adaptive Systems, School of Computing, University of Plymouth, Drake Circus, Plymouth PL4 8AA, United Kingdom

Describing the position of objects in space necessitates a mapping between the spatial representation(s), computed by the visual system, and the language processing system. However, it turns out that spatial description is influenced not only by *where* objects are in space, but also by the *functions* that objects afford, and the functional relations between objects. For example, the preposition *at* in *the woman is at her desk* indicates not only that the woman is in close proximity to the desk (a topological-geometric relation), but that she is likely to be working there (an extra-geometric functional relation). Indeed, there is much empirical work showing that meaning of spatial prepositions across a range of languages involves the instantiation of both geometric *and* extra-geometric factors (e.g., Carlson-Radvansky & Radvansky, 1996; Coventry, Prat-Sala & Richards, 2001). However, how geometric and extra-geometric constraints combine is an open question. Regier and Carlson (2001) present a computational account, the attentional vector sum (AVS) model, which grounds the preposition *above* in a mechanism analogous to population vector codes in the neural model of Georgopolous *et al* (1986). However, Regier and Carlson deal only with geometric computations over the visual scene.

We present a new computational model which attempts to deal with the spatial prepositions *in, on, over, under, above* and *below* and extends processing of the visual scene to include functional factors parasitic upon object knowledge. One possibility is that object knowledge can be used as a means of weighting parts of geometric processing, as is suggested by Regier, Carlson and Corrigan (in press). In contrast, Coventry and Garrod (in press) suggest that separate geometric and extra-geometric processes are operational in parallel, and come together in a situation model. Our approach introduces cognitive-functional constraints by extending Ullman s (1984) notion of visual routines to include operations on *dynamic* rather than static visual input (cf. Cavanagh *et al*, 2001). We use neuropsychologically-inspired implementations of connectionist models (cf. Regier, 1996). Based on evidence of motion and spatial-frequency processing in areas V1-V4, the MT, and interactions from regions implicated in object-recognition, such as the IT cortex (Edelman, 1999), we construct a model which might account for extra-geometric *and* geometric factors in one computational system. Developmental accounts of an infant s understanding of concepts such as geometry (spatial relations), dynamics (e.g. gravity, containment and object constancy), and object individuation and identification constrain the training of relevant parts of our model. To give an example, the containing part of a mug is usually taken to be the part the liquid is poured into, and not the semi-circular handle. By watching interactions between mugs and liquids, we induce a dynamic visual routine , and a representation of the object over time. These routines and representations can then be deployed in future processing, for example, to generalize to similar objects *in the absence* of functional interactions. Initial results show that the computational model performs similarly to reference data, obtained from new experimental data on spatial preposition comprehension tasks.

Acknowledgments

This research was funded by EPSRC grant number GR/N38145, awarded to the first two authors.

References

Carlson-Radvansky, L. A. & Radvansky, G. A. (1996). The influence of functional relations on spatial term selection. *Psychological Science, 7(1),* 56-60.

Cavanagh, P., Labianca, A.T., & Thornton, I.M. (2001) Attention-based Visual Routines: Sprites, *Cognition, 80,* 47-60.

Coventry, K. R. & Garrod, S. C. (in press). *Saying, seeing and acting. The psychological semantics of spatial prepositions.* Taylor Francis: Psychology Press.

Coventry, K. R., Prat-Sala, M. & Richards, L. V. (2001). The interplay between geometry and function in the comprehension of over , under , above and below . *Journal of Memory and Language . 44,* 376-398.

Edelman, S. (1999). *Representation and Recognition in Vision*. The MIT Press.

Regier, T. (1996). *The Human Semantic Potential. Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130(2),* 273-298.

Regier, T., Carlson, L. A. & Corrigan, B. (in press). Attention in spatial language: Bridging geometry and function. To appear in L. A. Carlson & E. van der Zee (Eds.), *Functional features in language and space: Insights from perception, categorization and development*. Oxford University Press.

Ullman, S. (1984) Visual Routines, *Cognition*, *18,* 97-159.

A Basis for a Rigorous Cognitive Science: Maintaining Context for Information Exchange between Modules in a Functional Hierarchy

L. Andrew Coward (lacoward@dijkstra.murdoch.edu.au)
School of Information Technology, Murdoch University
South St., Murdoch, WA 6150 Australia

A functional system acts upon itself and an environment in order to achieve internally specified objectives. A system is defined as functionally complex if four conditions are met. The first is that the system can perform many different behaviours. The second is that the system has multiple potentially conflicting objectives. The third is that a high degree of interaction is required between large volumes of information derived from current and past environmental and internal system conditions, from past behaviours, and from system objectives in order to determine appropriate behaviour at any point in time. The fourth is that limited time is available between when an environmental condition occurs and when a behavioural response must be completed.

By this definition, electronic image processing systems are functionally trivial (although difficult to design). The vast majority of information technology applications are very simple functionally. Functionally complex electronic systems are those which control large physical systems with no human intervention except as users of services provided. Such physical systems include aircraft simulators and telecommunications networks. The corresponding electronic control systems have millions of lines of software code designed by hundreds of engineers. Brains are biological examples of functionally complex systems.

Experience with functionally complex electronic systems demonstrates the need for system architectures which make effective use of resources, allow functional changes without side effects, and permit identification and repair of failure conditions. Analogous needs exist for biological brains. These needs force any system which performs a sufficiently complex combination of functions into a modular hierarchy defined by the requirement that modules on each level are roughly equal and information exchange between modules is minimized as far as possible (Coward 2000; 2001). In addition, careful attention must be paid to maintaining the context for such information exchange. Modules within such a hierarchy will not in general correspond with obvious system features, but can be used to relate high level system functions to detailed operations down to the device level.

Two types of information exchange are possible between modules. An unambiguous information exchange indicates that the currently appropriate system behaviour is within a specified set of possible behaviours with 100% confidence, and can therefore be interpreted as an instruction. A partially ambiguous information exchange indicates that appropriate system behaviour is probably within a specified set of possible behaviours, and can therefore be interpreted as a recommendation (Coward 2001).

The requirement to support unambiguous contexts forces a system into the memory, processing form of the von Neumann (or instruction) architecture ubiquitous in functionally complex electronic systems. However, heuristic definition of functionality (i.e. learning) is impractical with unambiguous information exchange.

The requirement to support meaningful although partially ambiguous contexts forces a system into the clustering, competition form called the recommendation architecture. Heuristic definition of functionality is possible, with clustering defining and detecting information combinations in system inputs and competition associating different sets of combinations with different behaviours. However, the requirement to maintain contexts is a severe constraint on both physical form and device algorithms (Coward 2000; 2001).

An implemented electronic version of a system with the recommendation architecture has demonstrated that learning is possible in such a system, and that learning can proceed with minimal effects on prior learning. There are a wide range of similarities between the structure and phenomenology of the implemented system and that of the mammal brain (Coward 2001).

The cortex with columns and areas resembles the clustering subsystem of the recommendation architecture, and the thalamus, basal ganglia and cerebellum resemble the required competition subsystems. REM sleep resembles the required process to ensure global minimization of information exchange. The phenomenologies of implicit and explicit memory resemble the different types of changes to information recording in clustering and competition (Coward 1999). The memory deficits introduced by damage indicate that the hippocampus manages resource assignment within the cortex (Coward 1990).

It is concluded that natural pressures have forced mammal brains into the recommendation architecture form, and that this architectural form can be used to understand the relationships between cognitive processes and physiology.

References

Coward, L. A. (1990). Pattern Thinking, New York: Praeger (Greenwood).

Coward, L.A. (1999). A physiologically based theory of consciousness, in Jordan, S. (Ed.), *Modeling Consciousness Across the Disciplines* (pp. 113-178), Maryland: UPA.

Coward, L.A. (2000). A Functional Architecture Approach to Neural Systems. International Journal of Systems Research and Information Systems, 9, 69 - 120.

Coward, L.A. (2001). The Recommendation Architecture: lessons from the design of large scale electronic systems for cognitive science. Journal of Cognitive Systems Research 2(2), 111-156.

Dynamic Interrelations Among Processing Efficiency, Working Memory, and Problem Solving: A Longitudinal Study

Andreas Demetriou (ademetriou@ucy.ac.cy
Department of Education, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus

A study is presented aiming to contribute to the integration of the information processing, the differential, and the developmental modelling of the mind into a comprehensive theory. This is a longitudinal study which investigated the relations between processing efficiency, working memory and problem solving from the age of 8 to 16 years. This study involved 113 participants, almost equally drawn among 8-, 10-, 12-, and 14-year olds at the first testing: these participants were tested for two more times spaced one year apart. These participants were tested individually with a large array of tasks addressed to processing efficiency (that is, speed of processing and inhibition), working memory (that is, phonological storage, visual storage, and the central executive of working memory), and thinking (that is, quantitative, spatial, and verbal reasoning).

Confirmatory factor analysis validated the presence of all of the above dimensions and indicated that they are organized in a three-stratum hierarchy. The first stratum included all of the individual dimensions mentioned above. These dimensions are organized, at the second stratum, in three constructs, namely processing efficiency, working memory, and problem solving. Finally, all second-order constructs are strongly related to a third-order general factor. This structure is stable in time.

Structural equation modelling indicated that the various dimensions are interrelated in a cascade fashion so that more fundamental dimensions are part of more complex dimensions. That is, speed of processing proved to be the most important aspect of processing efficiency and it is strongly related to the condition of inhibition, indicating that the more efficient one is in stimulus encoding and identification, the more efficient one is in inhibition. In turn, processing efficiency is strongly related to the condition of executive processes in working memory, which, in turn, are related to the condition of the two modality-specific stores (phonological and visual). Finally, thinking was related to both processing efficiency and working memory, the central executive in particular.

These findings provide only partial support to the basic positions in psychometric theorizing concerning the role of the components of processing capacity in the functioning of thinking. Specifically, none of these components alone is the crucial factor in the functioning of thinking. Rather, they additively contribute to its functioning. Moreover, a considerable amount of variance in different domains of thinking remains unexplained by these general factors. Therefore, any theory of intelligence must be able to account for the organization and functioning of these domains.

All dimensions appeared to change systematically with time. Growth modelling suggested that there were significant individual differences in attainment in each of these dimensions. Moreover, development affected differently each of them as well as their interrelation. Mixture growth modelling suggested that there were four types of developing persons, each being defined by a different combination of performance along these dimensions. Some types were more efficient and stable developers than others. These analyses indicated that processing efficiency is a factor that explains developmental differences in problem solving whereas working memory explains individual differences. Modeling by logistic equations uncovered the rates and form of change in the various dimensions and their reciprocal interactions during development. A developmental model is proposed to account for these findings.

References

Demetriou, A., Christou, C., Spanoudis, G., & Platsidou, M. (in press). The development of mental processing: Efficiency, working memory and problem solving. *Monographs of the Society for Research in Child Development*.

Tutoring Real-Time Dynamic Task Performance: Using ADAPT to Augment Pilot Skill Acquisition

Stephanie M. Doane (sdoane@ra.msstate.edu)
Daniel W. Carruth (dwc2@ra.msstate.edu)
Engineering Research Center, Mississippi State University
2 Research Boulevard, Mississippi State, MS 39762 USA

Background

This research examines the role of comprehension-based cognitive processes in the acquisition of skills in real-time dynamic task environments. A theoretically-based model of pilot instrument flight (ADAPT) is used as the student model component (VanLehn, 1988) of an intelligent tool for training real-time complex task performance. ADAPT is a computational model of action-planning with an architecture based upon Kintsch's (1994; 1998) construction-integration theory of comprehension. ADAPT's learning mechanisms are used to model instrument flight skill acquisition (Doane, Sohn, McNamara, & Adams, 2000) and to select instructions intended to optimize pilot performance.

In previous research, rigorous tests of ADAPT's predictive validity compared performance of individual pilots to that of their respective models (Doane & Sohn, 2000; Doane, 2001; Sohn & Doane, 2002). Individual pilots were asked to execute a series of flight maneuvers using a flight simulator, and their eye fixations and control movements were recorded in a time-synched database. Models of the 25 individual pilots were constructed and used to simulate pilot execution of the same flight maneuvers. The time-synched eye fixations and control movements of individual pilots and their respective models were compared. The results suggest that the model explains and predicts a significant portion of pilot visual attention and control movements during flight as a function of piloting expertise.

Current Research

Current research is focused on incorporating ADAPT into a prototype training system that can identify training opportunities. In the training system, human pilots accomplish flight maneuvers using a graphical flight simulator. Their eye movements are tracked by an ASL oculometer and their control movements and flight performance are recorded by flight simulator software. The flight simulator and oculometer data are time-synched, passed to the ADAPT model for analysis, and then ADAPT selects instructions that optimize pilot comprehension of their current task environment.

Of particular interest for cognitive science is how ADAPT uses performance data to make inferences about individual pilot knowledge, skill, and focus of attention, and the ability of the model to run simulations in real-time to predict future pilot actions and to select instructions that optimize future performance. Instructions will be delivered to pilots via agents that use verbal (e.g., voice instructions via earphones) as well as nonverbal forms of communication. The major challenge will be to accomplish this goal without disrupting pilot performance. We will measure subject response to instruction and the time course of their learning and relate these back to the a priori individual differences in pilot performance

This research will make a theoretical contribution to our understanding of the role of comprehension-based cognitive processes in real-time complex and dynamic task performance. It will also make practical contributions to training technologies that can be used to augment the acquisition of complex task performance skills.

Acknowledgments

The authors wish to thank the Office of Naval Research and Dr. Susan Chipman, Program Officer for the Training-Related Science and Technology Program for their support of this research under grant N000140210152.

References

Doane, S. M. (2001). ADAPT: Predicting user action planning. In M. Smith, G.Salvendy, D. Harris, & R. Koubek (Eds.), *Proceedings of the 9th International Conference on Human-Computer Interaction, Vol.1.* (pp. 322-326). New Orleans, LA: Erlbaum.

Doane, S. M., Sohn, Y. W., McNamara, D. S., & Adams, D (2000). Comprehension-based skill acquisition. *Cognitive Science*, *24*, 1-52.

Doane, S. M. & Sohn, Y.W. (2000). ADAPT: A predictive cognitive model of user visual attention and action planning. *User Modeling and User Adapted Interaction, 10*, 1-45.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York. NY: Cambridge University Press.

Kintsch, W. (1994). The psychology of discourse processing. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 721-739). San Diego: Academic Press.

Sohn, Y.W., & Doane, S.M. (2002). Evaluating comprehension-based user models: Predicting individual user planning and action. *User Modeling and User Adapted Interaction, 12(2-3),* 171-205.

VanLehn, K. (1988). Student modeling. In M. C. Polson & J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 55-76). Hillsdale, NJ: Erlbaum.

Implementing Latent Semantic Analysis in Learning Environments with Conversational Agents and Tutorial Dialog

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Xiangen Hu (xhu@memphis.edu)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Brent A. Olde (baolde@memphis.edu)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Matthew Ventura (mventura45@hotmail.com)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Andrew Olney (aolney@hotmail.com)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Max Louwerse (mlouwers@memphis.edu)
Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152 USA

Donald R. Franceschetti (dfrncsch@memphis.edu)
Department of Physics, University of Memphis, CAMPUS BOX 523390, Memphis, TN 38152 USA

Natalie Person (person@rhodes.edu)
Department of Psychology, Rhodes College, 2000 N. Parkway, Memphis, TN 38112 USA

We have been developing learning environments with animated conversational agents. The agents manage a mixed-initiative dialog between the learner and the computer system either by a direct conversational interaction or by serving as a navigational guide on a web site. Two of the systems simulate human tutors by (a) presenting difficult questions that require deep reasoning, (b) attempting to comprehend the learner's typed input, (c) formulating dialog acts that are sensitive to the learner's contributions, and (d) speaking to the student with the animated agent. AutoTutor teaches computer literacy whereas Why/AutoTutor teaches conceptual physics (Graesser, VanLehn, Rose, Jordan, & Harter, 2001). The Human Use Regulatory Affairs Advisor (HURAA) teaches officers in the military about the ethical use of human subjects on a web site with a search facility that accesses documents through questions posed in natural language.

All three systems have used Latent Semantic Analysis (LSA) as its primary representation of world knowledge. LSA is a statistical technique that compresses a large corpus texts into a space of 100-500 dimensions (Landauer, Foltz, & Laham, 1998). The K-dimensional space is used when evaluating the similarity between any two bags of words, with values ranging from 0 to 1. From the standpoint of AutoTutor and Why/AutoTutor, one bag of words is the set of assertions that a student expresses within a dialog turn; the other bag of words is the content of the curriculum script for a particular topic. From the standpoint of HURAA, one bag of words is the learner's query in natural language and the other is a paragraph in the document space. LSA has generally been successful in evaluating the quality of student explanations, in evaluating the quality of student assertions in tutorial dialog, and in the retrieval of documents from natural language queries. Successes and failures of LSA are identified in these three learning environments.

Acknowledgements

This research was supported by the Office of Naval Research (N61339-01-C1006), the Institute for Defense Analyses (AK-2-1801), the National Science Foundation (REC 0106965), and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR, IDA, DoD or NSF.

References

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*, 39-51.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Human-Automation Interaction Strategies

Stephanie Guerlain (guerlain@virginia.edu)
Department of Systems and Information Engineering, University of Virginia
151 Engineers Way, Charlottesville, VA 22904-4747 USA

One approach to limiting the consequences of error-prone human performance is to automate as much as possible in a system. However, accidents and near-misses have occurred when automation fails to perform as desired, and the people "supervising" the automation have trouble determining the state of the automation, the state of the underlying process being controlled, or the implications of how changes to the state or to the automation parameters will affect overall performance. Other classic problems with automation include loss of human skill as tasks become automated and brittleness (the automation works well for the situations for which it is designed but can otherwise give up control or attempt a solution that is completely inappropriate).

Often, much effort goes into the design of computerized algorithms, but relatively little effort is put into the user interface. To design explicitly for *mixed-initiative interaction,* one needs to design a system where both the automation and the human operator have the capability to guide or perhaps even take over control of the system being controlled and that both the human and the automation each has the information and communication means necessary to make his/her, or its own "judgments" about the situation and to guide and perhaps critique the other's behavior. Clearly, due to the well-known differences between information systems' and humans' strengths, weaknesses, and means to: 1) sense information, 2) make judgments, and 3) execute actions, both the types of information required and the means for gathering and communicating that information will necessarily be different for each type of agent. Recent research has suggested certain strategies for safer automation design (assuming humans must monitor or guide the automation's behavior). These are:

1. Interactivity (allow humans to generate alternative automated *and* manual solutions, with the automation providing a comparison across all these solutions) (Guerlain, 2000).
2. Include user-initiated notification and critiquing. User-initiated notification (Guerlain & Bullemer, 1996) allows the human operator to set up temporary, context-sensitive "monitors" and to define who to be notified (person or system) and what to do when such conditions are met. These alerts can be process-specific, temporal, or a combination of the two. User-initiated notification can be turned into a critiquing strategy (Fischer, Lemke, Mastaglio, & Morch, 1990; Guerlain et al., 1999; Silverman, 1992) when these types of context-sensitive alerts are programmed in at design time (e.g., not by the operator, but by the engineer or knowledge expert), and are designed to be more generic and continuous monitors for faulty or important conditions that are in general rare, but would require operator or automation attention.
3. Use appropriate representation aiding and workspace navigation techniques, to minimize errors and difficulties associated with excessive cognitive integration and to maximize effective decision making. The goal of representation aiding is to represent relevant domain, task, and system constraints through visual properties of the display, and thus encourage people to perceive these relationships with little cognitive effort. Workspace management refers to the window manipulation, command input, and navigation activities required when working with computer-based systems(Guerlain, Jamieson, Bullemer, & Blair, 2002).

These techniques have been successfully applied across diverse domains, such as petrochemical, medical, and military. These solution strategies are by no means foolproof, but they are as generic as the problem of how to design for safety when humans and automated agents are involved.

References

Fischer, G., Lemke, A., Mastaglio, T., & Morch, A. (1990). Using critics to empower users. *CHI '90 Human Factors in Computing Systems Conference*, ACM: New York.

Guerlain, S. (2000). Interactive advisory systems. *Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium*, Savannah, GA, 166-171

Guerlain, S., & Bullemer, P. (1996). User-initiated notification: A concept for aiding the monitoring activities of process control operators. *Proceedings of the 1996 Annual Meeting of the Human Factors and Ergonomics Society*, Philadelphia, PA, 283-287

Guerlain, S., Jamieson, G. A., Bullemer, P., & Blair, R. (2002). The MPC Elucidator: A case study in the design for human-automation interaction. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 32*(1), 25-40.

Guerlain, S., Smith, P., Obradovich, J., Rudmann, S., Strohm, P., Smith, J., Svirbely, J., & Sachs, L. (1999). Interactive critiquing as a form of decision support: An empirical evaluation. *Human Factors, 41*(1), 72-89.

Silverman, B. (1992). Survey of expert critiquing systems: Practical and theoretical frontiers. *Communications of the ACM, 35*(No. 4), 107-127.

Statistical Learning, Implicit Memory, and Phonology

Prahlad Gupta (prahlad-gupta@uiowa.edu)
Department of Psychology, University of Iowa
Iowa City, IA 52242 USA

John Lipinski (john-lipinski@uiowa.edu)
Department of Psychology, University of Iowa
Iowa City, IA 52242 USA

This paper argues that (1) implicit memory is based on statistical learning; (2) phonological learning of word forms is based on implicit memory, and therefore that (3) phonological learning of word forms is statistical learning.

Implicit memory as statistical learning

A theoretical analysis of two key implicit memory phenomena, skill learning and repetition priming, shows that a number of apparent dissociations between them are misleading (Gupta & Cohen, 2002). First, it can be shown that the fact that skill learning but not repetition priming follows the power law of practice follows from the mathematical definitions of these constructs, and that this dissociation is therefore artifactual. Second, it can be shown that the presence or absence of correlations between these phenomena is also artifactual, and also follows from their definitions. Behavioral dissociations between these phenomena therefore cannot be regarded as evidence of a processing dissociation between them. Further, a statistical learning based computational model can be shown to account for specific empirical data, exhibiting a classic profile of skill learning and repetition priming, as well as a number of apparent dissociations between these phenomena (Gupta & Cohen, 2002). These theoretical and computational analyses provide complementary evidence that skill learning and repetition priming are aspects of a single underlying mechanism that supports implicit memory. The computational simulations suggest that this mechanism has the characteristics of statistical learning.

Phonological learning as implicit memory

The hypothesis that phonological learning of word forms is based on implicit memory predicts that implicit memory tasks employing nonwords (i.e., novel phonological word forms) should yield a typical profile of skill learning and repetition priming (e.g. Gupta & Dell, 1999). A typical multiple-repetition implicit memory task was devised in which participants were presented with nonwords. Some of the nonwords in each block of stimuli appeared only once during the experiment while other nonwords appeared in every block. Participants were simply required to repeat each stimulus as soon as it was presented. Performance functions were very similar to those in standard implicit memory tasks, exhibiting classic skill learning and repetition priming. These findings suggest that implicit memory plays a role in the learning of phonological forms, which in turn suggests a role for statistical mechanisms in phonological learning.

Further evidence of the role of distributional statistics in phonological learning comes from a second manipulation in the study. If distributional statistics play a role in the learning of phonological word forms, then a word form's frequency-weighted neighborhood density (N) should impact repetition priming. To test this hypothesis, half of the nonword stimuli had a high N and half a low N. Neighborhood density was found to have a significant impact on learning of the nonwords.

Phonological learning as statistical learning

The effects of neighborhood density provide new evidence that phonological learning is affected by the distributional statistics of the environment. The presence of classic skill learning and repetition priming effects in nonword repetition provides complementary evidence regarding the nature of the underlying learning, suggesting it is based on implicit memory. The theoretical and computational analyses suggest that implicit memory is based on statistical learning mechanisms. Together the present results provide new evidence that learning novel phonological forms is based on statistical mechanisms.

Acknowledgments

We wish to thank Rochelle Newman, Kirrie Ballard, Gary Dell, Jean Gordon, and Larissa Samuelson, for helpful discussion of aspects of this work.

References

Gupta, P., & Cohen, N. J. (2002). Theoretical and computational analysis of skill learning, repetition priming, and procedural memory. *Psychological Review*, *109*, 401–448.

Gupta, P., & Dell, G. S. (1999). The emergence of language from serial order and procedural memory. In B. MacWhinney (Ed.), *The emergence of language*, 28th Carnegie Mellon Symposium on Cognition. Hillsdale, NJ: Lawrence Erlbaum.

Mental Visualizations and External Visualizations

Mary Hegarty (hegarty@psych.ucsb.edu)
Department of Psychology, University of California, Santa Barbara
Santa Barbara, CA 93106 USA

Recent advances in computer technology and graphics have made it possible to produce powerful visualizations of scientific phenomena and more abstract information. There is currently much excitement about the power of these computer visualizations in activities such as scientific discovery, search of information spaces, and education (e.g., Card, Mackinlay & Schneiderman, 1999; Gordin & Pea, 1995).

In this presentation I will define a visualization, very broadly, as any visual-spatial display in which information is communicated by the spatial arrangement of elements in the representation. Computer visualizations are often dynamic. However static graphs, diagrams and maps are also examples of visualizations. Visualizations often depict physical phenomena that are spatial in nature, such as the development of a thunderstorm. They can also depict more abstract phenomena, such as the flow of information in a computer program or the organization of information on the world wide web. A visualizations can exist both internally, in the mind of an individual (as a mental image) or as an artifact printed on paper or shown on a computer monitor that can be viewed by an individual.

Whereas our ability to internally visualize has probably not changed significantly in recent history, technology has significantly improved our ability to create external visualizations. It is probably not surprising therefore that current research on the role of visualization in thinking focuses on external visualizations. Cognitive scientists have made important contributions to research on external visualizations (e.g., Larkin & Simon, 1987; Scaife & Rogers, 1996). There has also been an important tradition of research on internal visualization within cognitive psychology (e.g. Kosslyn, 1994). However, psychological research on internal visualization has been more concerned with the nature of the imagery system than in its role in thinking and reasoning

The purpose of this paper is to explore possible relationships between internal and external visualizations and their role in thinking and reasoning. One possibility is that external visualizations can substitute for internal visualizations. This view assumes that one person can create an external visualization of phenomenon, a second person can view that visualization, and as a result the second person will have the same internal represenation as the person who created the visualization. If this were true, an external visualization could act as a "prosthetic" for people with poor spatial visualization abililities. A second possibility is that external visualizations augment internal visualizations, that is, provide information or insights that are additional to those that can be provided by internal visualizations. For example, an external visualization might show a more complex process than can be internally visualized within the limited capacity of visual-spatial working memory. A third possibility is that the ability to internally visualize might be a requirement for comprehending and using external visualizations. In this case, gaining insight from a visualization would depend on the same skills as internally visualizing. A fourth possibility is that viewing external visualizations enhances the development of internal visualization abilities. These possibilities are not mutually exclusive.

This paper will be informed by my research on mental animation of static diagrams (Hegarty, 1992) and on learning from animated displays (Hegarty, Narayanan & Freitas, 2002). It will discuss implications for education and training.

References

Card, S. K., MacKinlay, J. D. & Schneiderman, B. *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann.

Gordin, D. N. & Pea, R. D. (1995) Prospects for scientific visualization as an educational technology. *The Journal of the Learning Sciences, 4,* 249-279.

Hegarty, M. (1992). Mental animation: Inferring motionfrom static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18(5)* 1084-1102.

Hegarty, M. Narayanan, N. H. & Freitas, P. (2002). Understanding Machines from Multimedia and Hypermedia Presentations. In J. Otero, A. C. Graesser & J. Leon (Eds.). *The Psychology of Science Text Comprehension*. Lawrence Erlbaum Associates.

Kosslyn, S. M. (1994). *Image and Brain*. Cambridge, MA: MIT Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science, 11,* 65-99.

Scaife, M. & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies, 45,* 185-213.

Modeling Aviation Crew Interaction Using a Cognitive Architecture

Robert W. Holt (bholt@gmu.edu)
Jeffrey T. Hansberger (jhansber@gmu.edu)
Ronald S. Chong (rchong@gmu.edu)
Deborah A. Boehm-Davis (dbdavis@gmu.edu)
Department of Psychology, George Mason University
4400 University Dr., Fairfax, VA 22030 USA

Holt (2001) proposed developing Scientific Information Systems to construct and validate theory concerning complex multi-person systems. Holt described a process of successive cycles of theory refinement using information in databases. Holt, Boehm-Davis, and Beaubien (2001) discussed the development of theory for describing crew performance in the aviation domain by statistically analyzing performance measures. These inductive, theory-building approaches require good data and analyses. Unfortunately, obtaining good quality measures may be difficult in domains such as aviation which are complex, dynamic, and multi-person (Holt, Johnson, & Goldsmith, 1997; Holt, Hansberger, & Boehm-Davis, in press).

An alternative approach is to carefully extend theory from a field closely related to the focus of research and subsequently validate it. This study was focused on aviation crew performance using flight deck automation during the descent phase of flight. The theory that was extended to this domain was the ACT-R 4.0 cognitive architecture (Anderson & Lebiere, 1998). The ACT-R architecture was extended to describe the highly procedural nature of crew performance in this context (e.g. checklists, Standard Operating Procedure, etc.). The initial development of this model focused on an ACT-R model of the Pilot Flying (PF) who had to receive directives from Air Traffic Control (ATC), decide on how to use the automation to achieve flight goals, and monitor the success or failure of actions.

Based on lessons learned from this initial effort, the approach was extended to constructing a crew model with a simulated PF and Pilot Not Flying (PNF). These crew members were simulated by separate ACT-R models based on a cognitive task analysis of the duties for each person. The simulated task scenario was the time period just before and after Top of Descent (TOD) in the descent phase of flight. The PNF tasks included verification and programming of the Flight Management System (FMS) computer as well as gathering appropriate information for completion of the flight. The PF monitors and flies the aircraft except for required briefings and responses.

Required aspects of crew interaction such as crew communication (e.g. briefings, acknowledgments) were implemented by a communication link between the PF and PNF simulations using a multi-model extension of ACT-R. Simulated communications involved goals, specific actions, or situational facts and features.

The linked PF and PNF models were evaluated by manipulating the simulated expertise of the crew. Expertise was simulated by changing ACT-R parameters and structures. Specifically, higher expertise was simulated by combinations of high strength of associative links for procedural behavior, higher working memory capacity, and cognitive strategies such as the systematic reactivation of goals cued by external stimuli such as a checklist.

One advantage of using the cognitive architecture was that a complete profile of cognition and performance could be measured for each simulation run. Model performance measures include total time for all tasks, average time for each task, checklist steps skipped, repeated, or performed out-of-order, automation programming delayed, skipped, or incorrect, and the omission of required communications.

Qualitative results such as step skipping, repetition, and intrusion of incorrect steps were observed at lower levels of simulated expertise. Emergent results included crew miscommunication, differential situation awareness, and forgetting relevant goals under certain conditions of delays and interruptions. The precise profile of performance differences for different levels of crew expertise can be used to develop assessment items, strategies, and guidelines for assessing performance of commercial crews.

Acknowledgments

This research was supported by grant NAG 2-1289 from the NASA, and 99-G-010 from the FAA.

References

Anderson, J.R. & Lebiere, C. (1998). *The Atomic components of thought*. Mahwah, NJ: Erlbaum.

Holt, R. W. (2001). *Scientific information systems.* Aldershot, UK: Ashgate Publishing.

Holt, R. W., Boehm-Davis, D. A., & Beaubien, J. M. (2001). Evaluating resource management training. In E. Salas, C. A. Bowers, & E. Edens (Eds.), *Improving teamwork in organizations: Applications of resource management training.* Mahwah, NJ: Lawrence Erlbaum Associates.

Holt, R. W., Hansberger, J. T., & Boehm-Davis, D. A. (in press). Improving rater calibration and performance in aviation. *International Journal of Aviation Psychology*.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997). Application of psychometrics to the calibration of air carrier evaluators. *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting,* 916-920.

Promoting Transfer through Case-Based Reasoning: Rituals and Practices in the Learning by Design Classroom and Evidence of Transfer

Janet L. Kolodner (jlk@cc.gatech.edu)
College of Computing, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA 30332-0280 USA

Our goal in education is to help students learn content and skills in ways that allow them to use what they are learning in new and novel situations. We want them to learn for transfer. The cognitive literature focuses on the cognition, or individual reasoning, needed to learn for transfer (see, e.g., Bransford et al., 1999). The socio-cultural literature focuses on the social interactions that are important for learning skills and practices in such a way that they can be performed in new and different contexts (e.g., Lave & Wenger, 1991). Both agree the building blocks for engaging in skills and participating in practices in an expert way are developed over time and in stages and require a variety of experiences and reflection on them. The computational models of cognition that come out of case-based reasoning (CBR; Kolodner, 1993) allows us to redefine transfer as spontaneous reminding and use of previous experience in reasoning about a new situation (Kolodner et al., 2002). This interpretation of transfer suggests practices for the classroom that can promote transfer (Kolodner, 1997).

We ve designed a project-based inquiry approach to science learning for middle school called Learning by Design“ (LBD“; Hmelo et al., 2000; Kolodner et al., 1998, 2002), based on these principles. We've identified many of the affordances and potential affordances for transfer that project and problem-solving activities provide, and we've designed classroom rituals and practices that help teachers and students identify those affordances and act on them. In science education, there is a need for students to learn not only content but also the skills and practices of scientists — from measuring and observing to interpretation of data to justifying with evidence and explaining causally to communicating with others, planning investigative activities, and applying what s been learned. LBD focuses on helping students learn this full set of objectives.

CBR tells us that productive learning from experience requires timely feedback on one s experiences, interpreting that feedback and explaining what happened in light of one s goals and intentions, making connections between one s goals, plans, and explanations, and having the chance to try again. It emphasizes the iterative nature of learning and the centrality of explanation. LBD s activity structures and sequencing provide both affordances and scaffolding for such reasoning. Students learn within the context of design challenges that require iterative trial and refinement for achievement. It is also highly collaborative. They engage in a variety of public presentations (poster sessions, pin-up sessions, and gallery walks) where they present their ideas, interpretations, and experiences to their peers in an interactive forum. Preparing for a session requires making connections between one s goals, plans, and explanations. The public venue allows students to get help from their peers at explaining their results. It also provides students with a variety of examples that are then discussed with lessons that might be learned from the full set extracted. As they iteratively move toward better design solutions, they iteratively enhance their understandings of concepts and their abilities to engage in skills and practices.

Students spontaneously make reference to previous experiences over the course of several months of engaging in LBD activities, especially with respect to carrying out skills and practices. Our performance assessments show spontaneous reminding and use of both knowledge and skills, and LBD students are more capable than comparison students of engaging as scientists and collaborators (Kolodner et al., 2002). We propose that studying learning environments that encourage the natural use of case-based reasoning will increase our understanding of transfer.

Acknowledgments

This work has been supported by the National Science Foundation, the McDonnell Foundation, and the Woodruff Foundation. Many others are involved in this research, including Paul Camp, David Crismond, Barbara Fasse, Jackie Gray, Jennifer Holbrook, Lisa Prince, Mike Ryan, and many teachers. Thanks to them all.

References

Bransford, J. D. et al. (Eds.) (1999). *How people learn.* Washington, D. C.: National Academy Press.

Hmelo, C.E., Holton, D.L. & Kolodner, J.L. (2000). Designing to Learn about Complex Systems. *Journal of the Learning Sciences*, Vol. 9, No. 3.

Kolodner, J.L. (1993). *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufmann.

Kolodner, J.L. (1997). Educational Implications of Analogy: A View from Case-Based Reasoning. *American Psychologist*, Vol. 52, No. 1.

Kolodner, J. L. et al. (1998). Learning by Design from Theory to Practice. *Proceedings of the International Conference of the Learning Sciences (ICLS 98)*. Charlottesville, VA: AACE, pp. 16-22.

Kolodner, J. L., Gray, J. & Fasse, B. (2002). Promoting Transfer through Case-Based Reasoning: Rituals and Practices in Learning by Design“ Classrooms. *Cognitive Science Quarterly, Vol. 1.*

Schank, R. C. (1999). *Dynamic Memory Revisited.* Cambridge University Press: New York.

Dynamic Adaptation to Critical Care Medical Environment: Error Recovery as Cognitive Activity

Tate T. Kubose (kubose@dmi.columbia.edu)
Vimla L. Patel (patel@dmi.columbia.edu)
Desmond Jordan (daj3@columbia.edu)
Department of Medical Informatics, Columbia University
622 W. 168th St. VC-5, New York, NY 10032

Introduction

Early research on errors focused on studies of human reliability in engineering domains. Human components were considered as additional elements in the system, similar to other technical components. Just as technical safety is improved through the reduction of technical breakdowns, it seemed common sense to use a symmetrical rationale to improve safety through the reduction of human errors. In the last few years, Patel and colleagues have reported a number of studies that focused on understanding dynamic decision making in high velocity medical environments, namely intensive care and medical emergency units (Patel and Arocha, 2000; Patel, Kaufman, and Magder, 1996). These and other studies have identified the problems of post-hoc analysis in research on error detection and faults in such environments, that are characterized by high levels of urgency, uncertainty, and shifting, ill-defined, and competing goals. Although such investigations into human error sometimes necessitate a post-hoc analysis, and can be very informative in identification and reduction of future errors, such retrospective analysis presents several problems.

Methodology

The data were collected in the cardiothoracic intensive care unit at a large teaching hospital using naturalistic approaches. The methods used represent an extension of the information-processing, cognitive science tradition. In order to examine the nature of the interactions and negotiations occurring within the workflow of the ICU, we performed approximately four months of ethnographic observation to collect data about the employees and general workflow of the ICU. We then shadowed three nurses for the duration of their 12-hour shifts. During shadowing, we followed the nurse wherever they went, audio-recorded their conversations with other ICU team members, and took notes in a journal to record non-verbal activities. The audio-tapes were transcribed for later analysis.

Results

The data provided us with information about the daily patterns of communication within the ICU and insight into the professional and social relationships between staff members in conducting daily activities. Through protocol analysis of the transcriptions, we were able to characterize specific instances of errors made and circumstances which are highly susceptible to such errors. While many errors occurred, most of them were often detected and resolved very quickly, either through (1) communication between team members or (2) feedback from the ICU environment. For instance, continued use of a sedative to deal with patient pain was quickly rejected as a treatment plan when one team member realized that it was contributing to liver failure. While one ICU staff member alone may have missed this error, with it perhaps leading to an adverse event, the ICU team was able to, through their interactions, identify the potential for error and take steps to prevent it. Although this situation was managed with error recovery, instances such as these can be missed and contribute collectively to more serious errors.

Discussion and Conclusion

We suggest that mistakes are an inevitable, cognitively useful phenomenon that cannot be totally eliminated. We view human errors as products of cognitive activity regulated in a broader context of adaptation to one's environment and work activities. In this view, where human activity is seen as dynamic adaptation to the work environment, most errors can be considered as the price paid for making compromises in trading off between various alternatives. Unlike the popular goal of achieving flawless performance (through development of error-free systems), our study argues for developing systems that are adaptive enough to allow for the specific nature of human errors.

Acknowledgments

This research was funded by US Army Grant #2000-36-U-of-Texas. We would like to extend our warmest gratitude to the nurses and doctors of the ICU, especially to the three nurses who shared an entire day of their lives with us.

References

Patel, V.L., Arocha, J.F. (2000). The nature of constraints on collaborative decision making in health care settings. In E. Salas & G. Klein (Eds.), *Linking Expertise and Naturalistic Decision Making*. Mahwah, NJ: Lawrence Erlbaum Associates: 78-91.

Patel, V.L., Kaufman, D.R., & Magder, S.A. (1996). The acquisition of medical expertise in complex dynamic environments. In K.A. Ericsson (Ed.) *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games.* Hillsdale, NJ: Lawrence Erlbaum Associates: 127-165

Applications of Latent Semantic Analysis

Thomas K Landauer (landauer@psych.colorado.edu)
University of Colorado at Boulder, CB 344 Boulder, CO 80309
and Knowledge Analysis Technologies, LLC

Latent Semantic Analysis (LSA) treats language learning and representation as a problem in mathematical induction. It casts the passages of a large and representative text corpus as a system of simultaneous linear equations in which passage meaning equals the sum of word meanings. Solution by Singular Value Decomposition (SVD) and dimension reduction produces a high-dimensional vector representing the average contribution to passage meanings of every word, and thus of the similarity between any two passages. LSA simulates human language understanding with surprising fidelity. Combining LSA with other statistical language modeling methods increases its practical scope. A variety of tests and applications illustrate its power, limits, and raise interesting theoretical issues.

Examples from Previously Published Results

LSA Improved IR 10-30% by recognizing documents of similar meaning but different words (Dumais, 1991); powered automatically constructed cross-language information retrieval (Landauer and Littman, 1990); mimicked the 10 words/ day vocabulary acquisition rate of children (Landauer & Dumais, 1997), and college student learning of psychology from textbooks (Landauer, Foltz & Laham, 1998) as measured by multiple-choice tests; simulated human categorization and similarity ratings (Laham, 2000), enabled simulations of predication and metaphor. (W. Kintsch, 2001); predicted paragraph comprehension differences caused by variation in S-S coherence; predicted which texts students would learn most from as a function of their prior knowledge (Rehder et al.; Landauer, Foltz & Laham, 1998); and improved summarizing skills by automatic componential feedback (E. Kintsch & Steinhart, 2000).

New Tests, Advances and Applications

LSA now scales to ca. 100 million word corpora by larger computer memory and new algorithms. Systems based on LSA measure the quality of sentences written to contextually define a word, r = .81 with expert ratings; connect by conceptual meaning each of a million paragraphs of an e-library; power collaborative learning environments that automatically alert participants to relevant contributions of others and assess contributions; enhance technical manuals to improve learning and speed performance; from text about tasks, occupational histories, etc., help guide career choice, fill jobs, and assemble optimal teams; combined with other statistical language models, score essays as accurately as expert human readers and provide componential feedback and plagiarism detection.

Some Implications, Limitations, and Issues

Successes to date disprove the poverty of the stimulus argument for lexical meaning and recast the problem of syntax learning, but leave much room for improvement. Size matters. The largest text corpora used in these applications equals one student's reading through high school; spoken language experience is an order of magnitude greater. Semantic atoms are not only single words; idioms need lexicalization. Syntax surely matters; LSA ignores word order. LSA's knowledge resembles intuition; people also use language for logic. Relations to other input matter. Perceptual and intentional experience contribute to meaning representation. (However, whether these bases are essential, more fundamental or involve different representational mechanisms is an open question. LSA represents perceptual phenomena vicariously, e.g. color relations. Demonstrations that people think in other modes, or that LSA does not exhaust linguistic meaning do not question LSA's validity, but call for more modeling, testing, and integration.

Possible Avenues for Research and Resolution

Similar inductive methods have been applied in perception (e.g. by S. Edelman, 1999)., opening a road to integrating language and perception. New models with learning of sentential order based meaning are needed. Simon Dennis's new unpublished model is a serious contender.

Acknowledgments

Funding support from ARI, AFRL, ONR, NSF, DoEd.

References

Landauer, T. K. (in press). On the computational basis of learning and cognition: Arguments from LSA. In B. H. Ross (Ed) *The Psychology of Learning and Motivation.* New York: Academic Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

For all other references cited plus demonstrations and public LSA research tools, see http://LSA.colorado.edu and http://www.Knowledge-Technologies.com/

Modeling the Development of Lexicon with DevLex: A Self-Organizing Neural Network Model of Lexical Acquisition

Ping Li & Igor Farkaš (pli, ifarkas@richmond.edu)
Department of Psychology, University of Richmond, Richmond, VA 23173, USA

It is well known in developmental psycholinguistics that young children produce a significant amount of speech errors at a certain stage, which is often associated with U-shaped learning (Bowerman, 1982). Not only do children show errors in morphological development (e.g., the well-known past tense errors), they also display errors in word naming and in spontaneous word production. Various accounts have been offered for these errors, from the child's confusion of related semantic fields, to the child's semantic reorganization, to the child's inability to retrieve the correct items in lexical memory during production. In this study, we investigate the nature and origin of these errors as a function of the organization and reorganization of a developmental lexicon, using DevLex as the basis of our modeling.

DevLex is a self-organizing neural network model that has the following properties (Farkaš & Li, 2002). The model consists of two self-organizing maps interconnected by associative links. The two maps attempt to capture the organization of word meanings (the semantic map) and word pronunciations (the phonological map). The semantic map is a dynamically growing network that learns word representations derived from word cooccurrence matrices in child-directed speech. It organizes word representations incrementally, adding new units to areas with higher lexical density. Semantic map is connected with phonological map that is pre-trained on the PatPho representations (Li & MacWhinney, 2002) of the 550 toddler word-list from the CDI lexical norms (Dale & Fenson, 1996). At every iteration during simulation, the semantic and phonological representations of words are simultaneously presented to both maps. Through self-organization (using Kohonen's algorithm), the network forms an activity on the phonological map in response to the phonological input, and an activity on the semantic map in response to the semantic input. For every word being presented, the model simultaneously learns associative connections between the two maps through Hebbian learning. DevLex is evaluated with respect to its accuracy in semantic representations and in productions (via associative connections from semantics to phonological output).

Our simulation results indicate that DevLex is able to model and provide insights into a range of phenomena in the early lexical acquisition. In particular, the onset of errors (measured as number of words confused in the network) reflects a number of important factors that govern lexical development. (1) The lexical categories of nouns, verbs, adjectives, and closed-class words have different profiles over the course of lexical growth, and their relative proportion, type frequency, and token frequency all affect the kinds of errors made in the network. This finding confirms Bates et al's (1994) argument on the important role that word category composition plays in lexical acquisition. (2) Number of words confused in the semantic map and in production is directly related to word density, measured as the amount of words mapped onto the nearest neighborhood of the target word. In addition, lexical confusion occurs more often for nouns than for other word categories, because of a "noun-bias" in the early vocabulary. This result shows that the source of children's naming errors may be the tight competition among similar neighbors in densely populated regions of the lexicon, consistent with the view that word density can predict the speed and accuracy in children's lexical access (Charles-Luce & Luce, 1990). (3) The rate of vocabulary expansion influences the rate of lexical confusion: the more related words that the network has to learn within a given period, the more likely it will show inaccuracies in the semantic map and in production. This pattern is consistent with the hypothesis that rapid increase in the rate of new words predicts the increase in children's naming errors (Gershkoff-Stowe & Smith, 1997). In sum, DevLex provides a new connectionist model that can simulate a developmental lexicon and relate to realistic language learning with self-organizing principles.

Acknowledgments

This research is supported by NSF (grant #BCS-9975249 awarded to P.L.).

References

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman (Eds.), *Language Acquisition: The State of the Art.* Cambridge: Cambridge University Press.

Farkaš, I., & Li, P. (2002). Modeling the development of lexicon with a growing self-organizing map. In H.J. Caulfield et al. (Ed.), *Proceedings of the 6th Joint Conference on Information Sciences* (pp. 553-556), JCIS/Association for Intelligent Machinery, Inc.

Li, P., & McWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments and Computers*. (in press)

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments and Computers, 28*, 125-127.

Bates, E. et al. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21*, 85-123.

Charles-Luce, J., & Luce, P.A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language, 17*, 205-215.

Gerskoff-Stowe, L., & Smith, L.B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology, 34*, 37-71.

Where Do Problem-Solving Strategies Come From?

Marsha C. Lovett (Lovett@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Research has shown that people often generate problem-solving strategies in a new domain by processing example solutions. However, this approach presumes the existence of some related strategies for processing examples in that new domain. The question then becomes: Where do *those* strategies come from? An important aspect of processing examples is knowing which features of the example problems are structurally significant and which features are superficial. Indeed, much research on expert-novice differences highlights experts' great advantage in properly representing and categorizing problems in terms of deep features (e.g., see Chi, Glaser, & Farr, 1988). But how can problem solvers new to a domain achieve this skill? In many domains, solvers have preconceptions of which features are significant and which are superficial. When these preconceptions are on target, i.e., the presumed-significant features are indeed relevant to the solutions and the presumed-superficial features are irrelevant, then learning by example can proceed effectively and efficiently. However, when solvers' preconceptions do not match reality for a given domain — either because their preconceptions mis-map features to the significant-superficial distinction or because their preconceptions are too weak to enable encoding of the relevant features — then learning is impeded.

Background Research

In previous work, Lovett and Schunn (1999) demonstrated that the same task with different superficial features could lead participants to generate very different strategies and, depending on their individual strategies, achieve very different learning gains. Specifically, in one version of the task, participants tended to encode their choices in terms of a single feature — whether each choice had the same color as the preceding stimulus — whereas in the other version of the task, participants were not biased to any particular feature. The experiment was then designed so both task versions would be best solved using a common, structurally important feature, and this was *not* the same-color feature salient in version 1. As predicted, performance was degraded in version 1. This was attributed to the difficulty — for participants in verson 1 — of learning to ignore a preconceived-relevant feature and having to generate new strategies that did not use this feature. Lovett and Schunn presented a process model, called ReCyCLe, of how features enter and leave one's task representation and, hence, how strategy sets evolve. One prediction of theReCyCLe model is that solvers are more likely to change their representation when their current stsraties' success rates are low.

Goals and Method

The current studies attempt to replicate this previous work under slightly different conditions and to address two additional questions:

(1) What, if any, are the local triggers for problem solvers to change their representations and generate new strategies?
(2) What is the role of explicit instruction (e.g., hints) in helping solvers to adjust their strategy sets?

Question (1) was addressed by asking a subset of participants to provide talk-aloud protocols and comparing coded occurrences of strategy-generation or strategy-change events to similar profiles among the non-protocol participants. Question (2) was addressed by manipulating if, when, and how participants would receive a textual hint about important features to include/exclude in their representation of the task. Also varied in this experiment was the degree of success of the best strategy. In particular, the best strategy's success rate could be increased/decreased by decreasing/increasing the overall randomness of the task environment. This manipulation is, simply construed, a task difficulty manipulation.

Results

Regarding the first question, results suggest that, at least when they are asked to talk aloud, participant engage in a considerable amount of explicit strategy (or hypothesis) generation. And, while participants tend to launch anew strategy immediately following a problem-solving failure rather than success, this is not always the case. Regarding the second question, results suggest that, when solvers mismap features (i.e., when they consider the superficial features in a domain to be significant and vice versa), an explicit hint can help problem solvers more quickly incorporate the structurally important task features in their representations and strategies. Even with such hints, however, performance is still aided by further problem-solving practice. The instructional implications of these results will be discussed.

References

Chi, M. T. H., Glaser, R., & Farr, M. J., (1988). *The nature of expertise.* Hillsdale, NJ:Erlbaum.

Lovett, M. C., & Schunn, C. D. (1999). Task representations, strategy variability, and base-rate neglect. *Journal of Experimental Psychology: General, 128,* 107-130.

Is There a Decision Bias For Information From Internally Consistent Sources?

Shenghua Luan (herome@ufl.edu)
Department of Psychology, University of Florida
P.O. Box 112250, Gainesville, FL, 32611 USA

Robert D. Sorkin (sorkin@ufl.edu)
Departments of Psychology and Industrial & Systems Engineering, University of Florida
P.O. Box 112250, Gainesville, FL, 32611 USA

Jesse Itzkowitz (itz@ufl.edu)
Department of Psychology, University of Florida
P.O. Box 112250, Gainesville, FL, 32611 USA

A person faced with a decision often obtains opinions from other sources. These information sources may be composed of several individual sub-sources. The sub-sources may be partially correlated and may differ in their level of expertise.

This study asked how decision makers weigh the estimates received from different sources when those sources varied in their internal consistency and individual expertise. We paid people to perform a graphical decision task while aided by simulated information sources. Each participant observed a graphical display of a signal-plus-noise or noise-alone event and made an estimate of signal likelihood. The participant then was shown likelihood estimates generated from two simulated information sources. The participant then made a yes-no decision about the occurrence of signal on that trial. A monetary payoff was contingent on the accuracy of this yes-no decision.

The estimates from each information source consisted of likelihood ratings generated by four sub-sources. Thus, on each trial the participant was shown 8 likelihood estimates to aid in her decision, four estimates from information source "A" and four estimates from information source "B". In order to estimate the decision weight that the participant gave to each source, we constructed a multiple linear regression model that related the participant's initial estimate and each source's average estimate, to the participant's final decision.

In different conditions of the experiment, we manipulated the overall information value of a source and the level of expertise and pair-wise correlation among a source's sub-sources. Source expertise was manipulated using the following formula adapted from Sorkin and Dai (1994):

$$d'_{source} = \left[\frac{m s^2_{d'}}{1-r} + \frac{m(m_{d'})^2}{1+r(m-1)} \right]^{1/2}$$

where $d\ell_{source}$ is the detection index (aggregate expertise) of the source, m is the number of sub-sources in the group, r is the correlation among the sub-source estimates, $s^2_{d'}$ is the variance of the sub-sources' expertise and $m_{d'}$ is the average detection ability of those sub-sources.

For example, one condition tested which of two equal-information sources (i.e., two sources that have the same overall detection ability, *d'*) would be given the higher weight: the one whose four sub-sources had partial pair-wise correlations and high sub-source *d'*s, or the one whose four sub-sources had zero pair-wise correlation and lower sub-source *d'*s. The results indicated that participants gave a significantly higher weight to the information source that had the higher consistency and higher component expertise, even though the information available from the two sources was identical. This bias was mainly evident on trials when the aggregate opinions of the two sources disagreed. Other conditions compared performance with sources that had different overall information values as well as different levels of sub-source expertise. In these conditions, the participants tended to overweigh the information from the sources having the higher information value and higher level of sub-source expertise. These biases reflect the participants' sensitivity to across- and within-trial differences in the accuracy and internal consistency of information sources.

Acknowledgments

These experiments were conducted as part of the requirements for the M.S. degree for the first author. This research was partially supported by a grant from the Air Force Office of Scientific Research to the second author.

Reference

Sorkin, R. & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes*, **60,** 1-13.

Understanding and Scaffolding Constructive Collaboration

Naomi Miyake and Hajime Shirouzu
({nmiyake, shirouzu}@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodate, Kaizucho, Toyota, 470-0393 JAPAN

Collaborative situations have started to serve as promising knowledge-building environments. Cognitive science should provide theoretical bases for them, by explaining mechanisms of how collaboration leads the participants to deeper, more conceptual understanding.

Findings of empirical studies

Our previous studies (e.g. Miyake, 1986) indicate that during constructive interaction, 1) each participant's problem interpretation and solution paths are based on each individual's prior knowledge, 2) collaboration provides different perspectives, through both others' comments and self-criticisms, and 3) the monitor views the task-doing situation from a slightly abstracted plane, which contributes to the accumulation of different perspectives.

To expand these, we have recently further analyzed the collaborative process of solving a simple fraction problem (Shirouzu, et al, 2002). The task was to get two-thirds of three-fourths on a square sheet of origami. More than ninety percent of time the subjects, both solos or pairs, either folded or marked the paper to solve it but did not calculate. When asked to solve the subsequent problem with the order of fractions reversed, the solos kept the same strategy to solve.

However, more than sixty percent of the pairs shifted to the arithmetic calculation in their second trial. The shift was a gradual one, involving three re-interpretations. Figure 1 schematically shows the shift, from the left to right. The most externally oriented, two-step strategy requires first folding the paper into four. Upon doing this, one could either start making two-thirds, or re-interpret the just-completed three-fourths as already having three equal-size rectangles, which eliminates the physical necessity of second folding. Similarly, one could re-interpret the two-thirds of the designated three-fourths as two-fourths of the original square. Re-interpreting this as one-half often led our pairs to realize that the problem was soluble by calculation.

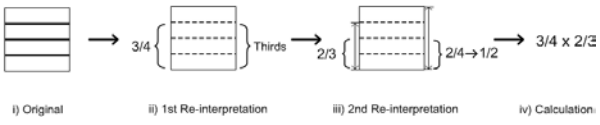


Fig. 1 Gradual re-interpretation of getting 2/3 of 3/4.

Furthermore, it was the monitor who re-interpreted, for 100% of time from level 1 to 2, and 60% from level 2 to 3.

Supporting constructive collaboration

Based on these observations, we propose that the basic components of constructive collaboration are the externalized traces of the task-doer's cognitive workings and their re-interpretations by the monitor, in verbal forms. In verbalizing the re-interpretation, the monitor reclaims the task-doing, on traces of which the previous task-doer can monitor to produce yet another level of abstracted verbalization. This iterates and produces a variety of re-interpretations, which each individual can use to restructure their own internal knowledge. The same picture should apply to inter-group collaboration, where the re-interpretation could be accumulated and become a resource for further abstraction in the entire community.

Summarizing the above, for a collaborative situation to work as a learning-enhancing environment, it is desirable that 1) it assures each individual's conceptual foundation, 2) it entertains the role exchange between task-doing and monitoring to help produce different solutions and their re-interpretations, all slightly more abstract than the previous ones, in roughly the order of their abstraction, 3) which in turn helps each participant to gain an abstract perspective. The social aspect of collaboration appears to motivate the integration of such a variety of solutions and interpretations.

Technological augmentation

Technology can augment implementation of these conditions by providing support to enhance externalization and re-interpretation. To take a simple example, the process of reading can be made visible by having subjects place cards with sentences onto a two-dimensional space. Video-recording of first stages of learning complex skills can be cut into segments and commented on to identify necessary steps. Notes can be shared, and the memos and linkages among them can be stored in the chronological order of their production, so that the production process itself can become a target for later scrutiny. We have been developing and testing such systems, and the context to use them, in a university setting to teach cognitive science to undergraduates (Miyake, et al., 2000; 2001). Use of such systems inevitably changes the way we teach and the students learn, requiring new methods to assess the effects and providing us with a new resource for further research on real-world understanding.

Acknowledgments

Supports are from CREST2000 by JSTC, JSPS Grants No. 265 and 09680380 to the first author and JSPS Fellowship to the second author.

References

Miyake, N. (1986) Constructive interaction and the iterative processes of understanding, *Cognitive Science*, *10* (2), pp.151-177.

Miyake, N, &, Masukawa, H., (2000). Relation-making to sense-making: Supporting college students' constructive understanding with an enriched collaborative note-sharing system. *ICLS 2000*, pp. 41-47.

Miyake, N., Masukawa, H., & Shirouzu, H., (2001). "The complex jigsaw as an enhancer of collaborative knowledge building in undergraduate introductory cognitive science courses, Euro-CSCL2001, pp.454-461.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002) Cognitively active externalization for situated reflection. *Cognitive Science, 26* (4).

Learning from Worked-Out Examples via Self-Explanations: How it Can(not) be Fostered

Alexander Renkl (renkl@psychologie.uni-freiburg.de)
Department of Psychology, Educational Psychology, University of Freiburg
Engelbergerstr. 41, D-79085 Freiburg, Germany

Learning from worked-out examples is of major importance for initial skill acquisition in well-structured domains such as mathematics and physics. However, only those learners who actively explain the rationale of the solution steps presented in the examples to themselves profit from this learning method ("self-explanation effect", Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Unfortunately, most learners are to be characterized as passive or superficial self-explainers (Renkl, 1997a). From an educational perspective, two main questions arise: (1) How can productive self-explanations be fostered? (2) When should instruction move from the self-explanation of worked-out steps to actually solving problems for heightened speed and skill accuracy? Both of these questions were addressed in a series of experiments.

With regard to fostering self-explanations, four issues were investigated: (a) Setting situational incentives: The main idea was that if most learners do not spontaneously generate elaborated self-explanations, it might be helpful to put them into the role of a tutor for another learner. This should motivate them to increase their explanation activities (e.g., Renkl, 1997b). (b) Training and prompting: Self-explanation activities can be induced by a training or by prompting self-explanations at worked-out steps (Renkl & Atkinson, in press; Renkl, A., Stark, R., Gruber, H., & Mandl, H., 1998). (c) Support by instructional explanations: A number of studies have shown that it is difficult to provide effective instructional explanations during example study. However, on the other hand, relying only on self-explanations also has some restrictions (e.g., proneness to errors). Therefore, a set of principles highlighting how to effectively support example study by instructional explanations was developed and empirically investigated (Renkl, in press). (d) Structuring learning materials: Learning materials (i.e., examples and problems) can be designed in order to induce active and well-focused self-explanations, for example, by giving the learner the opportunity of problem-solving experiences before example study (Stark, Gruber, Renkl & Mandl, 2000). The main findings of a series of experiments on these issues can be summarized as follows. The setting of situational incentives has shown not to be very promising. The training and prompting as well as designing learning materials can substantially foster self-explanations and, thereby, learning outcomes. In addition, well-designed instructional explanations can further enhance learning.

For structuring the transition from example study in early phases of skill acquisition to problem solving in a later stage, we developed a fading rationale by which problem-solving elements are successively integrated into example study until the learners are able to solve problems on their own. The effectiveness of fading has been shown in several experiments (Renkl, Atkinson, Staley, & Maier, in press). Presently, we adapt the fading procedure to the learners' prior knowledge level.

Based on the results of our experimental research program, an instructional model of acquiring skills from examples and problems is proposed. It is argued that different learning goals are to be achieved in subsequent stages of skill acquisition (e.g., understanding vs. automation). Therefore, instruction should induce different learning activities during the course of skill acquisition. How these activities can be instructionally fostered can be derived from our experimental findings.

Acknowledgments

This research was funded by the *Deutsche Forschungs-gemeinschaft* (Re 1040/1-1, 1-2, 4-1, 9-1, Ma 978/5-2).

References

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Renkl, A. (1997a). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1-29.

Renkl, A. (1997b). Lernen durch Lehren [Learning by teaching]. Wiesbaden, Germany: DUV.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*, 90-108.

Renkl, A. (in press). Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction*.

Renkl, A. & Atkinson, R. K. (in press). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist.*

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (in press). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education.*

Stark, R., Gruber, H., Renkl, A., & Mandl, H. (2000). Instruktionale Effekte einer kombinierten Lernmethode: Zahlt sich die Kombination von Lösungsbeispielen und Problemlöseaufgaben aus? [Instructional effects of a combined learning method: Does the combination of examples and problems pay off?] *Zeitschrift für Pädagogische Psychologie, 14*, 206-218.

Category Use: Learning and Understanding Categories

Brian H. Ross (bross@s.psych.uiuc.edu) and Seth Chin-Parker (chinpark@s.psych.uiuc.edu)
Beckman Institute and Department of Psychology, University of Illinois
405 N. Mathews Ave., Urbana, IL 61801 USA

Categories are crucial for a large number of cognitive activities, such as classification, inference, problem solving, and explanation. They provide an important means for allowing us to benefit from past experiences. Because of this importance and involvement across a wide variety of intelligent activities, category learning has long been a central research topic in cognitive science, cognitive psychology, and machine learning.

Most of the research on category learning has focused on classification learning, how to assign items to categories. Although classification is an important part of category learning, it is clearly not the only part. In addition, this near-exclusive focus may be limiting our understanding in at least three ways. First, we learn categories in many different ways and how we go about learning categories is likely to have a large influence on what we learn. Thus, a full understanding of category learning requires examining multiple ways of category learning. Second, the focus on classification has led to finding a strong influence of feature diagnosticity, those features that distinguish the categories being learned. Although diagnosticity is an important influence on category representation, we clearly know much more about categories than what distinguishes them. However, given that many items consist of a large number of features and relations that might not be very diagnostic of the category, how do we determine what information to include or not to include in a category representation? Third, in many cases our knowledge of categories does not rely solely on observable features and relations, but on deeper underlying similarities of why the category members go together. It is not clear how classification learning promotes the learning of this type of category understanding.

Recently, there has been a variety of research examining the different ways people learn and use categories (for reviews see Markman & Ross, 2002; Solomon, Medin, & Lynch, 1999), which addresses these limitations of the focus on classification. First, studies have investigated how different ways of category learning might influence the representation (e.g., Anderson, Ross, & Chin-Parker, 2002; Yamauchi & Markman, 1998, 2000). Second, work has examined how nondiagnostic information relevant to other uses of categories might be learned when the focus is not on classification (Chin-Parker & Ross, 2002a, b; Ross, 1997, 1999). Third, research has begun to investigate the understanding that derives from using categories in different ways. Some of my work has examined category learning during problem solving with three different types of tasks—decoding formulas applied to coded messages, mathematical equations, and letter-string transformations (e.g., Ross, 1997, 1999; Ross & Warren, 2002). The results suggest that not only can such learning lead to additional (nondiagnostic) information in the category representation, but it also allows the learning of abstract relations that may help learners to understand the underlying coherence among category members. For example, in the decoding task, learners are able to classify later coded messages on the number relations learned during decoding, even when the relations are fairly abstract (such as the difference between two numbers being less than zero).

Acknowledgments

This research was supported by the National Science Foundation, Grant NSF SBR 97-20304.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30*, 119-128.

Chin-Parker, S., & Ross, B. H. (2002a). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition, 30*.

Chin-Parker, S., & Ross, B. H. (2002b). Diagnosticity in category learning by classification and by inference. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.*

Markman, A. B., & Ross, B. H. (2002). Category use and category learning. *Manuscript under review.*

Ross, B. H. (1997). The use of categories affects classification. *Journal of Memory and Language, 37*, 240-267.

Ross, B. H. (1999). Post-classification category use: The effects of learning to use categories after learning to classify. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 743-757.

Ross, B. H., & Warren, J. L. (2002). Learning abstract relations from using categories. *Memory & Cognition*.

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences, 3*, 99-105.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 776-795.

Relating Properties of Human Memory to Cortico-Hippocampal Architecture

Lokendra Shastri (shastri@icsi.berkeley.edu)
International Computer Science Institute
1947 Center Street, #600; Berkeley, CA 94704 USA

Introduction

Episodic memory (Tulving, 1995) refers to our ability to remember events and situations in our daily lives and acquire memories of specific events by reading a newspaper or watching a newscast. There is a broad consensus that the hippocampal system (HS) consisting of the hippocampal formation and neighboring cortical areas in the medial temporal lobe plays a critical role in the encoding and retrieval of episodic memories (Squire, 1992; Cohen & Eichenbaum, 1995; Nadel & Moscovitch, 1997). But how the HS subserves this mnemonic function in concert with cortical circuits is not fully understood.

Although a number of computational models have been proposed to explain how the HS might support episodic memory function, several key representational problems have remained unsolved. In particular, most existing models view an item in episodic memory as a *feature vector* or as a *conjunctive code* that binds together the components of memory, but as argued in (Shastri, 2002; 2001), such a view is inadequate for encoding events and situations.

SMRITI

SMRITI (**S**ystem for **m**emorizing **r**elational **i**nstances from **t**ransient **i**mpulses) is a computational model of episodic memory that demonstrates how a cortically expressed transient pattern of activity representing an episode can be transformed rapidly into a persistent and robust memory trace in the HS as a result of long-term potentiation (Shastri, 2001; 2002).

SMRITI explicates the representational requirements of encoding events and situations, proposes a detailed neural circuit that satisfies these requirements, and demonstrates that the propagation of a suitable pattern of activity encoding an event can lead to the rapid and automatic formation of the requisite neural circuit within the HS.

The neural circuit required for encoding an episodic memory trace is fairly complex and idiosyncratic, but SMRITI shows that this complexity and idiosyncrasy is well matched by the complexity and idiosyncrasy of the architecture and local circuitry of the HS.

Predictions and Explanations

SMRITI predicts (i) the functional role of each HS component and some of the cortical areas interacting with the HS, (ii) the properties of cortically expressed event schemas underlying episodic memories, (iii) the sorts of memories that must persist in the HS for the long-term, (iv) the nature of memory consolidation, and (v) memory deficits that would result from cell loss in different HS regions and cortical circuits encoding semantic knowledge.

SMRITI also offers biologically grounded explanations of behavioral findings about human memory such as the fan-effect (Anderson, 1974) and the list-strength effect (Ratcliff, Clark, & Shiffrin, 1990). It is significant that no attempt was made to model these behavioral findings; the explanations for these phenomena arise directly from the biologically grounded architecture and structure of the model.

SMRITI makes specific behavioral predictions about the time required for retrieving memorized facts. For example, it predicts that the time to retrieve a fact, wherein an entity fills a given role, is affected primarily by the total number of facts memorized in which the entity plays the same role, and not by the total number of facts memorized about that entity. Thus SMRITI suggests a modified form of the fan-effect. SMRITI also predicts that retrieval times of facts pertaining to populated event schemas are qualitatively different from those of facts pertaining to unpopulated ones. Here, an event schema is heavily (lightly) populated if many (only a few) instances of the schema have been memorized.

This talk will present an overview of SMRITI, and discuss some of its key properties and predictions.

Acknowledgments

This work is funded by NSF grants 9720398 and 9970890.

References

Anderson, J. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6, 451-474.*

Cohen, N.J., & Eichenbaum, H. (1995). *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT Press.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, 7, 217-227.

Ratcliff, R., Clark, S. and Shiffrin, R. (1990) The list-strength effect: I. Data and Discussion. *Journal of Experimental Psychology: LMC:* 16:163-178.

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Science, 6,* 162-168.

Shastri, L. (2001) *Episodic memory trace formation in the hippocampal system: a model of cortico-hippocampal interactions* (Tech. Rep. TR-01-004). Berkeley, CA: International Computer Science Institute.

Squire, L.R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195-231.

Tulving, E. (1995) Organization of Memory: Quo Vadis? In M.S. Gazzaniga (Ed.), *The Cognitive Neuroscience*. Cambridge, MA: MIT Press.

What Happened to the Imagery Debate?

Peter P. Slezak (p.slezak@unsw.edu.au)
Program in Cognitive Science, University of New South Wales
Sydney NSW 2052, AUSTRALIA

Déjà Vu All Over Again?

Zenon Pylyshyn's (2002) recent return to the fray means that at least one thing may be said with certainty about the imagery debate: Despite Kosslyn's (1994) claim to have resolved the controversy, there has been no progress at all. Worse still, if Pylyshyn's null hypothesis is right, we don't have a viable theory of imagery of any kind. The 'tacit knowledge' rival to pictorialism, is not itself an alternative theory but rather an indication of the direction in which an adequate theory might be sought - that is, as a theory of high-level belief or knowledge representation.

Pylyshyn's central criticism of pictorial theories echoes Descartes (1637), who insisted that it is enough that the mind adequately *represent* the properties of the world and does not have to *share* them. In the same vein, S. Edelman (1998), recently says nobody thinks that a mental representation of a cat is furry. Perhaps not, but it is telling that such views must be repeatedly refuted throughout the history of speculation about the mind.

For some reason the case for spatial properties has seemed much more persuasive than the same point regarding furriness. In view of their compellingness, such mistakes evoke Kant's (1781) distinction between mere errors and certain deeper, inherent cognitive illusions. Thus, I disagree with Pylyshyn only regarding his optimism in hoping that, by repeating his powerful arguments loudly and slowly, he might succeed this time where he has failed before. Sufficient grounds for my skepticism is the fact that the Imagery Debate is perhaps the most remarkable modern duplication of controversies concerning the nature of 'ideas' which have persisted not just for thirty years but since the seventeenth century. In this recent re-enactment, Pylyshyn has played Arnauld (1683) against Kosslyn's Malebranche (1712) See Slezak (1992, 1995, 2002).

Of course, Pylyshyn is not vindicated merely because he was anticipated by Descartes and Arnauld. The striking historical parallels suggest that the fundamental problems at stake do not arise in any essential way from the data of modern experiments and computational theories. Indeed, just as we would expect in this case, we see a recurrence of the same perplexities not only throughout history, but also in more or less independent domains of cognitive science today.

What these doctrines have in common is the mistake of assuming that we apprehend our mental states rather than just *have* them. It is clear why such an implicit conception leads to positing a representational format - sentences or pictures - which is paradigmatically the sort of thing requiring an external, intelligent observer – the notorious homunculus. Computer simulation of certain theories does not necessarily prove pictorialism innocent of this charge. As Rorty put it, there is no advance in replacing the little man in the head by a little machine in the head. As Pylyshyn argues, resort to neuroscience is no help either.

Despite the jaundiced views of "philosophical" arguments (as distinct from "strictly empirical science") expressed by some pictorialists, Pylyshyn's critique suggests there remain grounds for Wittgenstein's (1953) gibe "in psychology there are experimental methods and *conceptual confusion*".

References

Arnauld, A. (1683/1990). *On True and False Ideas.* Trans. S. Gaukroger, Manchester: Manchester University Press.

Descartes, R. (1637). Dioptrics. *The Philosophical Writings of Descartes.* Trans. J. Cottingham, R. Stoothoff & D. Murdoch, Cambridge: Cambridge University Press.

Edelman, S. (1998). Representation is Representation of Similarities, *Behavioral and Brain Sciences*, 21, 449-498.

Kant, I. (1781). *The Critique of Pure Reason*, Trans. N. Kemp Smith, New York: St. Martin's Press.

Kosslyn, S.M.. (1994). *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, Mass.: MIT Press.

Malebranche, N. (1712/1997). The Search After Truth. Trans. T. Lennon & P. Olscamp, Cambridge: Cambridge University Press.

Pylyshyn, Z. (2002). Mental Imagery: In Search of a Theory. *Behavioral and Brain Science*, in press.

Slezak, P. (1992). When Can Images Be Reinterpreted: Non-Chronometric Tests of Pictorialism, *Proceedings of 14th Conference of the Society for Cognitive Science,* Hillsdale, N.J.: Lawrence Erlbaum, 124-129.

Slezak, P. (1995). The 'Philosophical' Case Against Visual Imagery, in P. Slezak, T. Caelli and R. Clark eds., *Perspectives on Cognitive Science: Theories, Experiments and Foundations,* Norwood, N.J.: Ablex Publishing Corporation, 237-271.

Slezak, P. (2002). The Tri-Partite Model of Representation *Philosophical Psychology*, In press.

Wittgenstein, L. (1953). *Philosophical Investigations*, Oxford: Basil Blackwell.

On the Origins of Perceived Sameness in Shape

Linda B. Smith (smith4@indiana.edu)
Psychology and the Program in Cognitive Science, Indiana University,
1101 East 10th Street, Bloomington, IN 47405

Background

What defines sameness in shape? A precise definition has proved elusive despite considerable theoretical and empirical efforts across several disciplines. However, a theory of shape is crucial to explaining human object recognition. The theoretical problem is that real instances of real categories are rarely ever the exact same shape. For example, rocking chairs, stuffed chairs, and desk chairs are the "same shape" only under some highly abstract description of shape. The present paper reports developmental evidence suggesting that this abstract description of object shape is a product of early category learning. The experiments focus on the period between 18 and 24 months of age, a period in which children progress from producing few object names (less than 100) to producing many (on average more than 200).

.

Experiment 1

There were two types of test stimuli: 3 dimensional lifelike replicas and 3-dimensional shape caricatures of the same things as illustrated in Figure 1. There were also two dependent measures of object recognition: (1) Recognitory play -- a child would be credited with recognizing an object as a phone if the child pretended to dial a number and/or talk on the object and (2) Name comprehension -- a child was credited with recognition if the child could select the target from among distractors given the name of the object. The children were divided into two groups according to noun vocabularies --- those with less than 100 nouns in their vocabulary and those with more than 100 nouns. Both groups of children recognized the Lifelike objects -- both by the play measure and the name comprehension measure. However, only the children with larger vocabularies recognized the Shape Caricatures. The fact that young children with few names for common do not recognize shape caricatures despite their accuracy in recognizing richly detailed instances of the same category indicates that the abstract representation of shape is a developmental product. The fact young children who are only slightly more advanced in their category knowledge recognize these shape caricatures suggests that early category learning plays a role in forming the processes of shape recognition.

Experiment 2

One possibility is that children learn to recognize shape caricatures, category by category. Alternatively, the developmental changes may be more general, changing how children perceive shape similarities for novel as well as known objects. This question was addressed in a second experiment. Children were introduced to a lifelike but (for young children) novel object, for example, an artichoke. The children were taught the object's name. On the critical test trial, three shape caricatures were presented to the child, one of which was a shape caricature of the originally named exemplar. The child was asked to indicate the named object, for example, "Where's the artichoke here?" If children must master the relevant shape properties category by category, then this task should be very hard because the caricatured artichoke only preserves some aspects of the original shape. If, however, children, are developing general perceptual skills that apply to novel shapes, then children who recognize the caricatures of familiar objects might also recognize the caricatures of novel ones. The results support this second possibility. Children with more than 100 object names in their productive vocabulary readily recognized the caricature of the newly learned noun. Children with fewer than 100 object names did not. These results strongly suggest that children are learning something general about the shape similarities relevant to object recognition and categorization.

Conclusion

The findings indicate that a complete theory of shape and object recognition will be a developmental theory. The relation of these results to contemporary theories of object recognition will be discussed.

Figure 1: A Lifelike and Caricature phone.

Acknowledgments

This research was supported by NICHD R01 HD 28675-08.

The Origins, Development, and Nature of Argument Understanding

Nancy L Stein (n-stein@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Avenue Chicago, IL 60637 USA

Elizabeth R. Albro
Department of Psychology, Wheaton College
Norton, MA USA

The early emergence and development of argumentation skill are the topics of this presentation. We argue that the mental schemas used to understand interactive arguments are influenced by the desire to maintain, dominate, or dissolve a relationship as much as they are by the desire to persuade and understand another person s position. Goals about personal relationships influence reasoning, thinking, and decision-making throughout arguments and negotiation. What appears to be illogical reasoning or irrational behavior is often quite rational and coherent, when the personal goals of the arguers are revealed. Personal goals also influence the outcomes of a negotiation and memory for what is said during the verbal interchange.

Data from three developmental studies will be presented to illustrate the relationship between personal-social goals and the content, organization, outcome and memory for an argument. We present different types of empirical evidence in support of our hypothesis, and we compare our early emergence hypothesis to the claim that argumentative skill emerges late in childhood and early adolescence. In support of our "early emergence" hypothesis, we focus on situations that are personally meaningful to young children and those that impact directly on their goals, beliefs, and well-being. We show that even the youngest children entering into an argument are able to generate and think about positive and negative reasons for pursuing different courses of action or for holding specific sets of beliefs.

We show, however, that argumentative thinking has an inherent bias that can be seen in adults' thinking as well as in young children's thinking. Arguers generally have more supporting knowledge for their own position than they do for their opponent's position. They also have more knowledge about the problematic aspects of their opponent's position than they do about their own position. Thus, they support a particular stance because they perceive more benefits accruing from their own position versus another.

We discuss the learning strategies that ameliorate this bias, both in social and in academic settings. We argue that current instructional strategies are often aimed at the wrong level of knowledge acquisition, in terms of teaching students how to write good arguments. The rhetorical concept of argument is often insensitive to the ways in which argument knowledge is stored psychologically. Most arguers, even adults, lack accurate knowledge about another's position. The focus for us, in terms of instruction and learning, has more to do with values, concerns, and beliefs underlying a position, the necessity to put each position on an equal footing, and the willingness to consider the legitimacy of different goals.

Therefore, we discuss the cognitive and emotional effects of a mediated conflict resolution training procedure. Our mediated instruction focuses on explanations for holding positions, plans for generating new goals, and strategies for adding conditions to favored goals to make them acceptable to an opponent. The effect of participating in meditated training is an increased understanding and accuracy of the opponent s position. By increasing understanding for the other s position, the participant incorporates input from the other, thereby increasing the new words and concepts that occur in thinking and reasoning. The cognitive and language effects of mediation will be discussed and contrasted with the effects of self imposed compromises and negotiations that do not entail compromise.

References

Stein, N.L. & Albro, E.R. (2001) The Origins and Nature of Arguments: Studies in Conflict Understanding, Emotion, and Negotiation. *Discourse Processes*.

Constructive Perception: An Expertise to Use Diagrams for Dynamic Interactivity

Masaki Suwa (suwa@sccs.chukyo-u.ac.jp)
Information and Human Activity, PRESTO, JST &
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodachi, Kaizu, Toyota, Aichi 470-0393 Japan

Diagrams Provide Dynamic Interactivity

By diagrams, we mean diagrammatic representations people use externally to their mind. They include pictures, sketches, charts, graphs and scribbles on napkins. Past literature (e.g. Anderson, Meyer and Olivier Eds, 2002) indicated that diagrams play facilitatory roles in inference and problem-solving; they reduce working memory load, serve as retrieval cues to evoke relevant information that might not otherwise be retrieved, promote inference by enabling perceptual judgements, and/or provide visuo-spatial cues for proper understanding of the structure of a problem. To serve these functions, the interpretation of the diagram needs to be static; it must stay the same in order not to introduce error in the operations performed from the diagram.

However, what diagrams could provide is not limited to such static interactivity. Rather, people using diagrams are encouraged to interpret them dynamically; the same appearance of parts of a diagram, especially when the diagram is vague and ambiguous, could evoke different interpretations at different times, dependent on what other elements surround the parts in focus at the moment or what the person has been thinking of. The situated cognition view (e.g. Clancey, 1997) corroborates this phenomenon. Dynamic interactivity of this sort afforded by diagrams is beneficial because it often enables dynamic construction of new thoughts on the fly in a situated manner.

A typical situation is design. Designers draw freehand sketches, often vague and ambiguous ones, and thereby see new features and relations among elements that they have drawn, ones not intended in the original sketch (Schon, 1983). These unintended discoveries promote the dynamic construction of new ideas and refine current ones. In recent years, we have explored ways that designers use sketches to dynamically construct design thoughts. Using the technique of protocol analysis, we examined the cognitive processes of experienced designers as they design through sketching. These protocols showed that the discovery of unintended perceptual features in sketches becomes a significant impetus for the generation of new ideas. Moreover, the generation of new ideas, in turn, was likely to become an impetus for further discovery of unintended perceptual features, so that each component process drives the other (Suwa, Gero and Purcell, 2000).

Constructive Perception to Benefit from Dynamic Interactivity

Dynamic interactivity, however, is by no means automatic when a diagram is available. To make it happen in using diagrams requires some cognitive skill, i.e. what we call constructive perception. By constructive perception, we mean self-awareness of the ways that perception underlies interpretations of diagrams. The self-awareness allows searching for other ways to perceive, enabling reorganization of the diagram to promote novel interpretations. We have found that this skill is useful in two different domains (Suwa, Tversky, Gero and Purcell, 2001). One is the design domain. During a conceptual design process, an experienced architect was likely to make unintended discoveries when he reorganized perception using this skill voluntarily. The other is in the task of multiple interpretations of ambiguous drawings. Novices instructed about this skill generated more interpretations from a single ambiguous drawing, and exhibited slower rate of decline of generation of interpretations over time, than those not instructed. Moreover, we have found that experienced designers are superior to laypeople in this skill (Suwa and Tversky, 2001). These findings raise two issues, one cognitive and the other didactic. What constitutes the expertise of constructive perception? How can people be trained to use it? Research on these will promote successful use of diagrams in people's intellectual activities, e.g. in learning environments.

Acknowledgments

I am grateful to Barbara Tversky for insightful discussions.

References

Anderson M., Meyer, B. & Olivier P. (eds) (2002). *Diagrammatic representation and reasoning,* London: Springer.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations.* Cambridge: Cambridge University Press.

Schon, D. A. (1983) *The reflective practitioner.* New York: Basic Books.

Suwa, M., Gero, J & Purcell, T. (2000). Unexpected discoveries and S-invention of design requirements: important vehicles for a design process. *Design Studies*, *21*, 539-567.

Suwa, M., & Tversky, B. (2001). Constructive perception in design. In J. S. Gero & M. L. Maher (Eds.) *Computational and Cognitive Models of Creative Design V*, Sydney: University of Sydney.

Suwa, M, Tversky, B, Gero J. & Purcell, T. (2001). Regrouping parts of an external representation as a source of insight. *Proceedings of the 3rd International Conference on Cognitive Science (pp.692-696).* Beijing, China: Press of University of Science and Technology of China.

Literary Cognition and Aesthetic Computing

Akifumi Tokosumi (akt@valdes.titech.ac.jp)
Department of Value and Decision Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Norikazu Yoshimine (yosimine@shokoku.ac.jp)
Shonan Kokusai Women's College
802 Engyo, Fujisawa, Kanagawa 252-0850, Japan

Literary text as a cognitive science problem

As a natural expansion from the tradition of natural language processing research, this paper presents our approach toward the poetic and artistic aspects of the human cognitive system. Our mental processes wouldn't stop when they understand the meaning of text. We always evaluate and appreciate them. Sometimes we contemplate, mediate and are moved. Cognitive science needs to invent vocabularies to describe these aspects of mental processes.

We assume that the interesting natures of literary texts are projecting following important problems:

(a) Successful literary texts attract readers, give motivations to read through, evoke subjective emotional experiences of "moved". (emotion problem)

(b) Many literary texts deal with fictitious contents. Transfer of factual knowledge may not be a major task of those texts. (knowledge problem)

(c) Successful literary texts receive high evaluations as artistic artifacts. Society of intelligent agents has mechanism to support such social activities. (art problem)

Based on psychological evidences, we have proposed an emotion oriented natural language processing model, particularly, a *wish* generation mechanism and an *aesthetic emotion* evocation mechanism (Tokosumi, 2001, Yoshimine and Tokosumi, 2001). The present paper is an attempt to give a partial answer to the emotion problem, and suggests possible solutions for the knowledge problem and the art problem.

Cognitive Computation hierarchy

The concept of affective computing (Picard, 1997) has advanced the way we talk about cognitive activities. Literary computing as a subordinate of aesthetic computing may bring new perspectives into the way we talk about language processing (Fig. 1)

Aesthetic Computing
literary computing, poetic computing, music computing, ...
Tokosumi (2001), Yoshimine and Tokosumi (2001)
Affective Computing
emotion/affect representation, emotional inference,...
Dyer (1983), Mueller (1990), Picard (1997)
Knowledge Computing
knowledge representation, case-based inference, ...

Fig. 1. Cognitive computation hierarchy.

In the set of programs called KEWP (Knowledge and Emotion Workbench Programs), we have been formalizing people's emotional experiences evoked by stories and other multimedia objects, which include content oriented emotions and aesthetic emotions.

Aesthetic emotions in literary experiences

Aesthetic emotions evoked by the quality of objects, such as linguistic expressions and other artistic forms, are important class of emotions yet to be investigated fully. Our treatment of aesthetic emotions described here is based on the cognitive theories of emotion (e.g. Frijda, 1986). We identify a cognitive appraisal component and an action readiness component for various aesthetic experiences and propose computational mechanisms to implement those components in the KEWP model.

Cognitive factors each component can deal with are:

(a) Cognitive appraisal component -- completeness, novelty, memory, ability recognition, competence, assimilation.

(b) Action readiness component -- possession, re-experience, creation, evangelic.

We also discuss the implication of the model as a competence-based neural architecture of the brain (Tokosumi et al., in press).

References

Dyer, M. G. (1983) *In-depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. Cambridge, MA: MIT Press.

Frijda, N. (1986) *Emotions*. Cambridge: Cambridge University Press.

Mueller, E. (1990) *Daydreaming in Humans and Machines: A Computer Model of the Stream of Thought*. Norwood, NJ: Ablex.

Picard, R. W. (1997) *Affective Computing*. Cambridge, MA: MIT Press.

Tokosumi, A. (2001) The Brain/Mind Machine: Toward modeling its wish generation processes. In Tadashi Kitamura (Ed.), *What should be Computed to Understand and Model Brain Function?* 43-51. Singapore: World Scientific.

Tokosumi, A., Noda, K., Anbo, T. and Matsumoto, N. (in press) A Competence-based Architecture for Aesthetic Emotions. *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics*. Hammamet, Tunisia.

Yoshimine, N. and Tokosumi, A. (2001) The Cognitive Model of Fiction Comprehending Control System. In N. Baba, L.C. Jain, and R. J. Howlett (Eds.) *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies: KES 2001,* 1228-33. Amsterdam: IOS Press.

Diagrams to Augment Cognition

Barbara Tversky[1], Julie Heiser[1], Paul Lee[1], and Jeffrey M. Zacks[2]

[1]{bt, jheiser, pauly@psych.stanford.edu
Department of Psychology
Stanford University
420 Jordan Hall
Stanford, CA 9435-2130

[2]{jzacks@artsci.wustl.edu}
Department of Psychology
Washington University
Campus Box 1125
One Brookings Drive
St. Louis, MO 63130-4899

Diagrams : A Cognitive Tool

Diagrams, such as maps, charts, graphs, and widely used as cognitive tools to promote memory and information processing, serving a variety of situated roles. They offload limited capacity working memory; they promote the use of space in inference and reasoning, they provide common ground for collaborative design (e. g., Kirsch, 1995; Larkin & Simon, 1987; Tversky, 2001).

One reason for the effectiveness of diagrams is that they map real or conceptual elements and relations to graphic elements and spatial relations in diagrammatic space. Diagrams have a rudimentary semantics and syntax. Diagrammatic elements, such as lines, blobs, crosses, and arrows have many possible interpretations derived from their geometric properties, but are disambiguated in context, much like the verbal concepts they approximate, such as relation and area. The elements can be combined in constrained ways to produce a multitude of meanings. This schematization has been a consequence of long term interactive situated use. Diagrams also use the spatial relations among elements to convey conceptual relations preserving varying levels of information, categorical, ordinal, interval.

Diagrams for Clarity

Diagrams can be used to organize and convey information and instructions, as in route maps and assembly directions. Here, the primary use is to instill prescribed information or linear actions. This will be exemplified by two projects on production and use of diagrams, one in route finding and the other in object assembly. In both cases, descriptions and depictions reveal the same underlying conceptual structure for traversing a route or assembling an object. These structures include both representations and procedures.

Diagrams for Creativity

Diagrams can also function to aid inference and promote creativity. Here, the goal is to come up with new ideas, ideas not anticipated by the designer of the diagram. This will be described by Suwa (2002; Suwa & Tversky, 2001) in a project on diagrams generated and used in design.

In both cases, diagrams are inevitably replete with ambiguity. In the former, context disambiguates, instilling clarity and avoiding confusion. In the latter, ambiguity is a resource for creativity.

Acknowledgments

We are grateful to Office of Naval Research, Grants Number NOOO14-PP-1-O649, N000140110717, and N000140210534 to Stanford University for support for this research.

References

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, 73, 31-68.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science,* 11, 65-99.

Suwa, M. (2002, in press). Constructive perception: An expertise to use diagrams for dynamic interactivity. *Proceedings of the Cognitive Science Society*.

Suwa, M., & Tversky, B. (2001). Constructive perception in design. In J. S. Gero & M. L. Maher (Eds.) *Computational and Cognitive Models of Creative Design* V, Sydney: University of Sydney.

Tversky, B, Zacks, J., Lee, P. U., & Heiser, J. (2000). Lines, blobs, crosses, and arrows: Diagrammatic communication with schematic figures. In M. Anderson, P. Cheng, and V. Haarslev (Editors). *Theory and application of diagrams.* Pp. 221-230. Berlin: Springer.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis, (Ed.), *Spatial schemas and abstract thought.* Pp. 79-111. Cambridge: MIT Press.

Papers

Representation Strength Influences Strategy Use and Strategy Discovery

Martha W. Alibali (mwalibali@facstaff.wisc.edu)
Department of Psychology
University of Wisconsin—Madison
1202 W. Johnson Street
Madison, WI 53706 USA

Tara L. Booth (boothtara@hotmail.com)
Department of Psychology
University of Wisconsin—Madison
1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

When attempting to solve a problem, individuals may activate multiple potential representations for that problem. Further, different representations may be activated more or less strongly. This study investigated how strength of problem representations is related to patterns of strategy use and strategy discovery. We hypothesized that the more strongly a particular representation is cued, the more likely participants should be to use a strategy that corresponds with that representation. Further, for individuals who do not initially have a corresponding strategy in their repertoires, the more strongly a particular representation is cued, the more likely participants should be to discover a strategy that corresponds with that representation. These hypotheses were investigated among adults solving word problems about constant change. The problems could be represented in terms of discrete change or continuous change. We varied two types of cues to discrete and continuous problem representations: linguistic cues and graphical cues. Both linguistic and graphical cues influenced strategy use, and the effects of the two cue types were additive. Among participants who did not use a continuous strategy at the outset of the study, discovery of a continuous strategy was relatively rare, and only participants who received a continuous graph tended to discover a continuous strategy. The findings suggest that it may be fruitful to consider problem representations as graded and variable rather than all-or-none.

Introduction

One step in the process of solving a problem is forming a mental representation of important features of that problem. Problem representations have been invoked to explain many aspects of people's problem-solving behavior, including success, solution times, strategies, and errors (e.g., Kotovsky, Hayes, & Simon, 1985; Lovett & Schunn, 1998). In the present study, we investigate links between problem representations and patterns of strategy use and strategy discovery.

Problem representations are sometimes conceptualized as integrated wholes, such that a particular representation is retrieved in its entirety from memory, and applied to the problem at hand (e.g., Larkin, 1983). Although this characterization may apply in some cases (e.g., for well-practiced problems), we suggest that in most cases, problem representations are constructed at the moment of solving, based on both perceivable features of the problem and on knowledge retrieved from memory about problem content or about particular problem-solving strategies (McNeil & Alibali, 2000). We further suggest that the knowledge activated in constructing a problem representation may be more or less strongly activated, and thus, aspects of the representation may be graded rather than all-or-none (see Munakata, McClelland, Johnson, & Siegler, 1997, for discussion).

There is some support in the literature for the notion that problem representations may be graded. Kaplan and Simon (1990) studied this issue in the context of the *mutilated checkerboard* problem. In this problem, the squares from two diagonally opposite corners of a checkerboard are removed, and the solver's task is to cover the remainder of the checkerboard with dominoes, each of which covers exactly two squares, or to prove that such a covering is impossible. Because the two diagonally opposite corners of a checkerboard are the same color (both black or both white), the covering task is indeed impossible; however, this fact is notoriously difficult for solvers to discover. In their experiment, Kaplan and Simon varied the strength of various cues to the "paired-ness," or parity, of the squares. Solvers were quicker to discover that the covering was impossible when adjacent squares were labeled "bread" and "butter" than when the squares were not labeled, or when they were labeled with terms that did not form a strongly associated pair ("black" and "pink"). The bread-and-butter cue to parity facilitated a stronger representation of this crucial problem feature, and this led to faster discovery of the problem solution.

The purpose of the present study was to investigate whether variations in the strength of problem representations can account for variations across solvers in patterns of strategy use. Several past studies have investigated the links between problem representation and strategy use (e.g., Alibali, Bassok, Solomon, Syc, & Goldin-Meadow, 1999; Morales, Shute, & Pellegrino, 1985; Siegler, 1976). However, to date, little research has examined how the *strength* of representations relates to patterns of strategy use. We hypothesized that the more

strongly a particular representation is activated, the more likely participants would be to use a strategy that corresponds with that representation. Further, for individuals who do not initially have a corresponding strategy in their repertoires, the more strongly a particular representation is activated, the more likely participants would be to discover a strategy that corresponds with that representation.

We also wished to examine the effects on strategy use of having multiple, incompatible representations that are simultaneously active. We hypothesize that the operative factor in determining which representation guides solution is the *relative* strength of a particular problem representation. Therefore, when multiple, potentially incompatible problem representations are simultaneously active, participants' performance should be more variable than when a single problem representation is active.

This study investigated these hypotheses among adults solving word problems about constant change (Bassok & Olseth, 1995). The problems could be represented in terms of either discrete, stepwise change or smooth, continuous change. The experiment varied two types of cues to discrete and continuous problem representations: linguistic cues and graphical cues. The linguistic cues were drawn from previous research on people's verbal descriptions of constant change problems (Alibali et al., 1999). The graphical cues were chosen based on previous research about graph comprehension (Zacks & Tversky, 1999), which indicated that line graphs cue representations of continuous changes in values, whereas bar graphs cue representations of discrete changes in values.

In some conditions in the present experiment, the linguistic and graphical cues converged on a single representation. In other conditions, linguistic cues alone were provided. In still other conditions, the linguistic cue pointed toward one representation and the graphical cue pointed toward the other representation. If stronger representations lead to more frequent use of a corresponding strategy, participants should use that strategy most often in the corresponding cues case, and least often in the conflicting cues case. The single-cue case should fall somewhere in the middle.

Method

Participants

Participants were 158 Introductory Psychology students at the University of Wisconsin—Madison. The sample included 58 males, 90 females, and 10 participants who did not disclose their gender. Most participants were either freshmen (58%) or sophomores (23%), and all had taken at least one semester of college-level mathematics. Students received extra credit points for Introductory Psychology in exchange for their participation.

Procedure

Students were tested in a small classroom in groups of 15 to 25. They were given up to 45 minutes to complete a set of 10 story problems. They were instructed to work the problems in the order presented and not to return to earlier problems after solving later ones. They were also asked to show all of their work and to circle their final solution for each problem. Students were not permitted to use calculators.

Materials

Students received a packet of 10 words problems about constant change, based on those used in prior studies (e.g., Alibali et al., 1999; Bassok & Olseth, 1995). The first 8 problems in each packet focused on quantities that changed continuously (e.g., rain falling, a tree growing), and these problems were the site of the manipulation. As seen in Table 1, the wording of the problems was varied to cue either a discrete representation or a continuous representation. Cues to the discrete representation included amount-like units for the initial and final quantities (e.g., 5 millimeters), mention of individual units of time (e.g., the 12 weeks), and explicit reference to the constant. Cues to the continuous representation included rate-like units for the initial and final quantities (e.g., 5 millimeters per week), mention of the entire period of time (e.g., the 12-week period), and explicit reference to rate.

In addition, as seen in Figure 1, the problems were accompanied either by bar graphs, by line graphs, or by no graphs at all. Thus, the study utilized a 2 (verbal cues: discrete or continuous) x 3 (graphs: discrete [bar], continuous [line], or none) between-subjects design. The final two problems were transfer problems that were the same across all conditions, and they utilized discrete content (e.g., plants per row in a garden), discrete wording, and no graph. Participants' performance on the transfer problems is not addressed in this paper.

Table 1: Sample Problems

Discrete Wording A sapling grows for 12 weeks. The number of millimeters it grows in each successive week increases by a constant from the number in the previous week. In the first week the sapling grows 5 millimeters and in the twelfth week it grows 137 millimeters. How many millimeters does the sapling grow in total over the 12 weeks?
Continuous Wording A sapling grows for a period of 12 weeks. The rate at which it grows increases steadily over the period, from 5 millimeters per week at the beginning of the first week to 137 millimeters per week at the end of the twelfth week. How many millimeters does the sapling grow in total over the 12-week period?

Coding

Each problem was initially scored as correct, incorrect, or no response. Next, the strategy that each participant used to solve each problem was coded, and all strategies were

classified as either discrete, continuous, or other (unclassifiable). Coding definitions are presented in Table 2. Most strategies in the "other" category were conceptually flawed attempts to solve the problems (e.g., adding or multiplying the initial and final values, or multiplying the initial value by the number of intervals and then adding the final value).

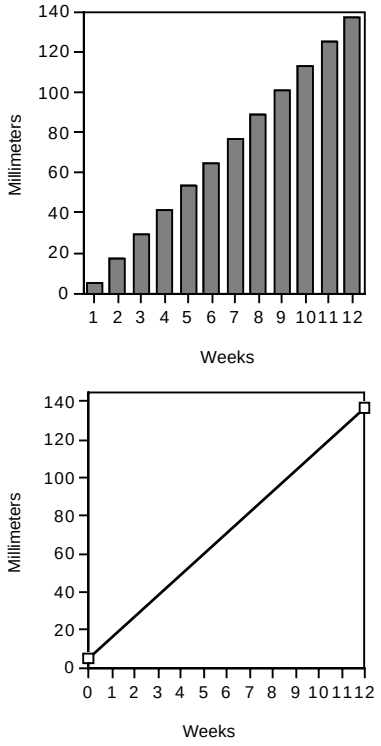


Figure 1: Sample bar and line graph.

Data from a subsample of 30 participants were rescored by a second coder to establish reliability. Agreement between the coders was 97% (N = 282 problems).

Table 2: Strategy Codes

Strategy	Definition
Discrete Strategies	
Sum	Participant finds the constant increase, calculates the value for each interval, and adds these values
Gauss	Participant adds values for initial and final intervals and multiplies this sum by half of the number of intervals
Continuous Strategies	
Average	Participant finds average value per interval and multiplies by number of intervals
Calculus	Participant sets up equation and integrates
Area	Participant calculates area using geometric methods (e.g., adds areas of rectangle and triangle)

Results

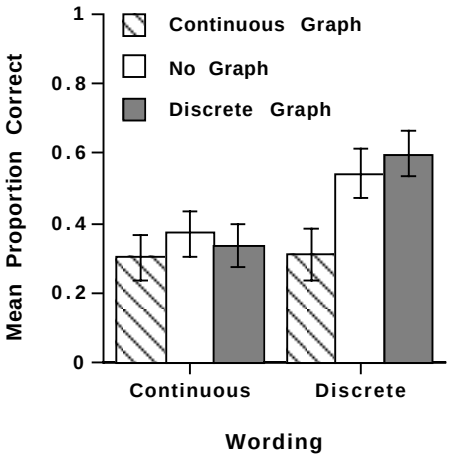
Our analysis focuses on the effects of cues to discrete and continuous representations on participants' overall level of performance, their strategy use, and their strategy discovery. The analyses reported here focus on the first eight problems, where the experimental manipulation occurred.

Problem solutions

We first examined whether variations in problem wording and graphs influenced whether participants solved the problems correctly. The interaction of wording and graphs was not significant, but there were main effects of both factors. As seen in Figure 2, participants who received problems with discrete wording were more successful than participants who received problems with continuous wording, $F(1, 152) = 7.22$, $p < .01$, despite the fact that all of the problems involved quantities that changed in a continuous fashion. Graphs also influenced success, $F(2, 152) = 3.61$, $p < .05$. Participants who received problems with continuous graphs performed most poorly, and participants who received discrete graphs and no graphs performed similarly well. Post hoc tests indicated that the discrete-graph and no-graph groups each differed significantly from the continuous-graph group, but they did not differ from one another.

Why were continuous wording and continuous graphs associated with poorer performance? Before addressing this question, we first consider patterns of strategy use.

Figure 2: Proportion of problems solved correctly by participants in each group.



Strategy use

Participants used discrete strategies much more often than continuous strategies in the dataset as a whole (63% vs. 16% of trials). Because of this, we used frequency of discrete strategy use rather than frequency of continuous strategy use as the dependent measure in our analysis of strategy use, to avoid possible floor effects in some of the cells of the design.

Once again, the interaction of wording and graphs was not significant, but there were main effects of both factors. As seen in Figure 3, participants who received discrete wording used discrete strategies more frequently than participants who received continuous wording, $F(1, 152) = 6.97$, $p < .01$. Graphs also influenced whether participants used discrete strategies, $F(1, 152) = 15.10$, $p < .001$. Participants who received continuous graphs used discrete strategies least often, and participants who received discrete graphs and no graphs used discrete strategies much more often. Post hoc tests indicated that the discrete-graph and no-graph groups both differed significantly from the continuous-graph group, but they did not differ from one another. Thus, both types of cues influenced participants' strategy use, and the effects of wording and graphs appear to be additive.

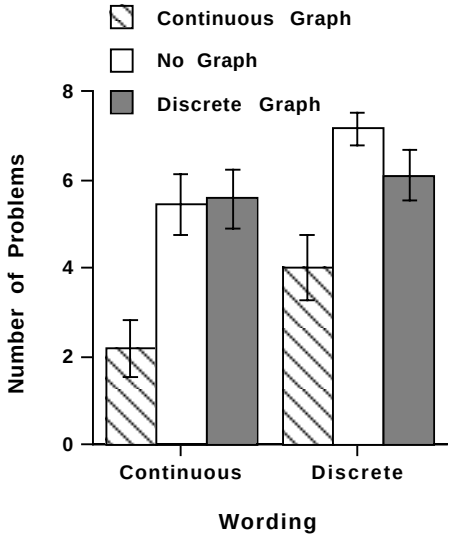


Figure 3: Number of problems solved using discrete strategies by participants in each group.

We now return to the question of why continuous wording and continuous graphs were associated with poorer problem-solving success in the initial analysis. The strategy analysis indicates that participants who received continuous wording or continuous graphs were less likely to use discrete strategies. Perhaps participants also applied discrete strategies more successfully than continuous ones, and this accounts for the performance differences seen in Figure 2. However, the success rates for discrete and continuous strategies did not differ. Participants succeeded on 50% of trials on which they used discrete strategies ($N = 1012$), and 52% of trials on which they used continuous strategies ($N = 238$).

To shed additional light on the performance differences, we examined the number of trials (out of 8) on which participants used discrete strategies, continuous strategies, or other, unclassifiable strategies, which were typically incorrect. These data are presented in Table 3. As seen in the table, participants who received continuous wording or continuous graphs used more strategies in the "Other" category than participants who received discrete wording and discrete graphs, or discrete wording and no graphs. Strategies in this category were often conceptually flawed attempts to solve the problems, suggesting that participants who received continuous wording or graphs were often at a loss as to how to solve the problems. They may have attempted various strategies in an effort to generate or discover a strategy compatible with the continuous problem representation.

Based on the data in Table 3, we next examined whether participants who received converging cues to a single representation used the corresponding strategies more often than participants who received only a single cue. Participants who received converging cues to continuous representations (i.e., wording and graphs) were indeed more likely to use continuous strategies than participants who received only a single cue (i.e., wording only), $F(1, 152) = 9.38$, $p < .01$. For discrete representations, there was no significant difference in discrete strategy use between participants who received converging cues to a discrete representation and participants who received only a single cue (indeed, the direction of the effect is opposite that predicted). This lack of difference suggests that the discrete representation was readily accessible even in the absence of the discrete graph.

Table 3: Mean Number of Strategies of Each Type Used by Participants in Each Group

		Strategy Category		
Wording	Graph	Discrete	Contin.	Other
Discrete	Discrete	6.07	0.74	1.15
Discrete	Contin.	3.96	2.00	1.46
Discrete	None	7.16	0.00	0.76
Contin.	Discrete	5.59	0.59	1.78
Contin.	Contin.	2.15	3.11	2.70
Contin.	None	5.46	0.92	1.54

We also compared the strategy use of participants who received converging cues to a single representation and participants who received cues to both representations. Participants who received converging cues to discrete representations used discrete strategies more often than participants who received discrete wording and continuous graphs, $F(1, 152) = 5.46$, $p = .02$, but they did not differ significantly from participants who received continuous wording and discrete graphs. Participants who received converging cues to continuous representations used continuous strategies more often than participants who received continuous wording and discrete graphs, $F(1, 152) = 12.62$, $p < .001$, but they did not differ from participants who received discrete wording and continuous graphs, although the effect was in the predicted direction, $F(1, 152) = 2.18$. This pattern of findings suggests that, for problems like those used in the present study, graphs may be more effective than wording as a cue to problem representations.

Strategy discovery

We next turn to the issue of strategy discovery. As noted above, continuous strategies were used relatively infrequently in the dataset as a whole. However, some participants who did not use continuous strategies at the outset of the session appeared to "discover" continuous strategies in the course of solving the eight problems. Note that we do not mean to imply that these participants invented continuous strategies "from scratch." Instead, we believe that participants realized that they could apply familiar techniques such as averaging or calculating area as a method for solving the constant change problems. In this sense, they "discovered" continuous strategies.

Was discovery of a continuous strategy facilitated by cues to a continuous representation? To address this question, we eliminated all participants who used a continuous strategy on the very first problem (N = 19), because those participants may have already had the continuous strategy in their repertoire before the session began, instead of discovering it during the session. We then examined the proportion of remaining participants in each group who used a continuous strategy on at least one of the remaining seven problems.

As seen in Table 4, the only groups in which a substantial proportion of participants discovered the continuous strategy were those that received continuous graphs. Thus, graphs were an important cue in fostering the discovery of a continuous strategy. It is also worth noting that the continuous strategy was discovered most often in the group that received converging cues to a continuous strategy, and indeed, was never discovered in the group that received only a single cue (wording) to a continuous strategy. Similarly, the continuous strategy was never discovered among the group that received continuous wording with a discrete graph. However, 18% of participants who received discrete wording with a continuous graph discovered the continuous strategy. On the whole, the data are compatible with the view that a stronger representation is more likely to lead to strategy discovery.

Table 4: Percent of Participants in Each Group Who Discovered a Continuous Strategy

Wording	Graph	% who Discovered Continuous Strategy
Discrete	Discrete	8
Discrete	Contin.	18
Discrete	None	0
Contin.	Discrete	0
Contin.	Contin.	26
Contin.	None	0

Discussion

In this study, both linguistic and graphical cues to problem representations influenced participants' strategy choices and strategy discovery. The overall analysis indicated main effects of both cue types on success and strategy use. Focused contrasts indicated that participants who received converging cues were more likely to use a target strategy than were participants who received a verbal cue to the target representation but a graphical cue to the alternative representation. For the continuous representation, participants who received converging cues were also more likely to use corresponding (continuous) strategies than participants who received the wording cue alone (i.e., with no accompanying graph). However, for the discrete representation, participants who received converging cues and participants who received wording cues alone used corresponding (discrete) strategies about equally often.

Although discovery of a continuous strategy was rare in the sample as a whole, graphical cues appeared to be especially important for strategy discovery. Continuous strategies were discovered most often in the group that received converging cues to a continuous representation, and second most often in the group that received discrete wording with a continuous graph. These findings underscore the importance of graphical representations in helping individuals construct mental models of problem situations. Our findings are compatible with Kalchman, Moss and Case's (2001) claim that line graphs are especially important in the development of understanding of mathematical functions.

Several aspects of the results converge to suggest that the "default" representation for constant change problems is one of discrete rather than continuous change. First, the large majority of problems were solved using discrete strategies. Second, participants who received bar graphs performed similarly to participants who received no graphs at all. This suggests that participants did not need the aid of the bar graphs in order to construct discrete mental models of the problem situations. They appeared to construct discrete representations spontaneously, even in the absence of the bar graphs. In contrast, participants who received line graphs performed quite differently from participants who received no graphs. The line graphs appeared to help participants construct continuous mental models of the problem situations, as the strategy discovery data suggest.

Why might discrete representations be more readily available to participants than continuous ones? One possibility has to do with the nature of the mathematical relations that are involved in working with the representations. Strategies compatible with discrete representations tend to rely on additive relations (e.g., summing the values for each increment), whereas strategies compatible with continuous representations tend to rely on multiplicative relations (e.g., multiplying the average value times the number of increments). Because additive relations are simpler and more fundamental than multiplicative ones, they may be noticed first. This hypothesis implies that individuals with strong mathematics skills should be especially likely to use continuous strategies a possibility we intend to examine in future work.

Even participants who received both linguistic and graphical cues to a continuous representation used continuous strategies relatively infrequently. The high incidence of unclassifiable strategies among participants who received continuous cues suggests that many

participants were unable to generate a strategy compatible with the continuous representation. It is possible that some of these unclassifiable strategies were generated based on hybrid representations that combined both discrete and continuous elements. If the "default" representation for constant change problems is discrete, as we argued above, then cues to a continuous representation may create a situation in which multiple, incompatible representations are simultaneously active. Indeed, many of the unclassifiable strategies included both additive components, reminiscent of discrete strategies, and multiplicative components, reminiscent of continuous strategies. A more detailed analysis of these unclassifiable strategies may shed light on processes of strategy construction.

It also seems worth noting that a small number of participants altered the presented graphs. On problems with line graphs, some participants drew lines down to the x-axis to "discretize" the graph, and on problems with bar graphs, some participants drew a line across the tops of the bars to "linearize" the graph. In this regard, it is interesting to note that participants who received discrete wording and a discrete graph were slightly more likely to use continuous strategies than participants who received discrete wording and no graph. It is possible that even a bar graph can sometimes cue a continuous representation, because the linear relationship between the variables is an emergent feature of the bar graph.

In sum, the present findings add to the body of literature elucidating the links between problem representation, strategy use, and strategy discovery. For constant change problems, graphical representations, and in particular, line graphs, were important cues to strategy use and strategy discovery. However, many participants in this study failed to discover a continuous strategy. The findings suggest that people often activate multiple representations for individual problems, and if these representations are incompatible, people may have difficulty generating an effective strategy for solving the problems. More generally, the present findings suggest that, to understand patterns of strategy change and strategy discovery, it is fruitful to conceptualize problem representations as graded and variable rather than all-or-none.

Acknowledgments

We thank Leslie Petasis for assistance with a pilot study that led up to this project, Maureen Kaschak for assistance with data collection and coding, and Nicole McNeil and the members of the Cognitive Development Research Group at the University of Wisconsin—Madison for helpful feedback and suggestions.

References

Alibali, M. W., Bassok, M., Solomon, K. O., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science, 10*, 327-333.

Bassok, M., & Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1522-1538.

Kalchman, M., Moss, J., & Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 1-38). Mahwah, NJ: Lawrence Erlbaum Associates.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology, 22*, 374-419.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17*, 248-294.

Larkin, J. H. (1983). The role of problem representation in physics. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 75-98). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lovett, M., & Schunn, C. (1998). Task representations, strategy variability and base-rate neglect. *Journal of Experimental Psychology: General, 128*, 107-130.

McNeil, N. M. & Alibali, M. W. (2000). Learning mathematics from procedural instruction: Externally imposed goals influence what is learned. *Journal of Educational Psychology, 92*, 734-744.

Morales, R. V., Shute, V. J., & Pellegrino, J. W. (1985). Developmental differences in understanding and solving simple mathematics word problems. *Cognition and Instruction, 2*, 41-57.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review, 104*, 686-713.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory & Cognition, 27*, 1073-1079.

Integrating Decay and Interference: A New Look at an Old Interaction

Erik M. Altmann (ema@msu.edu)
Department of Psychology
Michigan State University
East Lansing, MI 48824

Christian D. Schunn (schunn@pitt.edu)
Learning Research & Development Center
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

An old but important debate about human memory concerns whether decay (indexed by time) or interference (indexed by amount of distracting information) is the cause of forgetting. We argue, based on a simple functional analysis, that this is a false dichotomy. Both processes must be at work, in that distracting information must decay to allow the cognitive system to have any hope of retrieving target information amidst the unavoidable clutter of a well-stocked memory. This analysis predicts that subtle decay effects should be pervasive, even in data produced by interference theorists to show that decay was impossible. A re-analysis of data from Waugh and Norman (1965) does indeed reveal decay effects that were dismissed by the authors as inconsequential and have been ignored by most investigators since. We fit a formal model integrating decay and interference to the Waugh and Norman data, and to the decay data of Peterson and Peterson (1959) to show that one model provides an improved account of two ostensibly divergent data sets.

Introduction

> "Decay must be one of the most discredited theories in psychology, amongst many distinguished competitors."
> — Memory researcher Robert Bjork, Michigan State University, Sept. 27, 2000.

How is information lost from human memory? Of the many potential metaphors, the two main competitors have historically been decay (a process indexed by time) and interference (a process indexed by the amount of distracting information "cluttering up" the mental desktop).

Of these two metaphors, decay has been the less popular, as the quotation above suggests. Memory researchers have often simply not wanted to credit the idea that memory could deteriorate by a time-indexed biological process (e.g., Keppel & Underwood, 1962; McGeoch & Irion, 1952; Postman, 1971; Waugh & Norman, 1965). Evidence often cited against decay includes the slowdown of forgetting during sleep (e.g., Ekstrand, 1972), though to interpret this slowdown as evidence against decay one must assume that the decay rate is the same during sleep as during wakefulness. Given the controversial nature of what little we do understand about brain activity during sleep, it seems equally likely that the decay rate is different in different states of consciousness. Another argument against decay is based on the observation that time by itself cannot be causal. As famously put by McGeogh, "In time, iron may rust and men grow old, but the rusting and the aging are understood in terms of the chemical and other events which occur in time, not in terms of time itself" (McGeogh & Irion, 1952, p. 402). Today, many important memory theories exclude decay (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1992), and cognitive textbooks often present decay theory as a historical footnote rather than as an active hypothesis (e.g., Ashcraft, 2002; Galotti, 1999; Reed, 2000.)

But decay is far from a footnote. Since the original studies of Brown (1958) and Peterson and Peterson (1959), various approaches have been taken to try to isolate decay from interference (Reitman, 1974; Baddeley & Scott, 1971; Turvey, Brick, & Osborne, 1970). Interference theorists themselves discovered that retention interval moderates proactive interference (e.g., Loess & Waugh, 1967). Decay is represented in select memory theories (Anderson & Lebiere, 1998; Richman, Staszewski, & Simon, 1995) and interpretations of the literature (Anderson, 2000; Baddeley, 1990; Wickelgren, 1977). Finally, it is increasingly clear that McGeoch's polemic (quoted above) misses the mark, given converging evidence that decay has neural correlates. For example, Fuster (1995, p. 247) observes that firing rates of particular pyramidal cells in the monkey show decay "reminiscent of the well-known decay of human short-term memory." And decay in the form of "leak currents" is an integral part of neural network simulation of the hippocampus (O'Reilly & Munakata, 2000).

The current study aims to bolster the case for a general and functional decay process and, more specifically, to show that evidence for decay exists behind "enemy lines," in the very data that purportedly showed that decay was impossible. The paper is organized as follows. We begin with a functional analysis that argues that decay (or some similar, non-interference forgetting process) must function in memory if memory is to function at all. We then develop predictions of this analysis for a classic study in the literature on interference, the Waugh and Norman (1965) probe digit experiment. A re-analysis of the data from that experiment supports the prediction of subtle decay effects. To make the prediction quantitative, we fit the data with a model based on a memory theory that offers the building blocks to integrate decay and interference. Finally, we turn to the original Peterson and Peterson (1959) data set on decay, and fit the same model to it. The goal is to show that one model, with decay and interference represented as functionally interacting processes, parsimoniously accounts for two ostensibly divergent data sets.

A Functional Perspective on Decay

The functional argument for decay is perhaps best illustrated with an example. As one drives an automobile through various speed zones, it is important to mentally register each change in the speed limit and update memory accordingly. However, if each change in the speed limit contributed

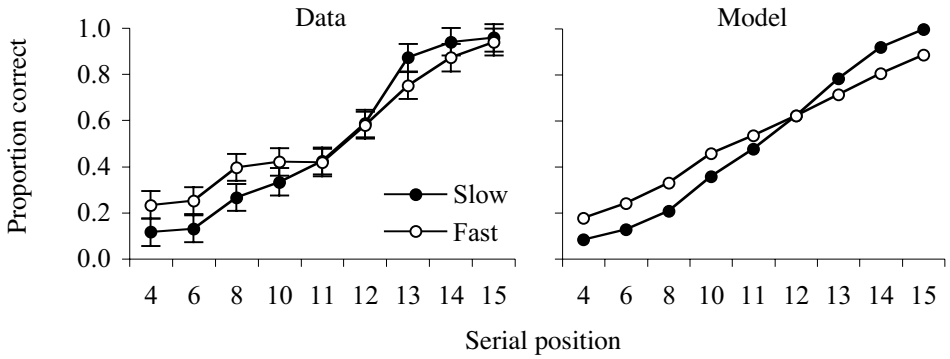


Figure 1. Left: Data from Waugh and Norman (1965), Exp. 1. Error bars are within-participants confidence intervals (Loftus & Masson, 1994) for the serial position x presentation rate interaction. Right: Fit of the functional decay model.

monotonically to interference in memory for speed limits, it would quickly become impossible to remember the current speed limit, whatever it might be. Although interference is clearly a potent source of forgetting, even at long delays (Keppel, Postman, & Zavortink, 1968), there must be some mitigating process if memory is to support our everyday needs (c.f., Luria, 1968).

Interference theorists themselves have been forced to acknowledge this functional need, and the literature consequently has known decay by other names. One example of decay being reinvented is the notion of forgetting through "stimulus fluctuation" (Estes, 1955; Gillund & Shiffrin, 1984; Landauer, 1975). The hypothesis is that changes in the environment gradually reduce the set of cues available to activate distracting information. However, this gradual environmental change in how distractors are cued is typically only specified at an abstract level, in terms that are functionally equivalent to a simple time-indexed process. A second example is the mystical sounding "spontaneous recovery" of previously extinguished distractors. Inherited from the behaviorists, this construct was applied by early interference theorists to explain, among other findings, the effect of retention interval in the Brown-Peterson paradigm (Keppel & Underwood, 1962).[1] Again, however, spontaneous recovery and decay seem functionally equivalent: In one case distractors gain strength relative to the target, and in the other case the target loses strength relative to distractors.

Thus, we argue that any non-interference forgetting indexed by time may as well be known as decay. What we propose to do is play out the functional implications and search for decay in time-based experimental manipulations where one might not otherwise expect to find it.

Revisiting Waugh and Norman (1965)

The classic data set of Waugh and Norman (1965) is discussed in many modern cognitive textbooks (including those cited above) usually to illustrate the importance of interference as a forgetting mechanism. In the *probe-digit* paradigm used in this experiment, the participant is presented with a list of digits and then asked to report one of them. The target digit (the one to report) is indicated by a probe digit given immediately after the list is presented. The probe also occurred exactly once during the list proper, and the target is the digit that followed the probe in the list proper. The serial position of the probe in the list proper changes randomly across trials, so accuracy across trials measures item retention as a function of serial position.

The experiment included a within-subjects manipulation of presentation rate to test whether the chronological age of the target item affected its retention. In the Fast condition, digits were presented once every 250 msec, and in the Slow condition they were presented once per second. The logic was that if decay caused forgetting, then Fast items should be more accurately remembered in response to the probe. If interference caused forgetting, then only serial position should affect retention. That is, late items in the list should be recalled better than earlier items, in both the Fast and Slow conditions, because late items suffered fewer intervening items before the probe, and hence should suffer less retroactive interference.

The Waugh and Norman data are plotted in Figure 1 (left panel), with serial position of the target along the x-axis. The curves represent the two presentation rates. The effect of serial position is readily apparent in both conditions, with later items recalled much more accurately than earlier items. This effect, and its similarity across the two conditions, was enough for Waugh and Norman to reject decay as cause of forgetting. They do note "a slight interaction" of serial position and presentation rate, but dismiss its importance — "the effect of rate is relatively small compared to the effect of serial position" — and report no statistics. Following their lead, no contemporary textbook that we have examined (including those cited above) even acknowledges the interaction, despite the fact that it fairly leaps off the page. The occasional investigator has observed the interaction and speculated about causes (Massaro, 1970; Hintzman, 1978; Wickelgren, 1977), but the theoretical significance of this interaction for decay theory has not been fully pressed.

To confirm the interaction statistically, we conducted a 9 x 2 repeated measures analysis of variance on the data.[2] The interaction of serial position and presentation rate is highly reliable, $F(8, 24) = 5.1, p < 0.001, MSE = .0033$.

[1] "The increase in [proactive interference] with increase in length of the retention interval may be accounted for by the recovering of extinguished interference associations" (Keppel & Underwood, 1962).

[2] We reconstructed the Waugh and Norman data by digitally scanning the individual participant data graphs in that report and overlaying a grid to estimate the actual numbers.

The Functional Decay Model

To explain the Waugh and Norman interaction, we modeled it using the central memory constructs of the ACT-R cognitive theory (Anderson & Lebiere, 1998). The model's fit is presented in the right panel of Figure 1. In addition to qualitatively capturing the interaction, quantitative measures of fit are quite close, with $RMSD = .054$ and $R^2 = .965$.

This *functional decay* model is based on the activation mechanism illustrated in Figure 2. The two curves in that figure plot the activation of each item in a list, just before the probe is presented. That is, the curves represent a snapshot of every potential target's activation immediately after all the potential targets have been presented. The curves are produced by the following activation function, adapted from ACT-R's Base Level Learning equation.

$$A = \ln(\frac{n}{\sqrt{T}}) \qquad \text{Equation 1: Base-level activation}$$

A is the activation of an item, n is the number of times the item has been retrieved from memory since it was encoded, and T is the age of the item (time from encoding to present). Equation 1 thus computes activation as a function of frequency of use. The premise behind this function is that historical need for information is a predictor of future need and thus should affect item availability (Anderson & Milson, 1989). Activation decays in this function as a power function of age of the item, T. The exponent of this function (-0.5) is, within ACT-R, a relatively constant parameter of the cognitive architecture governing the decay rate of a memory trace. However, this is not the only factor governing an item's base-level activation. For example, a rehearsal process could increase activation by increasing the value of the usage parameter, n. In the Waugh and Norman model, we fix n at 1, on the assumption that items were not differentially rehearsed. Waugh and Norman anticipated the possibility that differential rehearsal could confound their results, and instructed their participants to rehearse only the most recent item of the list, if they rehearsed at all.

In terms of the activation curves in Figure 2, Equation 1 (with n constant across items) predicts that the latest (most recent) item is the most active, the next most recent item the next most active, and so on. Items in the Slow condition are on average less active than items in the Fast condition, because a Slow item at a given serial position is older than a Fast item at the same serial position, so has decayed more.

In addition to the base-level activation governed by Equation 1, the second source of activation for a memory trace under ACT-R theory is priming through associative links. In other words, an item is activated associatively when cues to that item are themselves activated. Associative priming is how ACT-R must explain that fact that elements other than the latest (in the probe digit paradigm) can be retrieved at all. The background theoretical assumption is that, in response to a retrieval request from central cognition, the memory system delivers the trace that is the most active at that instant. In Figure 2, the item with the highest base-level activation is the last one in the list, so if base-level activation were the only activation a memory trace could have, then only the last item would ever be retrieved. This is clearly not how the cognitive system operates, in the probe digit paradigm or in general; people are perfectly able to retrieve thoughts other than the most recent. ACT-R implies that such retrieval depends heavily on retrieval cues delivering activation through associative links. In Bayesian terms, base-level activation reflects the influence of history (retrieval history, in particular, as captured by base-level activation), whereas associative priming from retrieval cues reflects the influence of the current context.

In the probe-digit paradigm, we assume that the role of associative priming plays out through an associative link between the probe and the target. That is, when the probe and the target co-occur in the list proper, this co-occurrence causes an associative link to be encoded between the two traces in memory. When the probe is re-activated at the end of the list, activation spreads from the probe to the target through this link, priming the target. This assumption is grounded in associatonism generally and various memory theories in particular (e.g., Gillund & Shiffrin, 1984), and in specific evidence that such associations are formed between neighbors in a list of random items (Nairne, 1983).

In the functional decay model, associative priming from the probe is implemented simply as a constant amount of activation added to the target. The effect of this priming is also illustrated in Figure 2, by the arrow labeled *priming*. Whereas the curves represent item activations immediately before the probe is presented, the elevated point at the head of the arrow represents the activation of a target when the probe is presented. (We have arbitrarily chosen the target to be the item at serial position 7, in the Fast condition.) The target item is shown to have much more activation than its neighbors, and also more activation than most recent list items (which would otherwise be the most active).

One other necessary model parameter, not represented in Figure 2, captures the differential effect (across the two presentation rates) of proactive interference due to previous trials. According to Equation 1, old items decay but

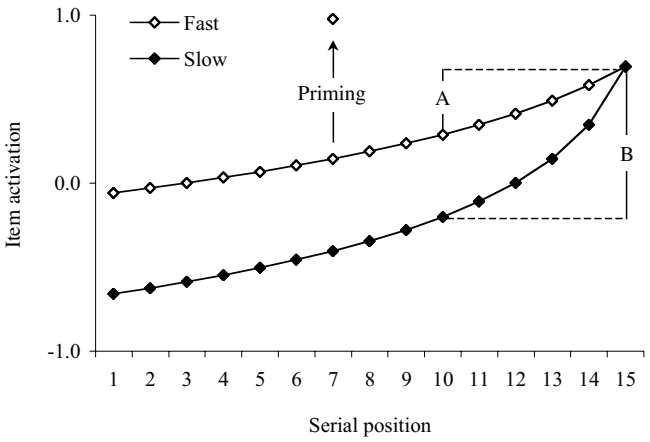


Figure 2: A snapshot of item activations from Equation (1) at the instant the full list of items has been presented. The arrow labeled *priming* indicates the magnitude of associative activation passing from probe to target (item 7, in this scenario) when the probe is presented. A and B are activation differences used to illustrate relative activation (see text).

nonetheless retain some activation well into the future, meaning that items from previous trials ("prior items") will continue to exist in memory as a source of proactive interference. However, the age of these prior items will differ across the Fast and Slow conditions, in that prior items in the Fast condition will retain more of their activation and hence cause greater proactive interference. This difference across the two conditions is captured in a prior-items parameter in the model. The parameter is implemented in terms of a single distracting element with an activation value that is estimated from the data.

The final step is to map item activations to the probability of retrieving a given target item. In ACT-R this mapping is the "soft-max" rule below, which predicts that the most active item has the highest probability of being retrieved (as we discussed above), but that other items can intrude from time to time. This rule determines the probability of retrieving a given item as a function of its activation relative to the activation of its distractors, and thus specifies the extent to which distractors interfere with the target.[3]

$$P_i = \frac{e^{A_i/s}}{\sum_j e^{A_j/s}}$$ Equation 2: Retrieval probability

P_i is the probability of retrieving i, the target, and A_i is the target's activation (from Equation 1). The quantity s represents the assumption that memory is susceptible to noise (i.e., transient fluctuations in activation levels).

Fitting Waugh and Norman (1965)

We can now describe how the model captures the interaction in Figure 1. One of the two basic patterns in the data is that late items are recalled better in the Slow condition than in the Fast condition. In the model, this effect results from late items having more *relative* activation in the Slow condition than in the Fast condition, as compared with distracting elements. By relative activation, we mean the difference in activation between the target and its distractors. Equation 2, which defines activation-based interference in ACT-R, predicts that the greater the difference in activation between the target and its distractors, the greater the probability of retrieving the target. In Figure 2, relative activation is represented by the differences A and B, which are differences in activation levels of items at two arbitrarily chosen serial positions (10 and 15). The difference A is between items 10 and 15 in the Fast condition, and the difference B is between items 10 and 15 in the Slow condition. When the target is a late item (i.e., 15), then the probability of retrieving it depends on the activation difference between it and earlier items (i.e., 10). This difference is larger in the Slow condition (distance B) than in the Fast condition (distance A). Thus, late items will be recalled more accurately in the Slow condition than in the Fast condition.

The second pattern in the data is that earlier items are recalled better in the Fast condition than in the Slow condition. In the model, this effect again results from relative activation, with target and distractor now reversing roles. With respect to the scenario in Figure 2, when the target is the earlier item (i.e., 10), then the distractor is the late item (i.e., 15). (Recall that although item 10 has less base-level activation than item 15, the associative priming illustrated in Figure 2 will compensate for this deficit and improve the probability of the target being retrieved.) The activation difference between target and distractor now favors the Fast condition, because item 10 in that condition faces a smaller activation deficit relative to its primary distractors (the later items in the list). Thus, earlier items will be recalled more accurately in the Fast condition than in the Slow condition.

Model parameters

The fit in Figure 1 depends on estimating three parameter values from 18 data points. Activation noise (s in Equation 2), was estimated at 0.19, a value in line with other applications of this equation (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998). The priming (associative activation) contributed by the probe digit was estimated at .83 units of activation. Finally, the prior-item activation (more specifically, the difference in prior item activation across the Fast and Slow conditions) was estimated at 1.1 units of activation. An Excel spreadsheet implementing the model and the two fits presented in this paper is at http://www.msu.edu/~ema/functionaldecay.

Fitting Peterson and Peterson (1959)

So far we have searched for and found evidence for decay in data on interference. Given our aim to integrate decay and interference functionally, we now turn to data on decay (Peterson & Peterson, 1959), and ask what role interference might play. The Brown-Peterson paradigm involves presenting a verbal item (e.g., a consonant trigram) and testing retention as a function of time. During the retention interval, verbal rehearsal is suppressed by a task like counting backwards. Figure 3 shows the data on recall accuracy from Peterson and Peterson (1959), Experiment 1. The x-axis shows retention interval and the y-axis shows proportion correct. The data show an even, negatively accelerating decline in accuracy with retention interval. This decline was interpreted to mean that maintenance of information in STM depended on active rehearsal, such that preventing rehearsal caused loss of information (Peterson & Peterson, 1959). Formal models of these data are not new (e.g., Baddeley, 1976, p. 130), but what is a novel integration is to explain these data on decay with the same processes that account for data on interference.

Our interpretation of the data in Figure 3 is that the current item (trigram) is represented in memory against a background of interfering items from previous trials. This kind of proactive interference has been demonstrated in a number Brown-Peterson studies (e.g., Dillon & Reid, 1969;

[3] Equation 2, termed the Chunk Choice Equation in Anderson and Lebiere (1998), plays a broader role in defining interference in our model than in ACT-R proper. In ACT-R, retrieval probability is a function both of Equation 2 and of a threshold parameter τ that specifies a minimum activation below which an item is invisible to the system. We assume no such threshold; the probability of retrieving an item is solely a function of that item's activation relative to the activation of distractors. We thus place greater emphasis on the role of interference from distracting information.

Keppel & Underwood, 1962; Wickens, Born, & Allen, 1963). In our model, this proactive interference plays the same role here as prior-item interference did in fitting the Waugh and Norman data. An element of this mental clutter can intrude when the system attempts to retrieve the target — and is more likely to, the more the target has decayed. Thus, relative activation is again the factor determining retrieval accuracy. Here, however, relative activation is a factor between trials only, whereas in the probe-digit model it was a factor between and within trials. The only other change to the model was to remove associative priming as a source of activation for the target, reflecting the absence of a specific probe in the Brown-Peterson paradigm.

As shown in Figure 3, the model again fits closely, with $RMSD$ = .027 and R^2 = .977. Fitting the model required estimating two parameters from six data points. Activation noise s (Equation 2) was estimated to be .34, which is again in the range used in other ACT-R models (Anderson et al., 1998). Prior item activation was estimated at -2.31 units.

Note that in fitting the Peterson and Peterson data we carried over the decay rate from the Waugh and Norman model (-.5). This illustrates the value of incorporating interference in a model of decay. A simple power-law decay model, without interference as a factor, is $P_i = a+bT_i^{\ d}$, with P_i the probability of retrieving item i, T_i the age of i, d the decay rate, and a and b free parameters. Fitting this model to the Peterson and Peterson data produces measures of fit $RMSD$ = .029 and R^2 = .973, and parameter values a = -19.8, b = 20.7, and d = -.14. Of particular interest is the decay rate, -.14. This deviates substantially from the value of -.5 that we carried over from the Waugh and Norman model, and from many ACT-R models before that (Anderson & Lebiere, 1998). Thus, although the *apparent* decay rate may differ from situation to situation, we propose that what varies is not the *architectural* decay rate but the background level of interference, which is situation-dependent and thus a more plausible source of variation. Importantly, this variable can also be estimated quantitatively, for example by counting the number of trials preceding the trial of interest (Keppel & Underwood, 1962).

In sum, if interference is indeed a primary mechanism of forgetting, then it would be odd if it played no role in forgetting in the Brown-Peterson paradigm. Our analysis suggests that decay by itself cannot cause forgetting — forgetting arises because decay takes place *relative to* background interference in memory.

Discussion

We propose that decay and interference are functionally related processes — decay of distractors mitigates the extent to which they interfere with the target. Playing out the consequences of this proposal makes functional sense of an empirical result that has lain largely dormant for a generation, absent the right theoretical framework in which to interpret it. We have also formalized the integration of decay and interference in another sense, by fitting the Waugh and Norman data set and its "opposite," the Peterson and Peterson data, with the same model. These model-fitting successes converge with our functional logic to argue that decay and interference must both operate in memory.

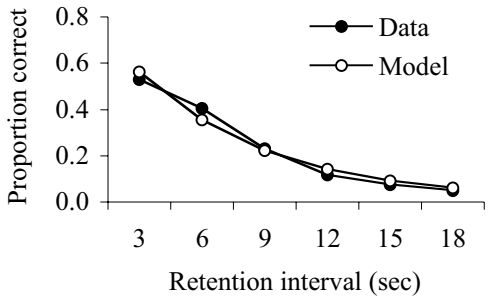


Figure 3: Data from Peterson and Peterson (1959), Experiment 1, and the fit of the functional decay model.

We began by hinting that if decay is important to the functionality of memory, then its effects should be found pervasively in behavioral data. Indeed, functional decay models similar to the one presented here have been applied to diverse domains. These include task switching (Altmann & Gray, 2002) and the time course of Stoop interference (Altmann & Davidson, 2001). Thus, the claim that decay is pervasive has some backing beyond our excursion here to the headwaters of the debate over forgetting mechanisms.

There are a number of caveats on the current work that will be important to address in the future. First, the Waugh and Norman interaction, though it seems visible in other probe digit data (Norman, 1966), needs to be replicated before we invest more in interpreting it. Second, our spreadsheet model needs to be implemented as a running simulation, to test whether we have missed important interactions among processes. Third, the model makes specific predictions about which distractor items should intrude in what proportions; later items should intrude more often, because they are more active. These predictions clearly need to be tested.

We should also note that the construct of relative activation underlying our model has other expressions in memory theory, such as the discrimination ratio (Baddeley & Hitch, 1993) and temporal distinctiveness (Neath, 1993). We would argue, however, that the grounding of the current model in ACT-R anchors it more directly to observable environmental processes. ACT-R is premised on the notion that memory is a mirror reflecting patterns of information need imposed by the environment. Thus, for example, retrieval frequency is a predictor of activation because it is also a predictor of impending need for that item. Consistent with the functional interpretation of decay, we favor a functional interpretation of memory generally, in which quantities like activation reflect the tasks that the memory system accomplishes for us.

Finally, we should emphasize that our claims about the importance of decay are not meant to conflict with the idea that interference is the dominant cause of loss of retrievable information in memory. Wherever they have been isolated, including in the Waugh and Norman data, the effects of decay are quite small (c.f., Reitman, 1974) compared to the effects of interference. Indeed, a small effect of decay is all that is functionally necessary to tilt retrieval probability toward the target, particularly when strategic memory processes like rehearsal are available for the system to manipulate target activation.

Acknowledgements

We thank the conference reviewers for valuable comments. CDS received support from ONR grant N00014-01-1-0321.

References

Altmann, E. M., & Davidson, D. J. (2001). An integrative approach to Stroop: Combining a language model and a unified cognitive theory. In *Proceedings of the twenty-third annual meeting of the Cognitive Science Society* (pp. 21-26). Hillsdale, NJ: Erlbaum.

Altmann, E. M., & Gray, W. D. (2002). Forgetting to remember: The functional relationship of decay and interference. *Psychological Science*, *13*, 27-33.

Anderson, J. R. (2000). *Learning and memory: An integrated approach* (2nd ed.). New York: Wiley.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory & Language, 38*, 341-380.

Anderson, J. R., & Lebiere, C. (Eds.). (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703-719.

Ashcraft, M. H. (2002). *Cognition* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Baddeley, A. D. (1976). *The psychology of memory*. New York: Basic Books.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Boston: Allyn & Bacon.

Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, *21*, 146-155.

Baddeley, A. D., & Scott, D. (1971). Short term forgetting in the absence of proactive interference. *Quarterly Journal of Experimental Psychology, 23*, 275-283.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*, 12-21.

Dillon, R. F., & Reid, L. S. (1969). Short-term memory as a function of information processing during the retention interval. *Journal of Experimental Psychology*,*81*, 261-269.

Ekstrand, B. R. (1972). To sleep, perchance to dream. In C. P. Duncan & L. Sechrest & A. W. Melton (Eds.*), Human memory: Festschrift in honor of Benton J. Underwood* (pp. 59-82). New York: Appleton-Century-Crofts.

Estes,W.K.(1955).Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145-154.

Fuster, J. M. (1995). *Memory in the cerebral cortex*. Cambridge, MA: MIT Press.

Galotti, K. M. (1999*). Cognitive psychology in and out of the laboratory* (2nd ed.). New York: Wadsworth.

Gillund, G., & Shiffrin, R (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1-67.

Hintzman, D. L. (1978). *The psychology of learning and memory*. San Francisco: W. H. Freeman.

Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528-551.

Keppel, G., Postman, L., & Zavortink, B. (1968). Studies of learning to learn: VIII. The influence of massive amounts of training upon the learning and retention of paired-associate lists. *Journal of Verbal Learning and Verbal Behavior*, *7*, 790-796.

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior*, *1*, 153-161.

Landauer, T. K. (1975). Memory without organization: Properties of a model with random storage. *Cognitive Psychology*, *7*, 495-531.

Loess, H. & Waugh, N. C. (1967). Short-term memory and intertrial interval. *Journal of Verbal Learning and Verbal Behavior, 6*, 455-460.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476-490.

Luria, A. R. (1968). *The mind of a mnemonist*. New York: Basic Books.

Massaro, D. W. (1970). Perceptual processes and forgetting in memory tasks. *Psychological Review*, 77, 557-567.

McGeogh, J. A., & Irion, A. L. (1952). *The psychology of human learning*. New York: MacKay.

Murdock, B. B. (1992). Serial organization in a distributed memory model. In A. F. Healy (Ed.), *Essays in honor of William K. Estes* (1; pp. 201-225). Hillsdale, NJ: Erlbaum.

Nairne, J. S. (1983). Associative processing during rote rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9, 3-20*.

Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Memory & Cognition, 21*, 689-698.

Norman, D. A. (1966). Acquisition and retention in short-term memory. *Journal of Exp. Psychology, 71,* 369-381.

O'Reilly, R. C., & Munakata, Y. (2000*). Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193-198.

Postman, L. (1971). Transfer, interference, and forgetting. In J. W. Kling & L. Riggs (Eds.), Psychology (3rd ed., pp. 1019-1132). New York: Holt, Rinehart, and Winston, Inc.

Reed, S. K. (2000). *Cognition: Theory and applications* (5th ed.). Belmont, CA: Wadsworth.

Reitman, J. S. (1974). Without surreptitious rehearsal, information in short-term memory decays. *Journal of Verbal Learning and Verbal Behavior*, *13*, 365-377.

Richman, H. B., Staszewski, J. J., & Simon, H. A. (1995). Simulation of expert memory using EPAM IV. *Psychological Review*, *102*, 305-330.

Turvey, M. T., Brick, P., & Osborn, J. Proactive interference in short-term memory as a function of prior-item retention interval. *Quarterly Journal of Experimental Psychology, 22,* 142-147.

Waugh, N. C. & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*, 89-104.

Wickelgren, W. A. (1977). *Learning and memory*. Englewood Cliffs, NJ: Prentice Hall.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*, 440-445.

Preventing Catastrophic Interference in Multiple-Sequence Learning Using Coupled Reverberating Elman Networks

Bernard Ans*, Stéphane Rousset*, Robert M. French† & Serban Musca*

**Experimental Psychology Laboratory*
Université Pierre Mendès-France, Grenoble 2 – CNRS UMR 5105
BP 47, 38040 Grenoble cedex 09, France
email: Bernard.Ans@upmf-grenoble.fr

† Quantitative Psychology and Cognitive Science
Université de Liège (Bât B32), Sart Tilman
4000 Liège, Belgique
email: rfrench@ulg.ac.be

Abstract

Everyone agrees that real cognition requires much more than static pattern recognition. In particular, it requires the ability to learn *sequences* of patterns (or actions) But learning sequences really means being able to learn *multiple* sequences, one after the other, without the most recently learned ones erasing the previously learned ones. But if catastrophic interference is a problem for the sequential learning of individual patterns, the problem is amplified many times over when multiple *sequences* of patterns have to be learned consecutively, because each new sequence consists of many linked patterns. In this paper we will present a connectionist architecture that would seem to solve the problem of multiple sequence learning using pseudopatterns.

Introduction

Building a robot that could unfailingly recognize and respond to hundreds of objects in the world – apples, mice, telephones and paper napkins, among them – would unquestionably constitute a major artificial intelligence *tour de force*. But everyone agrees that real cognition requires much more than static pattern recognition. In particular, it requires the ability to learn *sequences* of patterns (or actions). This was the primary reason for the development of the simple recurrent network (SRN, Elman, 1990) and the many variants of this architecture.

But learning sequences means more than being able to learn a single, isolated sequence of patterns: it means being able to learn *multiple sequences*, one after the other, without the most recently learned ones erasing the previously learned ones. But if catastrophic interference – the phenomenon whereby new learning completely erases old learning – is a problem with static pattern learning (McCloskey & Cohen, 1989; Ratcliff, 1990), the problem is amplified many times over when multiple sequences of patterns have to be learned consecutively, because each sequence consists of many new linked patterns. What hope is there for a previously learned sequence of patterns to survive after the network has learned a new sequence consisting of many individual patterns?

In this paper, we will present a connectionist architecture that solves the problem of multiple sequence learning.

Catastrophic interference

The problem of catastrophic interference (or forgetting) has been with the connectionist community for well over a decade now (McCloskey & Cohen, 1989; Ratcliff, 1990; for a review see Sharkey & Sharkey, 1995). Catastrophic forgetting occurs when newly learned information suddenly and completely erases information that was previously learned by the network, a phenomenon that is not only implausible cognitively, but disastrous for most practical applications. The problem has been studied by numerous authors over the past decade (see French, 1999 for a review). The problem is that the very property – a single set of weights to encode information – that gives connectionist networks their remarkable abilities of generalization and graceful degradation in the presence of incomplete information are also the root cause of catastrophic interference (see, for example, French, 1992).

Various authors (Ans & Rousset, 1997, 2000; French, 1997; Robins, 1995) have developed systems that rehearse on *pseudo*-episodes (or pseudopatterns), rather than on the real items that were previously learned. The basic principle of this mechanism is when learning new external patterns to interleave them with *internally-generated* pseudopatterns. These latter patterns, self-generated by the network from *random* activation, reflect (but are not identical to) the previously learned information. It has now been established that this pseudopattern rehearsal method effectively eliminates catastrophic forgetting.

A serious problem remains, however, and that is this: cognition involves more than being able to sequentially learn a series of "static" (non-temporal) patterns without interference. It is of equal importance to be able to serially learn many of *temporal sequences of patterns*. We will propose an pseudopattern-based architecture that can effectively learn multiple temporal patterns consecutively.

The key insight of this paper is this:

> *Once an SRN has learned a particular sequence, each pseudopattern generated by that network reflects the entire sequence (or set of sequences) that has been learned.*

From which our key result follows:

> When learning a new sequence, simple rehearsal with these sequence-encoding pseudopatterns will prevent catastrophic forgetting of the previously learned sequence(s).

We will use a connectionist architecture using two coupled "auto-associative recurrent networks (AARN)" (Maskara & Noetzel, 1992, 1993; Cleeremans & Destrebecqz, 1997) that pass information back and forth to each other by means of pseudopatterns. We will refer to auto-associative recurrent networks as Reverberating SRNs (RSRN), in order to emphasize the manner in which they use pseudopatterns to eliminate catastrophic interference in multiple sequence learning.

The remainder of this paper is organized as follows. We will briefly review the standard dual-network pseudopattern solution to the problem of catastrophic forgetting in static pattern learning. We will then show (Simulation 1) that multiple-sequence learning is particularly susceptible to catastrophic forgetting. We will then show how our pseudopatterns-based dual-network architecture can be used to effectively overcome catastrophic interference in multiple sequence learning.

Overcoming catastrophic interference with pseudopatterns

Before discussing catastrophic interference in multiple-sequence learning, we need to briefly describe what pseudopatterns are and how they can be used to reduce catastrophic interference in the simpler case of static pattern learning.

Assume we have a three-layer feedforward network that learns a number of binary patterns drawn from some distribution. Assume, thereafter, that these patterns are no longer available. How can one determine, even approximately, what the network has learned? Answer: By "bombarding" the input of the network with *random* binary vectors and collecting the associated output vectors. Each input-output pair of vectors produced in this way will constitute a *pseudopattern* that will be a reflection of the function previously learned by the network.

The basic idea to use pseudopatterns to reduce catastrophic interference is due to Robins (1995). It works as follows. The network learns a first set of patterns, $\{P_i\}$. Then before it begins to learn a second set of patterns, $\{Q_i\}$, noise is fed through the network to produce a set of pseudopatterns, $\{\psi_i\}$, as above. These pseudopatterns are added to the new patterns to be learned and the network trains on this larger set until all of the new patterns, $\{Q_i\}$, are learned to criterion. When the network is tested on the originally learned patterns, $\{P_i\}$, it has not forgotten them catastrophically. Had there been no pseudopatterns mixed in with the new patterns that were learned, the network would have completely forgotten the originally learned patterns.

Dual-Network Architectures

But where are these internally generated pseudopatterns stored so that they can be interleaved with the new patterns? One answer is to generate them on the fly in a separate network (French, 1997; Ans & Rousset, 1997, 2000). Let us consider one way that a single new pattern, Q, might be learned in this dual-network model. Assume that new patterns are learned by NET 1 (this can be considered to be the "Performance Network"), while previously learned patterns are stored in NET 2 (which could be considered to be the "Storage Network"). NET 1 learns Q as follows:

> i) pattern Q is input to NET 1, which modifies its weights once; ii) noise is input to NET 2, which generates a pseudopattern; iii) this pseudopattern is presented to NET 1, which modifies its weights once; iv) if Q has been learned to criterion, stop; otherwise go to i).

We call this the "awake state". Once the output error for pattern Q has dropped below criterion, we transfer the information in NET1 to NET 2 as follows:

> Loop N times: i) noise is input to NET 1, which generates a pseudopattern; ii) this pseudopattern is input to NET 2, which modifies its weights once.

We call this phase, when information is transferred to NET 2, the "sleep state." This will be the basis of all learning in the dual-network framework described in this paper.

"Reverberating" backpropagation

A reverberating backpropagation (RBP) network is a standard three-layer network that has a built-in autoassociator ("reverberator") for the input of the patterns to be learned (Ans & Rousset, 1997, 2000). We have shown this network in an "unfolded" manner (Figure 1). In this visualization, the output layer is divided into the "autoassociative" nodes for the input component of the patterns to be learned (on the left in Figure 1) and the "target" nodes for the targets of the patterns to be learned (on the right in Figure 1).

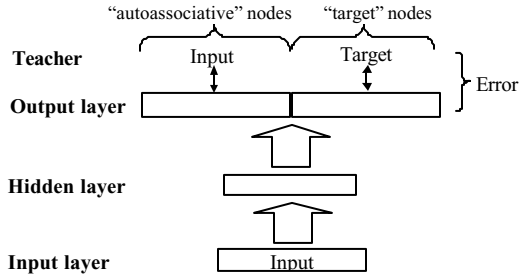


Figure 1. An RBP network

Assume the network is to learn a pattern $P: I \rightarrow T$, consisting of an input, I, and a target, T. I is presented on input and is sent through the network to the output layer. For all of the nodes in the output layer, an error is calculated. For the "autoassociative" output nodes,

the error is based on the difference between the network output and the original input, *I*, whereas for the "target" nodes, this error is based on the difference between the "heteroassociative" output and the desired target *T*. The errors associated with both the autoassociator and the target output nodes are then backpropagated through the network in order to change the network's weights.

"Attractor" pseudopatterns in an RBP Network

To generate pseudopatterns in a reverberating network, a random input i_ψ is presented to the input layer of the network and fed through to the output layer. The activation values of the "autoassociative" nodes in the output layer (nodes on the left of the output layer in Fig. 1) constitute a new input, i'_ψ, which is then sent through the network (the activation values on the "target" nodes in the output layer are ignored). This produces a pattern of activation on the autoassociative output nodes, i''_ψ, which is then presented to the input nodes of the network, and so on. After a number of reverberating cycles through the network, a final "reverberated" input, i^R_ψ, is sent through the network and the activation vector of all the output nodes (the "autoassociative" and the "target" output nodes), o_ψ, is used to produce a pseudopattern $\psi : i^R_\psi \rightarrow o_\psi$. The significant advantages of using an input autoassociator with a feedforward backpropagation network have been shown elsewhere (Ans & Rousset, 1997, 2000). Suffice it to say that the reverberation process transforms a pure random input, i_ψ, into an *attractor* of the system, i^R_ψ. An "attractor" pseudopattern produced provides a much better reflection of the old patterns than a pseudopattern produced from simple random noise on input. It is this reverberation technique that is largely responsible for the power of this technique.

Reverberating Simple Recurrent Networks (RSRN)

We will assume that the reader is familiar with the design of a Simple Recurrent Network (SRN, Elman, 1990). An RSRN (Figure 2) works very much like a standard SRN (Maskara & Noetzel, 1992, 1993; Cleeremans & Destrebecqz, 1997). Just as the RBP network involves adding "autoassociative" nodes to the output layer of a BP network, a reverberating SRN involves adding "autoassociative" nodes to the output layer of an SRN. The full input to the network consists of the "standard" input, i.e., the input from the sequence item, *S*(*t*), and the "context" input, *H*(*t* -1), which is the activation vector of the hidden layer associated with the previous item in the sequence, *S*(*t* -1).

"Attractor" pseudopatterns in an RSRN

The principle is identical to pseudopattern generation in an RBP network. Crucially, once the RSRN has learned a sequence of items, each "reverberated" pseudopattern generated by it reflects *the entire sequence learned by the recurrent network*. Each pseudopattern can be thought of as a very compact representation of the entire previously learned sequence.

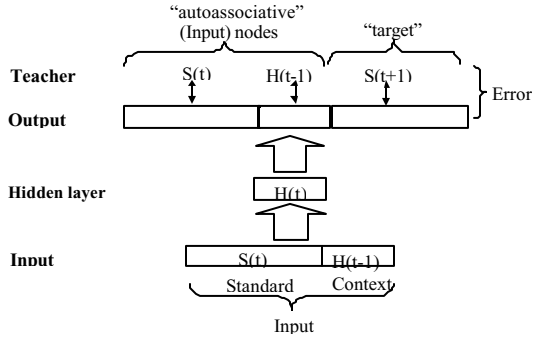


Figure 2. A Reverberating SRN (RSRN)

A dual-network architecture with self-refreshing memory to overcome catastrophic forgetting in multiple sequence learning

Ans & Rousset (1997, 2000) proposed a reverberating dual-network architecture with a self-refreshing memory to avoid catastrophic forgetting in static pattern learning. The basic dual-network architecture consists of two coupled RBP networks, denoted NET 1 and NET 2. NET 1 is the primary network that interfaces with the environment and that learns the new patterns. The secondary network, NET 2, is a "storage" network because information initially learned in NET 1 will ultimately be transferred to NET 2.

The basic principles of RSRN dual-network learning, where each of the networks is an RSRN, are virtually identical to those underlying dual-networks composed of RBP networks. Dual-network RSRNs are designed to learn multiple sequences of patterns.

Assume we have two identical RSRNs, NET 1 and NET 2. New sequences will be learned only by NET 1, while NET 2 will store the previously learned information. The learning procedure is similar to that of the basic RBP dual-network. A sequence, $S = S(0), S(1), ..., S(n)$ is presented to NET 1. The network makes a single pass through the entire sequence, updating its weights once for each item in the sequence. This defines one learning "epoch" corresponding to one presentation of all items in the sequence in order. Next, NET 2 generates a number of pseudopatterns (e.g., 10 per learning epoch). These pseudopatterns are *close to attractor states* of NET 2,

which makes them particularly good vehicles for information transfer from NET 2 to NET 1. For each NET 2 pseudopattern, NET 1 performs one feedforward-backpropagation learning pass. Once this is completed, a new learning epoch starts and NET 1 makes another pass through the sequence *S*. NET 2 generates new pseudopatterns, each of which is learned for one feedforward-backpropagation pass by NET 1. And so on. This is the awake state for an RSRN dual-network.

It is extremely important to notice that each pseudopattern generated by NET 2 is a *static input-output pattern that represents a dynamic state* (i.e., the previously learned sequence or sequences). This is what gives this system its power: there is no need to attempt to reproduce an entire pseudo-sequence that will then be interleaved with the new sequence being learned. Rather, we only need to interleave with the new sequence to be learned *non-temporal pseudopatterns*, each of which reflects (at least partially) the information in the entire previously learned temporal sequence (or sequences).

To transfer the sequence newly learned by NET 1 to NET 2, we again make use of pseudopatterns. This time the pseudopatterns are generated by NET 1 and learned by NET 2. For each pseudopattern generated by NET 1, NET 2 performs a single feedforward-backpropagation learning pass.

Overview of Simulations

We will present two simulations. The first will demonstrate the severity of catastrophic interference in multiple sequence learning in standard SRNs. The second will demonstrate that interleaving pseudo*patterns* (reflecting the whole previously learned sequence) with the new sequence effectively eliminates catastrophic interference.

A standard SRN network was used for the first simulation demonstrating the severity of catastrophic interference in multiple sequence learning. In the second simulation, a dual-network architecture consisting of two coupled RSRNs will be used. Each RSRN has an input layer with 100 "standard" input units (corresponding to the size of the items, $S(t)$, in the sequence) and 50 "context" units. The hidden layer consists of 50 units. The output layer consists of 150 "autoassociative" inputs that are identical to the input layer plus 100 "target" units (Figure 2).

Learning a given sequence consists of presenting it repeatedly to the network until each item in the sequence can predict the subsequent item with a predefined degree of precision. The network weights are updated by backpropagation once per presentation of each sequence item. A cross-entropy error function is used (Hinton, 1989; Plaut, McClelland, Seidenberg & Patterson, 1996) with a learning rate of 0.01, a momentum of 0.5 and a 1.0 bias term. All weights are randomly initialized between –0.5 and 0.5.

To create the "attractor" pseudopatterns, noise on input is "reverberated" 5 times before the actual pseudopattern that will be used is created. Ten pseudopatterns from NET 2 are interleaved with each epoch of the sequence learned by NET 1. During transfer of the information from NET 1 to NET 2, 10^4 pseudopatterns are used.

Measuring learning and forgetting

For each item, $S(t)$, of the sequence fed forward through the network, we calculate the difference between the activation values of "target" units in the output layer and the desired target item in the sequence, $S(t+1)$. We calculate this difference for each of the 100 "target" output units and count the number of output units for which the absolute value of this difference is above the learning criterion of 0.1. So, for example, assume a given item, $S(t)$, in the sequence is sent through the network. If, on the output layer, 14 of the "target" output units differ from the corresponding units of $S(t+1)$ by more than 0.1, we say that there are 14 "incorrect" units. A sequence is considered to have been learned if, for each of its elements, $S(t)$, the network produces a vector of "target" output values, each of which is within 0.1 of the corresponding element of $S(t+1)$. The overall measure of how well the network has learned (or forgotten) a sequence after a given number of learning epochs will be the total number of incorrect units over all items of the sequence.

Simulation 1: Catastrophic forgetting in multiple sequence learning

To illustrate the severity of catastrophic forgetting in multiple sequence learning, we will consider two sequences, *A* and *B*, and have an SRN attempt to learn them sequentially. The sequences are constructed as follows. Twenty-two distinct random binary vectors of length 100 are created. Half of these vectors are used to produce the first ordered sequence of items, *A*, denoted by $A(0)$, $A(1)$, …, $A(10)$. The remaining 11 vectors are used to create a second sequence of items, *B*, denoted by $B(0)$, $B(1)$, …, $B(10)$. In order to introduce a degree of ambiguity into each sequence (so that a simple BP network would not be able to learn them), we modify each sequence so that $A(5) = A(8)$ and $B(1) = B(5)$. First, sequence *A* is completely learned by the network. Then sequence *B* is learned and, during the course of learning, we monitor at regular intervals how much of sequence *A* has been forgotten by the network.

Fig. 3a shows the progression of the network's learning of sequence *B*. The number of incorrect units for each serial position of the sequence, as defined above, is shown. As expected, it is harder for the network to learn $B(2)$ and $B(6)$ since their immediate predecessors are identical and, in order to distinguish them, the network needs to additionally take into consideration the context of the preceding items in the sequence. After 450 epochs, the network has completely learned the entire sequence.

Then, during learning of sequence *B*, we

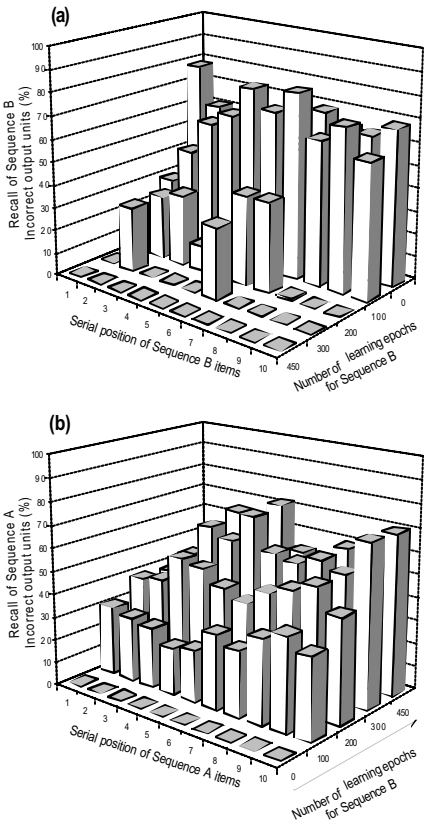


Figure 3. Catastrophic forgetting in an SRN during multiple sequence learning. **(a)**: Learning of sequence *B* (after having previously learned sequence *A*). By 450 epochs (an epoch corresponds to one pass through the entire sequence), sequence *B* has been completely learned. **(b)**: The number of incorrect units for sequence *A* during learning of sequence *B*. After 450 epochs, the SRN has, for all intents and purposes, completely forgotten the previously learned sequence *A*.

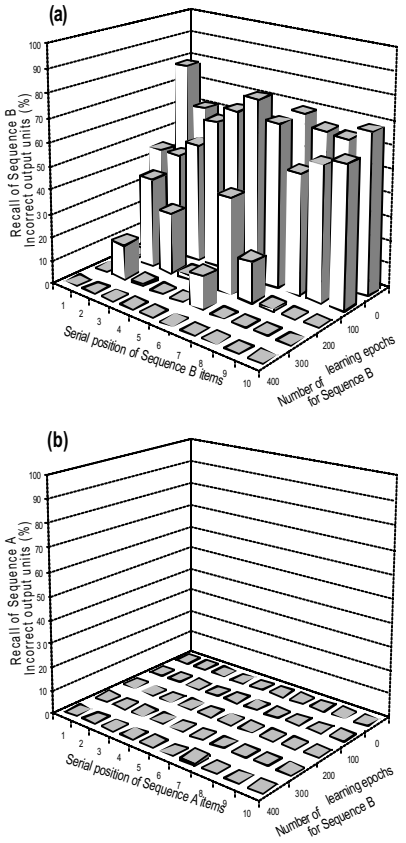


Figure 4. Recall performance for sequences *B* and *A* during learning of sequence *B* by a dual-network RSRN. **(a)**: By 400 epochs, the second sequence *B* has been completely learned. **(b)**: The previously learned sequence *A* shows virtually no forgetting. Catastrophic forgetting of the previously learned sequence *A* has been completely overcome.

monitored the forgetting of the previously learned sequence, *A* (Fig. 3b). Initially (i.e., before the network began learning sequence *B*), it can be seen that sequence *A* was completely learned. But very early on, as sequence *B* is being learned, the network's memory of sequence *A* is overwritten. By 5 epochs after beginning to learn the sequence *B* (not shown in Fig. 3b), the network gets an average of 40% of the units of the items of sequence *A* wrong. By 250 epochs, the network's performance on sequence *A* is essentially no better than chance and, by 450 epochs, sequence *A* is completely forgotten. In short, learning sequence B causes severe catastrophic forgetting of sequence A.

Simulation 2: Catastrophic forgetting is overcome with pseudopatterns

An RSRN dual-network architecture was used with the parameters indicated above. Both networks are initialized to random weight settings between −0.5 and 0.5. NET 1 then completely learns sequence *A* and then generates 10^4 pseudopatterns in order to transfer this learning to NET 2. (There are 2^{150} possible distinct states for the input layer of each network, and hence, there is a very little possibility that the random binary vectors used to produce the "attractor pseudo-input" would be actual input patterns already seen by the network.)

Now, NET 1 begins to learn sequence *B*. After each learning epoch (consisting of the entire sequence of items in *B)*, NET 1 receives 10 pseudopatterns from NET 2 and does one feedforward-backpropagation pass for each of them. (The number of pseudopatterns is not related to the length of the previously learned sequences and can be varied.)

Fig. 4a shows that the NET 1 does, in fact, learn sequence *B* completely by 400 epochs. In other words, for all items in sequence *B,* all of the units in the network output are within 0.1 of the desired target

output. Notice that the sequence items $B(2)$ and $B(6)$ are learned more slowly by the network. This was, of course, expected since these two items are preceded by identical items, hence creating ambiguity having to be solved by the temporal context. During learning of sequence *B*, we tested the performance of the network on all of the items of sequence *A*. Fig. 4b shows that *there is virtually no forgetting of sequence A as the network learns sequence B*. In short, catastrophic forgetting has been completely overcome by the coupled system of RSRNs using pseudopattern information transfer.

Concluding remarks

We have shown that the reverberating dual-network architecture, originally proposed earlier to overcome catastrophic forgetting in non-temporal pattern learning (Ans & Rousset, 1997, 2000) can be generalized to sequential learning of multiple temporal sequences. The basic principle of interleaving internally-generated pseudopatterns from a long-term storage network with patterns from the environment being learned by a second network has been developed elsewhere (Ans & Rousset, 1997, 2000; French, 1997; Robins, 1995).

When learning multiple sequences of patterns, it turns out that interleaving simple input-output pseudopatterns, each of which reflect *the entire previously learned sequence(s)*, reduces (or eliminates entirely) forgetting of the initially learned sequence(s).

Further, we demonstrate the power of a network architecture that allows us to produce "reverberated" pseudopatterns that are, in reality, attractors of the entire network and therefore reflect, in a highly compressed manner, the previously learned sequences.

We have demonstrated a technique that effectively allows multiple sequences to be learned consecutively. Of course, these networks can be made to forget gradually. This gradual forgetting depends on the size of the learning network and on the number, the overlap, the length and the complexity of the successively learned sequences. But the problem of sudden, "catastrophic" forgetting of previously learned sequences caused by learning a new sequence of patterns would seem to have been overcome.

Acknowledgments

This research was supported in part by a research grant from the European Commission (HPRN-CT-1999-00065) and by the French government (CNRS UMR 5105).

References

Ans, B. & Rousset, S. (1997) Avoiding catastrophic forgetting by coupling two reverberating neural networks. *C.R. Acad. Sci. Paris, Life Sciences*, 320, 989–997.

Ans, B. & Rousset, S. (2000) Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science*, 12, 1–19.

Cleeremans, A. & Destrebecqz, A. (1997). Incremental Sequence learning. In *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society*, NJ: LEA, 119-124

Elman, J.L. (1990) Finding structure in time. *Cognitive Science*, 14, 179–211.

French, R.M. (1992) Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science,* 4, 365–377.

French, R.M. (1997) Pseudo-recurrent connectionist networks: An approach to the 'sensitivity-stability' dilemma. *Connection Science,* 9, 353–379.

French, R.M. (1999) Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3, 128–135.

Hinton, G.E. (1989) Connectionist learning procedures. *Artificial Intelligence*, 40, 185–234.

Maskara, A., & Noetzel, A. (1992). Forced simple recurrent neural network and grammatical inference. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 420-425, NJ: LEA.

Maskara, A. & Noetzel, A. (1993). Sequence learning with recurrent neural networks. *Connection Science, 5,* 139-152.

McClelland, J.L., McNaughton, B.L. & O'Reilly, R.C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review,* 102, 419–457

McCloskey, M. & Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 24. New York: Academic Press, pp. 109–165.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115.

Ratcliff, R. (1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97, 285–308.

Robins, A.V. (1995) Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7, 123–146.

Sharkey, N.E. & Sharkey, A.J.C. (1995) An analysis of catastrophic interference. *Connection Science*, 7, 301–329.

A Cognitive Account of Situated Communication

Rita B. Ardito (ardito@psych.unito.it)
Bruno G. Bara (bara@psych.unito.it)
Enrico Blanzieri (blanzier@psych.unito.it)

Centro di Scienza Cognitiva, Universita' di Torino
via Po 14, 10123 Torino, Italy

Abstract

Communication is situated and it is situated in the environment, in the perception the participants have of the environment, and also in the mental representations they privately retain. This work addresses the problem of explaining the interaction between state of the physical world and mental states of actors involved in a communicative exchange. We have the goal of integrating physical world representations, such as space representation, physical co-presence of the actors and physical nature of the communication channel, with a cognitive account of communicative phenomena. We introduce the concepts of *scene*, *situation* and *scenery* for elaborating a theory for situated communication and give an account of the interaction between mental states involved in communication and the subjective representations of the state of the world.

1. Introduction

Communication and physical action are strictly connected. Since Austin's approach to Pragmatics based on speech-acts (Austin, 1962), the things one does through (not only) words, comprehension and generation of communication have been linked to the general framework of action. Conversely, the actions performed by an actor are better understood if considered as situated (Suchman, 1987; Clancey, 1998). Situations involve the physical world as well as the social world when other actors are present. The relationship between communication and action is complex. Actions performed in the physical social world are not necessarily communicative: in fact, they can either facilitate or impair communication itself, through a modification of the situational context. On the other hand, communication between actors can modify their intentions about performing actions in the physical world. Finally, the state of the physical world influences the possibility of performing actions and communicative acts.

Current theories on communication such as Relevance Theory (Sperber and Wilson, 1986) and Cognitive Pragmatics Theory (Airenti, Bara and Colombetti, 1993a; 1993b) deal with accurate descriptions of mental states and cognitive functions involved in communication. However, their applications to concrete situations is not straightforward because they do not give an account of how these mental states interact with the state of the physical world in order to produce the observed natural cases. In particular, Cognitive Pragmatics, developing an intuition of Wittgenstein (1953), introduces the notion of behaviour game, namely a shared plan between the actors, but it gives no accounts of *how* the behaviour game is played in a concrete situation.

This work addresses the problem of explaining the interaction between the state of the physical world and the mental states of actors involved in a communicative exchange. We have the goal of integrating physical world representations, such as space representation, physical co-presence of the actors and physical nature of the communication channel, with a cognitive account of communicative phenomena. Our work is consistent with the perspective outlined by Clancey (1998) of shifting from an abstract and disembodied concept of cognition to a situated one. In one case, this approach amounts to exploiting the properties of abstract shared plans in communication, in order to enrich them with the features of concrete situated actions.

Our proposal is based on the novel concept of *scenery* that we define using the notion of a shared plan, namely a *behaviour game* in Cognitive Pragmatics terms. For an actor, to know the shared behaviour game is crucial in order to grasp the meaning of a communicative act. The *scenery* relates at the representational level context and at the level of behaviour games in terms of preconditions and possible actions.

The paper is organised as follows: Section 2 discusses situated communication in the framework of situated cognition, situated action and Cognitive Pragmatics Theory; Section 3 introduces the concepts of scene, situation, scenery and scenario; Section 4 proposes a cognitive account of situated communication and in the last Section we draw some conclusion.

2. Situated communication

Communication is obviously situated and it is situated in the environment, in the perception the participants have of the environment, and also in the mental representations they privately entertain. The Exs. 1-10 report a short fiction story that describes a long series of interactions between Alice, a professor, and Bob, a Ph.D student who aims to meet to discuss the draft of his thesis proposal. Alice and Bob communicate in a wide range of different physical environments (roads, corridors, office, elevator, cafeteria)

using different media (phone, e-mail), manipulating different objects (phones, handles, chairs, buttons, cups of coffee), meeting different people (maintenance people, cafeteria staff, a colleague). Moreover, during their interactions Alice and Bob are affected by the subjective perception they have of environments, media, objects and people, by their representations and finally by their representations of their own representations (meta-representations).

The process of communication may be theoretically described in terms of shared plans. This approach is assumed in the work of Airenti, Bara & Colombetti (1993a; 1993b) while shared plans have been proposed by Grosz & Kraus (1996). The theory of Cognitive Pragmatics is based on the idea that co-operation is the key element for the communicative interaction. It assumes that two people who communicate co-operate, and their actions are at least partly shared in order to reach a common goal. The plans by which two interacting people base their co-operation are called *behaviour games*. Plans can be seen as trees of intentions, where the leaves are specified either as terminal, precise actions, or as intentions made specific according to the context.

Behaviour games enable people to select the correct meanings to be assigned to the linguistic and extra-linguistic moves of each participant in a communicative exchange. Therefore, to understand the actor's meaning it is necessary to infer the behaviour game the actor is referring to.

Another concept adopted by Cognitive Pragmatics Theory, and which plays an important role, is *shared belief*, namely a belief that a single individual thinks of sharing with the person he is talking to. In symbols, a shared belief can be represented as follows:

$$SH_{AB}\, p$$

meaning that agents A and B share the belief p.

It is worth emphasising that shared belief is a subjective mental state. In other words, it may happen that A believes p to be shared by B and A, whereas B does not believe p to be shared by A and B. Shared beliefs are mental states which allow each actor to take for granted the sharing of a series of beliefs with his/her interlocutor and to use this background in order to add new beliefs. Shared belief is considered a primitive mental state of communication just like a private belief.

Given two actors, their relationship is defined as the set of playable games. The theory considers the relationship from both a static and dynamic point of view: In order to be playable a game also needs to be valid within the present state of the world. Validity refers to the whole context: physical, social and cognitive. However, the theory does not link directly relationship dynamics, validity conditions and state of the world.

From a situated perspective, the notion of communication as plan recognition and shared knowledge was targeted in the influential book by Suchman (1987). Suchman clarifies the status of plans as "an artifact of our reasoning about action, not as the generative mechanism of action" (Suchman, 1987, p. 39), so plans do not determine actions in any strong sense. Suchman introduced the concept of situated action that describes the influence of the situation, environment included, on actions and communicative acts. However, adopting the situated cognition paradigm does not imply to accept that representations of the environment do not exist:

"Using the terms knowledge and representations synonymously, early situated cognition publications, including my own, say that 'representations are not stored in the brain'. A better formulation is that descriptions are not the only form of representation involved in cognition, and storage is the wrong metaphor for memory" (Clancey, 1998, p. 221).

The other form of non-descriptive representation, that Clancey refers to, emerges from the concept of direct perception (Gibson, 1979) supported by direct coupling between the agent and the world:

"In this interpretation of Gibson's idea of direct perception, directness means that the internal structures constitute and sustain their own space of configurations without *mediating* 'stuff' such as symbol strings representing the world. At this level of processing, outside stuff is neither brought inside directly nor mapped onto internal codes. Internal structures operate on their own changing properties. Higher levels of processing may *categorise* sensory configurations, but these are again only internal correspondences or relations between internal structures" (Clancey, 1998, p. 88).

However, the concept of situated communication introduced by Suchman emphasizes more the role of the whole set of actions performed by the actors as a situation or context for the conversational exchange, than the environment itself:

"When one takes situated language as the subject matter, however, the definition of the field must necessarily shift to communication under naturally occurring circumstances. And when one moves back far enough from the utterances of the speaker to bring the listener into view as well, it appears that much in the actual construction of situated language that has been taken to reflect problems of speaker performance, instead reflects speaker competence in responding to cues provided by the listener" (Suchman, 1987, p. 71).

Neither the shared-plan approach to communication presented above, nor the Suchman's notion of situated language are completely situated. The shared-plan approach follows the indications of a situated language provided by Suchman without considering the actions as situated or adopting her purely constructive notion of plan. The level of description is purely representational and, neither the environment nor the representations the partners have of the environment, are taken into account. On the other hand, Suchman takes into account the environment at the situated level but there is no trace of the environment in her representational notion of plan. The plan is a representation of actions that are in some sense unsituated. In other words,

both approaches lack in considering representations and meta-representations of the environment that are involved in communication and its interaction with actions.

For a complete theory of situated communication it is necessary to consider the interaction between environment and actions at all the levels: objective, directed perceived, representational and meta-representational. Considering environment and actions at each level guarantee the coherence with the situated cognition paradigm. Moreover, the approach should clarify some of the confusion generated by using the concept of "context" for all the levels.

Example

Alice is a professor. She is the tutor of Bob, a Ph.D student who is working on his thesis proposal.

1. >>From: Bob
   >>To: Alice
   >>Hello Alice,
   >>please find enclose the draft of my
   >>thesis proposal.
   >>Bob
   >
   >From: Alice
   >To: Bob
   >Hi Bob,
   >the basic ideas are rather good so the
   >revision will not take long.
   >what about a meeting on Tuesday at my
   >office at 10.00?
   >Alice
   >
   From: Bob
   To: Alice
   Hi Alice,
   See you then
   Bob

2. On Tuesday morning Bob is late for the meeting and he calls from his cell phone.
   A: Hello
   B: Hello it's Bob speaking. I'm late. Sorry, the traffic is heavy today
   A: Don't worry.

3. Bob arrives and Alice is not in the office. The door is open. He waits in the corridor. Alice arrives and invites Bob to enter. They enter and Bob closes the door of the office. Alice re-opens it. They sit at the desk and they start to discuss.

4. After a while a man of the maintenance service knocks on the door. He enters saying that there is an electric failure in the building and he has to control the sockets of the room. While the man checks the room they keep on discussing.

5. The man goes away closing the door. Alice and Bob continue the discussion.

6. After a while the man comes back with a colleague saying that he has probably found the failure. They start to remove the floor tiles talking and making noise. Alice says: "Let's go to the cafeteria". Bob says: "Ok".

7. While they are going downstairs they keep on talking about the proposal but they get stuck in the elevator. Alice presses the alarm button. Bob calls security with the cell phone. While they are waiting they talk about how to get out of there and how to keep cool. Eventually the doors open.

8. They head towards the cafeteria. Alice turns left and Bob stops in the middle of the sidewalk. Alice says: "There is a shortcut to the cafeteria. We can pass through the Maths department". They keep on talking about the elevator.

9. At the cafeteria Alice meets a professor colleague. She introduces him to Bob. The professor asks Bob what is his subject and the professor asks a lot of details about his thesis. Alice says that they are going to work on it right now. The professor goes on asking questions and making suggestions. She takes her coffee cup and leads Bob to a small table with only two chairs. They starts to discuss again.

10. Two cups of coffee later, Alice and Bob agree on the improvements required by the draft and end the discussion.

---

3. Scene, situation, scenery and scenario

We propose four different concepts (scene, situation, scenery and scenario), in order to reflect the integrate influences of environment and of actions at objective, directed-perceived, representational and meta-representational levels respectively.

The terms we adopted -scene, situation, scenery and scenario- require a justification with respect to their usual meaning. We adopted situation ("relative position or combination of circumstances at a certain moment" Marriem-Webster) in the sense introduced in the situated cognition literature (Clancey, 1998). The usual meaning of scenario refers to an hypothetical, possibly simulated, state of affairs ("an account or synopsis of a possible course of action or events" Marriem-Webster) that the reason why we reserved the word for the meta-representational level. The common-sense meaning of scene appears more concrete and real ("the place of an occurrence or action" Marriem-Webster) and we reserved it for the more objective level. Finally and in contrast with scene, we adopted scenery for giving emphasis to the representational level ("the painted scenes or hangings and accessories used on a theater stage" Marriem-Webster).

In particular, scene considers the world and its affordances. Situation considers the directly perceived world and the possible actions. Scenery considers the represented world and the plans, and finally scenario considers the meta-represented scenery and the simulated executions of plans. The last three levels roughly correspond, using Clancey terms, to structural coupling, categorical reference and symbolic interpretation (Clancey, 1998, p. 317).

A *scene* is a state of the world equipped with a set of affordances. For example, a scene can be A's kitchen and its affordances for cooking, eating, drinking, washing. From an objective point of view, given the state of the physical world, the state provides an affordance for an action if there exists an actor that can execute it in that state.

A *situation* is the direct perception that an actor has of a scene. Namely, a situation is the subjective "representation" produced by an actor A of a state of the physical world and of the actions that are possible from the point of view of an actor A. For example: A in A's kitchen perceives the room and the possibility of drinking from the tap. This means that the actor has a functional "representation" of the world that can include mental states. If the world includes mental states the situation can be perceived as shared. Note that a shared situation is not a situation in the shared knowledge but a situation that is directly perceived as shared. For example: A in A's kitchen perceives B in the room and the possibility for both of drinking from the tap.

Giving the affordances of a scene the possibility of an action will be perceived by an actor depending on her own experience of the physical world. Moreover, the possible actions from the point of view of an actor can be the result of complex processes involving, goals, plans, motivations, self-esteem, self-deceit and, perception of self, of the others and of self-in-the-world, with the relevant possible distortions.

A *scenery* is a subjective representation produced by an actor A of a state of the world and of a set of plans that it is possible to execute within the world. Given a state of the world, a plan is possible if: (i) the represented state of the world verifies the preconditions of at least one plan, (ii) the moves of the plan correspond to possible actions in the scene. The plan is said to be executable within the scenery and the scenery is said to host the plan.

For example: A retains {KITCHEN} as a scenery for the private plan [BREWING COFFEE]. An attributed scenery is a scenery attributed to another actor. For example: A entertains {KITCHEN} as attributed to B and as a scenery for the private plan [BREWING COFFE]. A shared scenery is a scenery within the space of the shared knowledge, and a shared scenery can host private or shared plans. For example: A entertains {KITCHEN} as shared between A and B and as a scenery for the private plan [BREWING COFFEE]; For example: A entertains {KITCHEN} as shared by A and B and as a scenery for the plan shared by A and B [COOKING PASTA TOGETHER]. Finally, a shared plan can be executable in a non-shared scenery. For example: If only A knows that there is pasta in his kitchen, A entertains {A's KITCHEN} as a scenery for the plan shared by A and B [COOKING PASTA TOGETHER].

A *scenario* is a subjective representation produced by an actor A that, possibly among other things, represents a scenery. For example: A entertains |B in A's KITCHEN| as a scenario representing B who entertains {KITCHEN} as a scenery for the private plan [BREWING COFFEE] and for the shared plan [COOKING PASTA TOGETHER]. A scenario can be a rather complex representation, possibly counterfactual or dynamic. In this sense our definition is consistent with the usual meaning of a hypothetical situation.

It is beyond our present goal to show how different approaches to context fit into this framework, but it is relevant to show how each concept can be considered a sort of "context" for actions or communicative acts. *Scene* can be considered as the context in an objective sense (e.g. the room the reader is in and its affordance for reading, writing etc.). *Situation* is the perceived context (e.g. the perception of the room the reader has now, while acting, namely reading), the here and now. *Scenery* emphasize the role of representation (e.g. a representation of the room the reader have or had and of the fact that there it is possible to read a paper). Finally *scenario* is related to context in the sense of encapsulable representations (e.g. the representation the reader had while thinking in the previous two examples). Our approach is consistent with the pragmatic approach to the relationship between context and relevance proposed by Ekbia & Maguitman (2001). Earlier cognitive pragmatics accounts of context (Bara and Bucciarelli, 1998) concentrated on the role of mental states and shared knowledge in the comprehension of a communicative act (Blanzieri and Bucciarelli, 1996a; 1996b).

4. A Theory for Situated Communication

In this section we use the concepts of *scene, situation* and *scenery* for elaborating a theory for situated communication. The aim of the theory is to give an account of the interaction between mental states involved in communication and the subjective representations of the state of the world.

The basic assumption is that from a cognitive point of view the three levels we hypothesize co-exist and co-operate. We assume that during situated communication the actors experience a flow of situations and each of these situations inform their actions. We also assume that representations like plans and sceneries can be mentally constructed as private, shared or attributed. Finally, the actor can entertain complex meta-representations (scenarios) involving sceneries. From an objective point of view the actors executes actions on a particular scene that can be perceived by the actors in different subjective situations and represented in different subjective sceneries.

What is relevant is the relation of the scenery with the shared-plans in terms of preconditions and possible moves. For example, given an actor like Alice in the Ex.1-10, her representation of her office in a University department {OFFICE} is a scenery for the plan [TUTORING SESSION], that has its preconditions verified and its moves are possible. Both {OFFICE} and [TUTORING SESSION] are shared between Bob and herself.

The scenery represents a state of the world including the communication channel. In Ex. 1 Alice and Bob communicate by e-mail and in Ex. 2 by phone. In both cases the scenery can include the remote presence of the actors and, in the case of e-mail, the asynchronous access to the messages.

It is worth noting that in a situated perspective any attempt to produce a representation of a situation, produces a scenery or a scenario. In fact, a situation is a direct perception, not a representation. In Ex. 2 we can only suppose the shared situation Alice perceives during the phone call. It will probably include Bob, the physical world Alice perceives through the phone and actions such as talking or listening. But Bob is stuck in the traffic so he has a private situation that includes himself facing a traffic jam with waiting or walking as possible actions. In any case in the attempt of representing the situation of the other agent, each actor entertains and attributes sceneries.

A *situation* is subjective, so it can change depending on whether a change in the scene occurs or not. A change of the scene produces a change of the *situation* if the actor perceives it. In Ex. 7 Alice and Bob realise they are stuck in the elevator and that changes their *situation*. A scene can change for external reasons or by means of an action performed by the actor or by the partner. In Ex. 7 the elevator stops for an external reason whereas Alice and Bob perform two actions (press the alarm button and call security) that changes the scene.

Scenery are subjective, hence they can be unrelated to the real scene. For instance, in Ex. 8 Alice attributes to Bob a scenery of the University that does not include the shortcut and using-the-shortcut as a possible action. Obviously, sceneries may also be private representations, permitting non-standard communication such as irony or deceit. In Ex. 2 Bob could lie about the traffic and have a private scenery that differs from the supposedly shared scenery he proposes to Alice. Being a representation, the scenery can also change by means of a communicative act without any change in the scene. In Ex. 8 Alice informs Bob that there is a shortcut, information which changes his scenery {UNIVERSITY}. Finally, a scenery change can be a goal of a behaviour game. In Ex. 6 Alice and Bob start to a play a behaviour game aimed to produce the scenery {THE TABLE AT THE CAFETERIA}.

A scenery hosts different plans, and conversely a plan is executable in different sceneries. For example, [SCIENTIFIC DISCUSSION] and [TUTORING SESSION] are shared plans playable in the scenery {OFFICE}. Other sceneries like for instance {THE TABLE AT THE CAFETERIA} can host some of those shared plans. A shared plan is in principle compatible with more than one scenery. Thus, a change in the scenery does affect the game, which normally will develop within the constraints of the new scenery. In Ex. 6 Alice and Bob consider [TUTORING SESSION] as playable in both {THE TABLE OF THE CAFETERIA} and {OFFICE}. A scenery is subjective, so different actors can consider different games as playable in a scenery. A more formal professor, for example, can consider [TUTORING SESSION] not playable at {THE TABLE OF THE CAFETERIA}.

In some case a game may be played only in a specific scenery (e.g. trial in court); in other cases a game is incompatible with a scenery, hence if the scenery is activated, the game will end (e.g. smoking in a high-school toilet is interrupted by the presence of the supervisor). A change of the scenery closes the game only if the new scenery does not host the game. In Ex. 4 Alice and Bob continue to play [TUTORING SESSION] after the entrance of another actor changed the scenery from {OFFICE} to {OFFICE WITH MAINTANANCE GUY}. The new scenery hosts the game so the actors can continue to play. This is the case also in Ex. 5 where the scenery {OFFICE WITH MAINTANANCE GUYS} has changed the scenery {OFFICE}. On the contrary, the modification of the scenery {OFFICE} to the scenery {OFFICE WITH MAINTANANCE GUYS} in the Ex. 6 interrupts [TUTORING SESSION].

The actors, by modifying the partners' sceneries, shape their relationship. To settle the validity conditions of a behaviour game, is an implicit way of controlling the relation between agents. Actor A make possible for herself and B to engage in game [G], by guaranteeing an adequate scenery. In fact, proposing the scenery for a game amounts to bidding that game (e.g. driving home a potential sexual partner). In Ex. 9 Alice chooses a table that modifies the shared scenery with the annoying professor in a way that prevents the playability of a game {THREE PEOPLE DISCUSSION}. The dynamic of the relationship produces effects in the long term, also affecting the basic relationship. In Ex. 3 the actions of opening or closing the door change the scenery dramatically. Actors modify the sceneries by means of actions that can be communicative acts, as noted in Section 3.

Given the co-presence of more than two agents, for each agent the third one can be part of the scenery, or can be involved in some behaviour game. In Ex. 4 and Ex. 6, the maintenance people are part of the modification of the {OFFICE} in a very natural way. In Ex. 9, the annoying professor tries to play [THREE PEOPLE DISCUSSION] in the scenery {CAFETERIA}. Alice changes the scenery to {THE TABLE OF THE CAFETERIA} that does not host [THREE PEOPLE DISCUSSION] and the game is closed. The professor does not join Alice and Bob and so he is not even part of the scenery anymore.

In order to understand the kind of phenomena our theory accounts for, it is interesting to note that a communication exchange produces actions that are either moves of the shared-plan or actions aimed to construct, maintain or modify the shared scenery. A shared-plan theory such as Cognitive Pragmatics accounts only for the changes of the scenery produced by the execution of the shared plan. In this case it would be possible to assume the existence of a general shared plan that gives an account of the whole sequence of actions. For example, the play of [TUTORING SESSION] in the scenery {THE TABLE AT THE CAFETERIA} could be considered as the execution of a more complex behavior game than [GOING TO A TABLE

AT THE CAFETERIA FOR A TUTORING SESSION]. This operation is not plausible, in particular when the modification of the scenery is a consequence of a private plan or of an external cause that changed the scene. In fact, the idea of scenery prevents the explosion of the number of the behavior games.

5. Conclusions

We have presented a theory based on the concepts of scene, situation, scenery and scenario that gives an account of the interaction between mental states involved in communication and representations of a state of the world. The adoption of a situated cognition paradigm motivates the introduction of the concepts. Differently from precedent approaches to situated communication, we emphasize the role of the environment and of the representations agents retain of the environment.

The theory presented in this work refers to the cognitive process of two actors involved in a communicative exchange. Therefore, we do not consider the effects and phenomena produced by the interaction of three or more people. This requires further work in order to bridge the gap between cognitive processes involved in communication and phenomena studied by social psychology.

Acknowledgements

The authors wish to thank Mauro Adenzato, Monica Bucciarelli, Lorenzo Pia, David Pickup and Georgia Zara for reading and the suggestions given. The names of the authors are in alphabetical order. This work has been supported by Ministero dell'Università e della Ricerca Scientifica e Tecnologica of Italy (Cofinanziato 2001: *Strumenti qualitativi e quantitativi per l'analisi della relazione psicoterapeutica*).

References

Airenti, G., Bara, B.G., & Colombetti, M. (1993a). Conversation and behaviour games in the pragmatics of dialogue. *Cognitive Science, 17*, 197-256.

Airenti, G., Bara, B.G., & Colombetti, M. (1993b). Failures, exploitations and deceits in communication. *Journal of Pragmatics, 20*, 303-326.

Austin, J.L. (1962). *How to do things with words.* London: Oxford University Press. [2nd ed. revised by Ormson, J.O., & Sbisà, M. London: Oxford University Press, 1975].

Bara, B.G., & Bucciarelli, M. (1998). Language in context: The Emergence of Pragmatic Competence. *Special Issue of Analise Psicologica: Cognition in Context*. In: Quelhas, A.C., & Pereira, F. (Eds.). Istituto Superior de Psicologia Aplicada, Lisbona.

Blanzieri, E., & Bucciarelli, M. (1996a). Reasoning processes underlying communication: Extracting the rules of the game from a connectionist network. *IX Conference of the European Society for Cognitive Psychology*, Wurzburg, 101.

Blanzieri, E., & Bucciarelli, M. (1996b). The evaluation of the communicative effect. *Proceedings XVIII Conference of the Cognitive Science Society*, San Diego, 501-506.

Clancey, W.J. (1998). *Situated cognition*. Cambridge, UK: Cambridge University Press.

Ekbia, H. R. & Maguitman, A. G. (2001). Context and relevance: a pragmatic approach. In Akman, V., Bouquet, P., Thomanson, R., & Young, R.A. (eds.) *Modeling and Using Context* LNAI 2116, Proceedings of CONTEXT2001. Berlin, Springer.

Gibson, J.J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Grosz, B., & Kraus, S. (1996) Collaborative Plans for Complex Group Action. *Artificial Intelligence* 86, 2, 269-357.

Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. London: Cambridge University Press.

Searle, J.R. (1975). Indirect speech acts. In Cole, P., &. Morgan, F.L. (eds.), *Speech acts: syntax and semantics, vol. 3*. New York, Academic Press.

Sperber, D., & Wilson, D. (1986). *Relevance*. Cambridge, MA: Harvard University Press.

Suchman, L. (1987). *Plans and situated action*. Cambridge, UK: Cambridge University Press.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Blackwell.

Ah-Ha, I Knew It All Along: Differences in Hindsight Bias Between Insight and Algebra Problems

Ivan K. Ash (iash1@uic.edu)
Jennifer Wiley (jwiley@uic.edu)

Department of Psychology
The University of Illinois at Chicago
1007 West Harrison Street (M/C 285)
Chicago, IL 60607, U.S.A.

Abstract

The present study investigated the role of restructuring in the solution of insight and incremental problems. Participants were presented with a series of insight and algebra word problems in a hindsight bias paradigm (Fischoff, 1975). Those who solved the insightful problems correctly showed increases in importance ratings on the key problem components. However, no increases in importance ratings were detected for the key problem components of algebra problems. These results are consistent with theories that propose that representational restructuring plays a fundamental role in the insightful problem solving process (Davidson & Sternberg, 1984; Ohlsson, 1992).

Restructuring and Insightful Problem Solving

A number of researchers have suggested that insight problems may be solved in a qualitatively different manner from incremental or analytic problems (Duncker, 1945, Davidson & Sternberg, 1984, Metcalfe & Wiebe, 1987, Ohlsson, 1992). For example, when Metcalfe and Wiebe (1987) had subjects make feeling of warmth ratings while solving algebra and insight problems, they found that while solving algebra problems, subjects' warmth ratings steadily increased towards the time of solution; whereas during the solution of insight problems subjects' warmth ratings remained low and suddenly increased right before they solved the problem. These results suggest the processes for solving insight problems are different from the systematic or analytical processes used for incremental problems. However, this only represents the *suddenness* of the insightful process and does not shed light on the causes of the solution patterns.

It has been suggested that insight problems differ from incremental problems in that the nature of the solvers' experience causes them to construct a representation of the problem that cannot lead to the correct solution (Duncker, 1945, Davidson & Sternberg, 1984, Ohlsson, 1992). In order to come to the correct solution, solvers must restructure their original conception of the problem. Proposed evidence for restructuring in insight problem solving has come from a wide range of empirical findings. For example, Dominowski and Buyer (2000) found decreases in re-solution time for those who correctly solve insight problems but not for those who were simply shown the answer. Knoblich, Ohlsson, and Raney (2001) analyzed eye movements during the solution of matchstick arithmetic problems and found evidence for impasses followed by increased fixations on components of the problems that were key to solution. Durso, Cornelia, and Dayton (1994) attempted to find an independent measure of restructuring by statistically modeling successful and unsuccessful problem representations after the problem solving session. They found that successful solvers representations centered on concepts key to the nature of the solution, whereas non-solvers representations centered on the principle characters in the story problem, which were not relevant to the solution.

These empirical studies suggest an insight problem solving process that involves the restructuring of the mental problem representation. However, these studies fall short of proving the existence of a restructuring process because they fail to directly measure representational change before and after solving across an individual and fail to compare problems that involve incremental solutions to the problems proposed to elicit insightful solutions. What is needed is an independent measure of restructuring in order to test theories that predict an insight process involving restructuring against theories which do not predict restructuring in any problem solving process (such as Weisberg's nothing special view, 1986). Furthermore, this method must be able to directly test whether more restructuring is involved in insightful than incremental problem solving. The present study uses a hindsight bias paradigm to produce an independent measure of the amount of restructuring involved in solving different types of problems.

Hindsight Bias

Hindsight bias is the observation that people with outcome knowledge of a situation falsely believe that they would have predicted the correct outcome (Hawkins & Hastie, 1990). Fischoff (1975) originally developed the basic paradigm. He had people read a narrative of a situation with or without receiving the outcome and then had them rate the probability of alternate outcomes as if they had no knowledge of the outcome. The general finding is that people with outcome knowledge unknowingly rate the outcome they were told as more probable than alternatives, as if they "knew it all along." Individuals who receive outcome knowledge also rate the sentences in the narrative that support the given outcome

as more important than those that do not, even though they are asked to ignore the given outcome. This same effect has also been shown in within-subject studies in which participants are asked to rate outcomes before they receive outcome information, and then asked to reproduce their original questionnaire ratings after receiving outcome information (Fischoff & Beyth, 1975). Most research suggests that hindsight bias is not due to motivational factors but involves cognitive processes that automatically restructure one's situation representation to accommodate the new information, leaving individuals unable to access or reproduce their original representation (Hawkins & Hastie, 1990).

This leads to the hypothesis that processes that cause more restructuring will lead to more hindsight bias. Restructuring theories of insightful problem solving make clear predictions about the nature of hindsight bias for insightful problems. If correctly solving insight problems involves restructuring of the problem space to come to a solution, then those who correctly solve insight problems should show hindsight bias on the problem components that are key to solution. However, those who fail to correctly solve the insight problems should not show any hindsight bias on those components.

If this restructuring occurs only as a result of an insightful problem solving process, then being shown the solution should not lead to hindsight bias of a similar nature to those who solve on their own. Algebra problems should be solved in an incremental fashion. Therefore, restructuring theories of insightful problem solving would predict no hindsight bias on the key components of algebra problems regardless of correctness of solution or being shown the answer. Finally, the "nothing special" or gradual transformation theory of insightful problem solving (Weisberg, 1986) predicts no differences in between insightful and incremental solution processes. Therefore, this theory would predict no difference in hindsight bias between insight and algebra problems or those who come to solution, and those who do not come to solution but are shown the correct answers.

The present study consisted of two sessions. Participants received a series of insight and algebra word problems. During session one, Ss first were asked to read through each problem carefully and rate each component of the problem on its importance in finding the solution. Then they were asked to attempt to solve each problem. Following the solution phase approximately half the Ss were shown the correct solution. After a one week interval Ss returned for the second session. At this time they were asked to attempt to remember their original importance ratings for each of the problem components. Hindsight bias was measured as the change in importance ratings between the two sessions, in favor of relevance for solution.

Methods

Participants. One hundred twenty eight introductory psychology students participated in this study to fulfill a class requirement. They were run in groups of 3 to 12.

Design. This study consisted of two one-hour sessions separated by a one-week interval. During the first session Ss rated the importance of the problem components of several insight and algebra problems. Next they attempted to solve each problem. After attempting to solve, half the groups were shown the solution. After a one-week interval Ss returned and were asked to reproduce their original component importance ratings. This resulted in a 2 session rating (session 1, session 2) X 2 shown answer (Yes, NO) X 3 solution type (no solution, incorrect, correct) X N components (varies by problem) mixed design. Session rating and problem components were within-subject variables, while shown answer and solution type were grouped between subjects.

Materials and Procedure. Participants completed the S1Q booklet first. The S1Q booklet consisted of six insight problems and four algebra word problems. Ss were instructed to carefully read each problem, but not to attempt to solve the problems. Instead, Ss were asked to rate each sentence or component of the problem on how important it is in finding the solution to the problem. Each sentence or component of the problem was listed one at a time followed by a rating scale that consisted of a 7.3 cm continuum. The far left side of the continuum was marked as representing "very unimportant," while the far right side was marked "very important." Participants were instructed to make a mark anywhere on the continuum that best represented their opinion of the particular problem component. Participants were allowed to work through the questionnaire booklet at their own pace but were allowed no more than 15 min. to complete all the ratings. The experimenter periodically reminded participants not to attempt to solve the problems throughout the rating phase. These ratings are the session 1 importance ratings.

Next participants completed the problem-solving packet (S1S). The S1S booklet presented the same 10 problems, each on its own page. Ss were instructed to attempt to solve each problem and that they would be given three minutes to complete each problem. The directions instructed the Ss to show all work, circle their final answer, and if necessary explain the solution using a few short sentences. Ss were instructed to work through the booklet in order, stop working on a problem at the experimenter's signal, and wait until the experimenter's signal to begin the next problem. Ss were given 3 min. to work on each problem. The Ss performance each problem was used to assign participant into one of three solution type groups (no answer, incorrect, correct) for each problem.

After attempting to solve all the problems, half of the groups were dismissed (shown answer no). The other half (shown answer yes) were shown step-by-step outlines of the solutions of each of the problems on an overhead projector. The experimenter read a script that explained each answer. Each problem explanation took approximately 1 min. At the conclusion of session 1 Ss were asked not to discuss the details of any of the

problems they saw or their solutions with anyone else in the Subject Pool and dismissed.

Session 2 occurred exactly one-week later in the same room at the same time. Ss were first issued the second session questionnaire packet (S2Q). This packet was identical to the first except that the participants were asked to attempt to reproduce their exact component importance ratings from the first experiment. These ratings are the Session 2 importance ratings. They were once again allowed to work through the booklet at their own pace but had no more than 15 min. to complete the entire memory test. Then, Ss once again attempted to solve each problem in the same manner as in session 1. Finally, participants were debriefed as to the nature of the study and once again asked not to discuss the study, problems, or solutions with anyone else in the Subject Pool.

Table 1: Number of Participants Per Cell by Problem.

	Shown Answer Yes			Shown Answer No			
Sol. type	NA	IN	CO	NA	IN	CO	Total
Train	14	27	30	11	16	30	128
Age	20	25	26	18	20	19	128
Triangle	22	23	26	20	17	20	128
Cups	6	23	24	11	26	25	115*

Note: NA = No Answer; IN = Incorrect; CO = Correct

* 13 Ss had to be dropped from this problem for missing data and/or marking the top importance level for all components on both sessions.

Results

Two algebra and two insight problems from the set of problems were chosen for the problem component analysis. These problems were selected on the basis of two criteria. The first criterion was due to a conceptual constraint. Each problem component or individual sentence had to contain only either information that was key to the solution or not. In other words, any one component of a problem must have been mutually exclusive from the other components and contain unique information that was only interpretable as important or unimportant depending on one's interpretation of the problem. The second constraint was a practical constraint. Each problem had to result in a moderate solution rate such that there would be a number of individuals in each of the solution type groups, and result in a similar number of individuals for each solution type across "answer shown" groups. The two insight and two algebra problems that best met these criteria were selected for analysis. The problems are presented in Figures 1-4, problem components are indicated by lower case letters. The number of participants in each cell for each problem is listed in Table 1.

Problem Component Ratings Problem component scores were coded by measuring the distance of the participants' mark on the continuum from the left end with a ruler. Lower scores indicate that a sentence or component in a problem was perceived as of little importance toward the correct solution, while higher scores indicate that a sentence or component in a problem is perceived of great importance in coming to the correct solution. Hindsight bias is the increase of an individual's importance rating for a solution-related problem component between the first and second sessions.

Restructuring theories of insightful problem solving predict that on insight problems, initial component importance ratings should reflect solvers' activation of an inappropriate problem representation. Those who overcome the impasse and correctly solve the problems need to restructure their representation of the problem to activate the correct operators or components for solution. This new representation should lead to hindsight bias in the importance ratings for the key components of the problem.

Those who fail to solve the problem should show no hindsight bias for key problem components, and may even increase their importance ratings for the original inappropriate components of the problem. In algebra problems no restructuring of the problem representation is necessary to come to solution, therefore no change in representation of the problem should be evident in the hindsight bias measures of algebra word problem components.

a) Two trains leave the same station at the same time. **b)** Each has enough fuel for a 2000 mile trip. **c)** The trains travel in opposite directions. **d)** One train travels 60 miles per hour, and the other 100 miles per hour. **e)** In how many hours will the trains be 800 miles apart?

Figure 1: Train Problem (Problem Type: Algebra)

a) Ann is twice as old as her son. **b**) They were both born in June. **c)** Ten years ago Ann was three times as old as her son. **d)** What are their present ages?

Figure 2: Age Problem (Problem Type: Algebra)

The triangle shown below points to the top of the page. Show how you can move 3 circles to get the triangle to point to the bottom of the page.

a)O

b)O c) O

d)O e)O f)O

g)O h)O i)O j)O

Figure 3: Triangle Problem (Problem Type: Insight)

The picture below is of six glasses. The first three contain liquid. Describe how you could make it so no two glasses containing liquid are next to each other, while keeping three of the six glasses full. To do this, you are only allowed to move **one** glass.

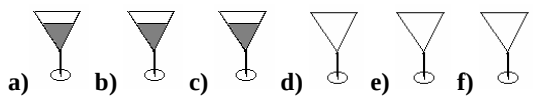


Figure 4: Cups Problem (Problem Type: Insight)

Train Algebra Problem The train problem (see Figure 1) involves constructing and solving equations in order to find how many hours it will take for two trains to be 800 miles apart. To solve this problem one must use information from sentences (components) **A, C, D,** and **E**. These sentences are therefore the *key components* in this problem. To investigate whether restructuring was involved in solving this algebra problem, a 2x5x2x3 (session x component x answer given condition x solution type (no solution, incorrect solution, correct solution)) ANOVA was conducted. Evidence for restructuring as measured by hindsight bias is revealed by interactions involving the session variable. Specifically, an increase in importance ratings on key components across sessions (initial rating vs. second rating) for those who correctly solve the problem is evidence for restructuring.

This analysis revealed a significant session x component x solution type interaction F (8, 484) = 2.69, p < .05, n^2 = .04. Follow-up analysis revealed that this small interaction was due to significant *decreases* in importance ratings on component **D**, and **E** for those who answered incorrectly and were shown the answer (t (26) = -3.71, p < .01, & t (26) = -2.78, p < .01 respectively). There was also a significant decrease in importance rating on component **E** for those who answered correctly and were shown the solution, t (29) = -3.21, p < .01 (see Table 2). If restructuring of the problem representation was necessary to solve this algebra problem one would expect to see *increases* in importance ratings on components A, C, D and E for those who correctly solve. This was not the case. Clearly there is no evidence of restructuring in those who solved the train algebra problem.

Age Algebra Problem The Ann and Son problem involves constructing equations from the given information to calculate Ann and her son's present age. Sentences (components) **A**, **C**, and **D** are key components for this problem (see Figure 2). To investigate whether restructuring was involved in solving this algebra problem, 2x4x2x3 (session x component x answer condition x solution type) ANOVA was conducted. This analysis once again revealed a significant session x component x solution type interaction F (6, 353) = 3.35, p < .05, n^2 = .05. However, follow up analyses revealed no significant differences between session 1 and session 2 ratings for any of the components in any of the groups (see Table 2). This interaction was most likely caused by the trend toward hindsight bias on component **C** by those in the correct solution/not given answer condition, and by a trend towards a *decrease* in importance ratings on component **D** in the incorrect/ shown answer group. However, interaction was small in effect size. These unsystematic trends are weak evidence for restructuring. Therefore, once again, we find no clear evidence restructuring in the solution of algebra problems.

Cups Insight Problem In order to solve this problem, one must pick up the glass in position **B** and pour the liquid into the cup in position **E.** Therefore, **B** and **E** are the key components of this problem. The insightful process theories would predict that the initial representation should be biased against viewing cup **E** as being key in finding the correct solution. This leads to the prediction that importance levels should increase for the two key components for those who correctly solve, while those who do not correctly solve should not show hindsight bias on only the two key components.

To investigate whether restructuring was involved in solving this problem, a 2x6x2x3 (session x component x answer condition x solution type) ANOVA was conducted. This analysis revealed a significant session x component interaction, F (5, 545) = 4.7, p < .05, n^2 = .02. Follow-up analysis revealed significant increases in importance ratings for both key components **B** (t (119) = 3.85, p < .05, n^2 = .11) and **E** (t (119) = 2.10, p < .05, n^2 = .03) across all participants. Even though a survey of the means (see Table 4) gives the impression that the groups that solved the problem correct are driving this effect, a significant interaction involving solution type was not detected. This result tends to follow the predicted pattern of hindsight bias that would be expected with regards to restructuring accounts of insightful problem solving, and stands in contrast to the algebra results that showed no evidence of restructuring.

Triangle Insight Problem To solve this problem, one must move the corner three circles around one position in order to make the triangle point down. The insightful process theories would predict that the initial representation should be biased towards viewing the array as a triangle and therefore solvers will inappropriately view the point of the triangle, or the top three circles, as more important. In order to solve, individuals will need to restructure their representation of the middle circles of the triangle as invariant whether the triangle points up or down. Therefore, the corner circles **A**, **G,** and **J**, are the key components for solving the problem (see Figure 3).

To investigate whether restructuring was involved in solving this insight problem, a 2x10x2x3 (session x component x answer condition x solution type) ANOVA was conducted. This analysis revealed a significant session x component interaction, F (9, 1098) = 4.05, p < .05, n^2 = .03; a significant session x component x answer condition interaction, F (18, 1098) = 2.04, p < .05, n^2 = .01, and a significant session x component x solution type interaction, F (18, 1098) = 1.91, p < .05, n^2 = .03. Follow-up analysis found significant increases in importance ratings, or strong positive trends, on each of the key components for those who successfully solved the problem regardless of whether they were shown the answer (see Table 5). This evidence supports the idea that those individuals who solved correctly had a different (more appropriate) problem representation on their second encounter with the problem.

Other significant increases were found on components **D** and **E** (both unimportant for finding correct solution) for those who came to no solution and did not receive the answer. In the groups that received the answers but did not find the correct solution, increases in importance ratings were detected on the entire bottom row of circles

(**G** through **J**). These findings suggest that being shown the solution to an insight problem can lead to a shift in representation. However this shift is not exclusive to the key problem solving components.

Discussion

These results suggest that hindsight bias can be used as an independent measure of the restructuring involved in solving different problems. These results demonstrate some of the characteristics predicted by insightful problem solving theories that involve mechanism of sudden restructuring of one's problem representation in order to come to solution (Duncker, 1945, Davidson & Sternberg, 1984, Ohlsson, 1992). Although all of the predicted interactions on insight problem component ratings were not found to be significant, there was evidence of an increase in importance ratings on key components in insight problems and not on algebra problems. Also there was evidence that these increases in importance ratings occurred in groups that had successfully solved the insight problem, regardless of whether they were shown the correct solution. The results on the insight problems contrast with those for the algebra problems, which showed little evidence of any increases on importance ratings on the key components. In examining the results it is clear that individuals initially were able to recognize which components of the algebra problems were key in solving the problem. However, even though all individuals were able to correctly report the importance of the key components, many still failed to correctly solve the algebra problems. This suggests that the locus of difficulty for the algebra problems did not lie in the representation.

While all the predicted interactions on the insight components were not found, it is helpful to keep in mind that this was a very stringent test for hindsight bias. The participants in this study were not asked to re-rate the questions during the second session of the experiment. They were asked to "reproduce" or remember their original ratings. It has been estimated that 2/3 of participants are actually able to remember their original rating in this type of within-subjects design (Fischoff & Beyth, 1975). Even though steps were taken to prevent this (ex. rating were done on continuums instead of Likert scales, one full week between sessions) this surely may have affected the final outcome. Also, all theories of insightful problem solving agree that restructuring is far more likely when subjects are naïve to the problems. In this design there was no way to separate out those who had experience with the problems and were able to solve through some other method than restructuring (i.e. memory search). On the same note, it may be possible that individuals did not have the expertise in algebra to solve the problems in a totally incremental fashion. In this design there was also no way to separate out those who experienced *partial insights* while solving the algebra problems (see Ohlsson, 1992). These factors may have also confounded the hindsight bias results. Despite all of these issues, there was clear evidence for differential amounts and patterns of hindsight bias between insight and algebra problems. Currently we are in the process of conducting more detailed studies using the hindsight bias paradigm on a wider range of problems, as well as using this method to investigate the role of impasse in the insightful problem solving process.

Acknowledgements

This research was completed as part of a Masters degree requirement at the University of Illinois at Chicago. Special thanks to Thieu Dang, Trina Kershaw, Amy Brodhead, Josephina DeAnda, Monica Nemiec, and Yessenia Cervantes for the assistance in running participants, and coding/entering data.

References

Davidson, J. E. & Sternberg R. J. (1984). The role of insight in intellectual giftedness. *Gifted Child Quarterly, 28*, 58-64.

Dominowski, R. L., & Buyer, L. S. (2000). Retention of problem solutions: The re-solution effect. *American Journal of Psychology, 113 (2*), 249-274.

Dunker, K. (1945). On problem solving. *Psychological Monographs, 58.*

Durso, F., Rea, C. & Dayton, T. (1994) Graph-theoretic confirmation of restructuring during insight. *Psychological Science, 5*, 94-98.

Fischoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under certainty. *Journal of Applied Social Psychology, 18*, 93-119.

Fischoff, B., & Beyth, R. (1975). " I knew it would happen"- remembered probabilities of once-future things. *Organizational Behavior and Human Performance, 13*, 1-16.

Hawkins, S. A. & Hastie, R. (1990). Hindsight biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107*, 311-327.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29(7*), 1000-1009.

Metcalfe, J. & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition, 15*, 238-246.

Ohlsson, S. (1992). Information processing explanations of insight and related phenomenon. In. M. Keane & K. Gilhooly (Eds.), *Advances in the Psychology of Thinking* (pp. 1-44). London: Harvester-Wheatsheaf.

Weisberg, R. W. (1986). *Creativity, genius, and other myths*. NY: Freedman.

Weisberg, R. W., & Alba, J. W. (1981). An examination of the role of "fixation" in the solution of several "insight" problems. *JEP: General*, 110, 169-192.

Table 2: Train Algebra Problem: Mean (SD) Importance Ratings by Session, Answer Group, and Solution Type

	Shown Answer Group						Not Shown Answer Group					
Sol.	No Answer		Incorrect		Correct		No Answer		Incorrect		Correct	
Sess.	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
A*	4.6(2.2)	4.9(1.9)	5.7(2.0)	5.3(2.0)	5.8(1.3)	5.8(1.8)	6.2(1.1)	5.4(1.9)	6.1(0.8)	5.5(1.7)	5.7(1.9)	5.8(1.7)
B	4.1(2.1)	.8(2.4)	3.3(2.5)	3.6(2.4)	3.3(2.5)	3.2(2.5)	3.5(2.8)	3.8(2.8)	3.6(2.9)	4.4(2.6)	2.9(2.6)	2.3(2.2)
C*	2.8(2.3)	3.9(2.6)	4.3(2.4)	4.1(2.4)	5.3(2.0)	5.0(2.1)	4.2(2.9)	5.5(1.5)	4.6(2.0)	5.3(1.6)	5.4(2.1)	4.8(2.4)
D*	6.1(0.7)	5.9(1.0)	**6.8(0.3)**	**6.3(0.7)**	6.5(1.0)	6.4(0.6)	6.3(1.7)	6.6(0.5)	6.4(1.0)	6.1(1.0)	6.5(0.5)	6.4(0.6)
E *	5.7(1.4)	5.2(1.8)	**6.5(0.7)**	**6.1(0.8)**	**6.4(0.8)**	**6.1(1.0)**	6.2(1.2)	6.5(0.7)	6.2(1.1)	5.9(1.0)	6.3(0.9)	6.3(0.9)

Note: Bold-faced cells denote a significant difference between the importance ratings on session 1 (S1) and session 2 (S2), $p < .05$. Asterisk (*) indicates key component.

Table 3: Age Algebra Problem: Mean (SD) Importance Ratings by Session, Answer Group, and Solution Type

	Shown Answer Group						Not Shown Answer Group					
Sol.	No Answer		Incorrect		Correct		No Answer		Incorrect		Correct	
Sess	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
A*	6.6(0.5)	6.0(1.3)	6.3(1.1)	6.3(0.8)	6.3(0.7)	6.3(0.6)	6.3(0.5)	6.2(0.6)	6.4(0.8)	6.2(0.8)	6.5(0.8)	6.6(0.5)
B	1.5(1.5)	2.2(2.0)	2.1(2.0)	1.8(1.9)	2.3(2.2)	2.2(2.2)	2.5(2.0)	3.1(2.7)	1.7(1.7)	2.1(2.2)	2.4(2.4)	1.6(2.1)
C*	6.3(0.7)	6.1(0.7)	6.3(1.0)	6.3(0.9)	6.2(0.8)	6.3(0.7)	6.2(0.6)	6.0(1.0)	6.4(0.7)	6.3(0.9)	*6.3(0.8)*	*6.6(0.5)*
D*	5.6(2.0)	5.2(1.8)	*6.1(1.1)*	*5.6(1.8)*	5.7(1.5)	6.1(0.8)	5.5(1.8)	5.8(1.5)	6.0(1.3)	5.9(1.3)	6.4(0.7)	6.5(0.6)

Note: No significant difference between the importance ratings on session 1 (S1) and session 2 (S2) were detected, $p < .05$. Italics indicates trend toward significance $p < .08$, $n^2 > .10$. Asterisk (*) indicates key component.

Table 4: Cups Insight Problem: Mean (SD) Importance Ratings by Session, Answer Group, and Solution Type

	Shown Answer Group						Not Shown Answer Group					
Sol.	No Answer		Incorrect		Correct		No Answer		Incorrect		Correct	
Sess.	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
A	4.1(2.3)	4.1(2.3)	**4.2(2.3)**	**3.4(2.7)**	3.8(2.5)	3.5(2.8)	4.1(2.4)	5.2(1.0)	3.6(2.5)	4.0(2.6)	3.6(2.5)	3.8(2.7)
B*	5.4(1.2)	5.6(1.1)	*5.0(2.1)*	*5.9(1.3)*	5.7(1.8)	6.1(1.2)	5.1(1.8)	5.4(0.7)	5.5(1.9)	5.9(1.4)	**5.3(1.8)**	**6.3(1.0)**
C	4.2(1.7)	4.1(2.0)	3.9(2.0)	3.6(2.4)	4.3(2.3)	3.9(2.6)	5.9(0.2)	4.9(1.7)	4.3(2.3)	4.5(2.5)	3.5(2.4)	3.8(2.6)
D	4.5(1.4)	4.5(1.8)	3.7(2.2)	3.8(2.3)	3.6(2.5)	3.5(2.5)	3.8(1.9)	3.5(1.9)	3.4(2.3)	3.1(2.2)	3.5(2.2)	3.3(2.5)
E*	*4.2(1.5)*	*4.9(1.6)*	4.1(2.1)	4.5(2.1)	4.6(2.3)	5.3(2.1)	4.4(1.7)	3.9(1.9)	4.4(2.2)	4.2(2.2)	**4.3(2.1)**	**5.8(1.6)**
F	4.0(1.9)	4.5(1.8)	3.7(2.3)	3.7(2.4)	3.5(2.3)	3.5(2.5)	3.4(2.2)	3.1(1.7)	3.3(2.4)	3.3(3.4)	3.9(2.1)	3.5(2.5)

Note: Bold-faced cells denote a significant difference between the importance ratings on session 1 (S1) and session 2 (S2), $p < .05$. Italics indicates trend toward significance $p < .08$, $n^2 > .10$. Asterisk (*) indicates key component.

Table 5: Triangle Insight Problem: Mean (SD) Importance Ratings by Session, Answer Group, and Solution Type

	Shown Answer Group						Not Shown Answer Group					
Sol.	No Answer		Incorrect		Correct		No Answer		Incorrect		Correct	
Sess.	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
A*	5.7(1.7)	6.0(1.9)	5.6(2.2)	6.0(1.4)	*5.3(1.6)*	*6.0(1.2)*	5.0(1.9)	4.8(2.3)	5.7(1.7)	6.0(1.2)	**5.3(1.9)**	**6.3(0.8)**
B	5.4(1.8)	5.0(2.1)	5.0(2.4)	5.4(1.9)	3.6(2.1)	3.3(4.5)	4.2(2.0)	3.8(2.4)	5.2(1.6)	5.0(2.1)	4.0(2.4)	3.5(2.5)
C	5.4(1.6)	5.1(1.9)	5.0(2.5)	5.2(2.0)	3.7(2.1)	3.4(2.5)	4.4(2.0)	3.7(2.3)	5.3(2.1)	5.2(2.0)	**4.3(2.1)**	**3.4(2.5)**
D	3.8(2.3)	3.7(2.4)	3.2(2.3)	3.0(2.2)	3.1(1.9)	3.3(2.5)	**3.3(2.0)**	**4.4(2.3)**	3.6(2.3)	3.9(2.4)	3.1(2.2)	3.1(2.5)
E	3.2(2.3)	4.0(2.4)	2.4(2.3)	3.1(2.4)	3.5(2.3)	3.1(2.5)	*3.1(2.0)*	*4.3(2.4)*	3.9(2.4)	3.5(2.5)	3.0(2.1)	3.0(2.5)
F	4.0(2.2)	4.2(2.4)	3.1(2.6)	3.7(2.5)	3.6(2.2)	3.5(2.6)	3.6(2.3)	4.5(2.3)	3.5(2.3)	4.1(2.4)	*3.7(2.4)*	*2.9(2.5)*
G*	**3.5(2.4)**	**4.8(2.0)**	**3.3(2.6)**	**4.5(2.1)**	**5.1(1.8)**	**6.0(1.2)**	4.4(2.2)	4.2(2.3)	4.1(2.4)	3.7(2.3)	*4.7(2.1)*	*5.7(1.7)*
H	**3.1(2.4)**	**4.5(2.3)**	3.4(2.6)	4.1(2.3)	3.5(2.0)	3.2(2.5)	4.1(2.1)	4.0(2.5)	4.2(2.4)	4.0(2.5)	3.7(2.2)	3.0(2.5)
I	*3.1(2.3)*	*4.1(2.4)*	*3.2(2.6)*	*4.3(2.3)*	3.7(2.2)	3.1(2.4)	4.0(2.1)	3.9(2.3)	4.1(2.4)	4.0(2.5)	3.8(2.2)	3.0(2.5)
J*	**3.4(2.5)**	**4.8(2.0)**	3.6(2.6)	*4.5(2.0)*	**4.9(1.8)**	**6.0(1.3)**	4.6(2.2)	4.3(2.3)	3.9(2.4)	4.4(2.4)	**4.9(1.9)**	**5.8(1.5)**

Note: Bold-faced cells denote a significant difference between the importance ratings on session 1 (S1) and session 2 (S2), $p < .05$. Italics indicates trend toward significance $p < .08$, $n^2 > .10$. Asterisk (*) indicates key component.

A Neurocognitive Model for Students and Educators

Michael Atherton (athe0007@umn.edu)
Department of Educational Psychology, 178 Pillsbury Drive SE
Minneapolis, MN 55455 USA

Abstract

Computer metaphors for cognitive processes have become dated and new models are required to help college students and classroom teachers interpret research in the neurosciences as it begins to impact the fields of education and psychology A model of cognition using a metaphor of neural activation is presented and supported by finding in the neurosciences.

Introduction

In 1986, Rumelhart and McClelland published Parallel Distributed Processing and revolutionized our conception of memory and thought processes. Their introduction of neural networks as a model of memory served as a starting point for the growth of neurologically based computational models. Today, cognitive and computational neuroscience research is continuing to enhance our understanding of cognition. However, while the neurosciences influence how researchers comprehend thought processes, there has been little change in the models we teach students in introductory psychology and education classes. What is needed are models that provide students with a framework for understanding the neural bases of cognition and are at the same time simple to communicate and comprehend.

The Need for a New Model

The information processing model of cognition has provided us with many useful characterizations of mental functions: sensory registers, short-term memory, working memory, and long-term memory. Its ultimate failure has been its inability to integrate these characterizations with research findings in attention, imagery, and reasoning to form a comprehensive model of cognition. This failure is caused in part because theoretical descriptions of behavior are weakly constrained and allow multiple valid interpretations of the same phenomenon. This variability in description has fractionalized cognitive research into narrow domains focused on particular aspects of mental activity. Thus, students study cognitive topics such as perception, memory, and learning that have little relation to each other, leaving them without an associative framework.

Another problem with the information processing model is that its conceptualizations of thought processes often clash with research findings in contemporary neuroscience. It is now clear that memories are not compartmentalized into boxes or transferred from location to location as they are in a computer. Such characterizations can lead students to inferences that are not always valid, and in education such inferences can lead to instructional methods that are not always effective. In the past ten years technological and methodological advances in the neurosciences have produced a wealth of research results that have greatly increased our understanding of the biological underpinnings of cognition. This research has already impacted traditional cognitive theories (Miyake & Shah, 1999), but what is needed are not models that have been updated to account for the new data. What is needed are models that are built from the neuron upwards, rather than from behavior downwards. Such models stand a better chance of providing an internally consistent integrative framework for understanding cognitive research.

This revolution in orientation from top-down to bottom-up analyses represent a fundamental shift in the science of cognition and as Kuhn (1962, p. 109) has pointed out, "...when paradigms change, there are usually significant shifts in the criteria determining the legitimacy both of problems and of proposed solutions." Thus, we see major issues in popular culture such as the division between mind and body become irrelevant as old axioms are rejected and new ones formed (Crick, 1994). This paradigm shift is well underway in the field of psychology (Gazzaniga, 1998) and cognitive science, but has been hampered in education by the dubious application of the neurobiological research; to wit, "brain-based learning" has become the phrenology of the new century (Bruer, 1999a). The fact that such questionable conceptualizations of cognitive neuroscience are being actively marketed to educational practitioners begs for models that are well grounded by research in the neurosciences. The purpose of this paper is to present one such model in the hope that it will inspire discussion within the educational community. I will begin by presenting the model as it might be presented to students and the following section will review the scientific justification for the model.

The Model

The cognitive model presented below is a synthesis of the current research in the neurosciences. It is proposed as a descriptive theory in which the complexity of some neurological processes has been simplified for

pedagogical advantage. For the most part I have tried to avoid issues that are the subject of on going debate, however I do make theoretical assumptions at some points to create an internally consistent model. Readers should keep in mind that what follows is a hypothetical description and that some, or many, aspects of the model have yet to be verified empirically.

Structure

Neurons The fundamental processing element in the brain is the neuron. A neuron is a cell that consists of dendrites, a body, and an axon. Signals are transmitted between neurons by chemicals called neurotransmitters at junctions between dendrites and axons called synapses. When neurotransmitters pass across a synapse from the axon of one cell to the dendrite of another they cause chemical changes in the dendrite and the body of the neuron and have an effect on the physical structure of the synapse. When a sufficient number of signals infringe on a neuron, they cause that neuron to release neurotransmitters at the synapses of its axons. When a neuron is in such a state, is receiving signals and releasing neurotransmitters, it is said to be *active*. The signals that a neuron sends can have either an excitatory and inhibitory effect. Excitatory effects cause other neurons to increase their activity; inhibitory effects prevent neurons from becoming active, i.e., from sending signals to other neurons.

Networks Neurons are highly interconnected; a single neuron can have thousands of synapses connecting to thousands of other neurons. Neurons that activate in common are linked together by synapses into networks. Networks of neurons can be local to a particular location in the brain, or they can be global and distributed across different regions of the brain. Local networks can be confined to an area a thousandth of an inch, whereas global networks can extend over distances of several inches and be comprised of multiple local networks. Networks can also be classified as task networks or control networks. A task network performs a specialized function such as the processing of the orientation of a line, of a color, or a phoneme. Task networks are also involved in more complex processing, such as the graphical representation of a word, or a human face. A control network, on the other hand, regulates the processing of a task network via connections that inhibit or activate neurons. This activation or inhibition by control networks functions to maintain, select, monitor, sequence, or integrate activity within task networks.

Control networks can be either local or global. Local control networks regulate task networks and are not necessarily physically distinct from them, but instead maybe spatially intertwined. Global control networks serve to regulate the activation between multiple task networks. Within the brain local control and task networks are generally located in regions known as the occipital, parietal and temporal lobes. Global control networks are generally located in the frontal lobes (Figure 1). General cognitive functions, such as auditory or visual processing that are executed by task networks are localized to particular areas in the brain such as the temporal and occipital lobes. While task and control networks maybe localized to different regions this does not mean that they are disconnected, rather as a general rule all classes of networks are highly interconnected.

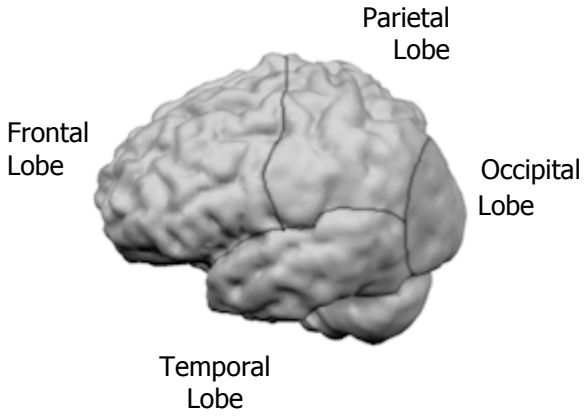


Figure 1: Global control networks are located in the frontal lobes; local task networks are located in the occipital lobes, parietal, temporal lobes.

Processes

With the neural and network structure outlined above traditional cognitive processes can be redefined in relation to their neurological processes. It is assumed that since thought processes are a function of brain activity, all of the concepts in traditional cognitive psychology can be redefined within such a framework. This paper covers fundamental topics such as learning, basic forms of memory, and attention.

Learning Learning is the process of making enduring changes in the relationships between neurons through the modification or creation of synapses. Learning is the intrinsic result of sensation, perception, thought, physical action, or the result of intention. Since thought or perception requires the activation of neurons, activation induces a change in the physical state of the synapses involved. When these changes are retained over time they form memories; when ephemeral they, along with the biochemical and electrical processes that cause them, form thoughts. Not all neurons exhibit the same properties; neurons and synapses differ in the rate at which changes are retained. Thus, long term physical changes in neurons located in areas that process sensory stimuli would occur slowly, whereas

changes in neurons storing life experiences would occur more rapidly.

Memory A memory is a stored pattern of synaptic connections within local networks or across an interconnected network of neurons. A single neuron might contribute to several memories by releasing varying amounts of neurotransmitters depending on how it is activated, but the activation of a memory requires interactions between neurons. That is to say, no one specific memory is stored in a single neuron, rather a memory is distributed across a network of neurons as a pattern of synaptic relations. Additionally, several memories may be stored in the same network by different patterns of synaptic connections.

Traditional cognitive models have identified different functional categorizations of memory. Traditionally, the major categorizations have been short-term, working, and long-term memory. Within the current model each of these can be defined by neurological processes. Short-term memory represents the dynamic patterns of activation in networks across the brain. These patterns of activation arise from transitory chemical and electrical properties of neurons lasting a few seconds. Working memory represents a combination of short-term memory processes along with other chemical and transient physical synaptic changes lasting several minutes. Long-term memory represents physical synaptic changes lasting from minutes to days, weeks, or years, thus long-term memory is viewed as scaled physical changes in synapses on a continuum across time, rather than as discrete storage locations. Physical changes in synapses may revert to a previous state in some types of neurons if not reactivated.

Memory Capacity Traditional short-term or working memory has been shown to have a limit on the number of items that can be held in consciousness simultaneously. In the current model memory capacity is regulated by both the physiological and structural aspects of neural networks. Physiologically neural activity is limited by the production and transmission of neurotransmitters and rate at which cells can produce signals. While neurophysiology places a limit on what can be held in neural networks at any single point in time, working memory capacity can be increased by adapting the underlying network structure for particular types of memories. Thus, experts in specialized areas such as chess have better working memory for meaningful configurations of pieces than do novices because they have tailored synaptic patterns to particular stimuli.

Forgetting In short-term memory forgetting occurs as cells stop sending signals and the concentration of neurotransmitters decays and returns to baseline. This process normally occurs within seconds unless activation is renewed. In long-term memory forgetting may occur when physical synaptic changes decay or as activation overlays new synaptic patterns over existing ones causing interference.

Attention Attention is the modulation of activation in a network by a control network. Modulation takes the form of either a rise or a reduction of activation. This modulation can occur across widely separated regions in the brain or locally within a specific part of the brain. Attention can be focused in different regions when a global control network or a combination of global control networks modulates activation within local task networks. Attention is a process which occurs when specific stimuli or tasks require specialized processing. Attentional processing may be required to keep particular memories active (maintenance), discriminate between similar stimuli (selection), when anticipating environmental stimuli (monitoring), when planning a particular action (sequencing), or when it is necessary to coordinate multiple responses (integration). Attentional modulation may be initiated from the bottom-up by neural signals originating in task networks that then activate global control networks, or from the top-down when global control networks execute a motivational goal.

Attention can also be focused in specific areas in the brain by local control networks without modulation by global control networks. This can occur within task networks when a local control network modulates activation. Many of the same functions executed by global control networks (maintenance, selection, monitoring, and sequencing) can be also be performed by local control networks. The process of transferring control from a global control network to a local one is called automation. Automation occurs by adjusting synaptic connections within a task network so modulation of activation is stimulated and responded to by activity within the task network. This adjustment of synaptic connections in task networks occurs through attentional modulation or by sensory activation. This is generally a slow process and may require both focused attention and repeated practice.

Retrieval from Memory Memory retrieval in traditional psychology is divided into two general classes: recall and recognition. Recall involves the explicit retrieval of information from memory. Recognition, on the other hand, only requires knowing if something has been previously encountered. In the current model retrieval is defined as the reactivation of a previously active network. Recall differs from recognition in that recall requires full activation of a local network by a global control network in a top-

down process, whereas recognition involves signaling a control network that some activation occurs in some local network and can occur as either a bottom-up or a top-down process.

An Activation Metaphor

Being able to imagine how thought processes occur in the brain can help teachers plan lessons and can also help undergraduate students tie together disparate concepts in psychology. The current model lends itself to a activation metaphor so that students can picture thought processes as dynamic intensifying and receding patterns of activation, much as neural activation is depicted by differences in blood oxygenation levels in the brain by functional magnetic resonance imaging (fMRI; see Figure 2).

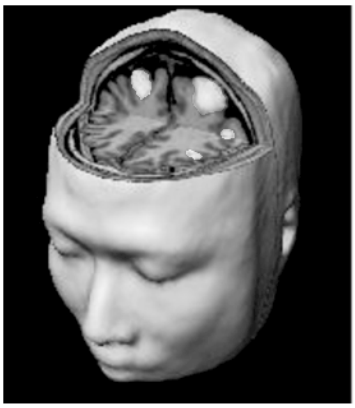


Figure 2: The brighter areas identify implied patterns of neural activation as detected by fMRI.

The processes of reading provides a good example of how levels of activation can be used to represent thought processes. When a word is first presented the visual areas in the rear of the brain become active. Activation is next seen in the word form area on the lower left side of the brain (Cohen, et al., 2000). At about this time there is an interaction between control networks in the left frontal areas and task networks in the lateral and posterior regions. There appears to be separate control networks for different aspects of reading such as grammar and those relating to semantics and these different networks appear to control specific task networks (Bokde, Tagamets, Friedman, & Horwitz, 2001; Poldrack, et al., 1999). Reading progresses as an interplay of activation between the control and task networks performing the functions of phonics, grammar, and semantics. Levels of activation in these areas can increase or decrease depending on the complexity of the task, when syntactic errors are encountered, or when semantic anomalies occur. Comprehension of longer passages of text appears to activate control and task networks on the right side of the brain (Robertson, et. al, 2000; St George, Kutas, Martinez, & Sereno, 1999).

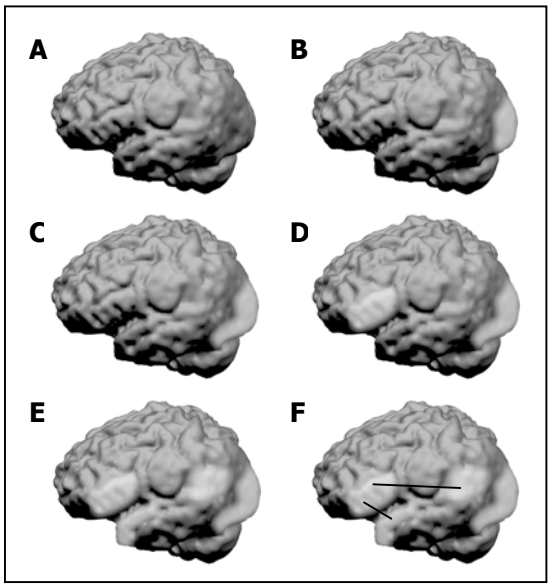


Figure 3: Idealized activation patterns during reading. (A) Displays the brain at rest. (B) Activation of visual areas. (C) Activation of the word form area. (D) Initial activation of frontal control networks. (E) Additional activation of phonological and semantic areas. (F) Interplay between control and task networks. The upper line indicates areas associated with phonological processing; the lower line with semantic processing.

While patterns of activation can provide an understanding of interacting networks, attention, short-term and working memory it should be kept in mind that long-term memory is the result by physical changes and is not necessarily reflected directly in the activation patterns, i.e., long-term memory is resultant effect of the activations.

Supporting Evidence

While some aspects of the model have not yet been verified empirically, most of the concepts are accepted by at least a portion of the neuroscientific community. In this section I will outline some of the research supporting the model.

The view that the changes in synaptic junctions are at least one of the major components of learning has been accepted for some time (Collingridge & Bliss, 1995; Larkman & Jack, 1995) and current research has been supportive of this hypothesis (Kennedy, 2000; Matus, 2000), however the exact mechanisms underling learning (Barinaga,1999) and forgetting (Berman & Dudai, 2001) remain the object of ongoing research. The assumption that learning occurs through the

formation of new synapses (Klintsova & Greenough,1999) is more debatable and has been questioned by Goldman-Rakic a leading researcher (as cited in Education Commission of the States, 1996, p. 11). A far more controversial conjecture has been the formation of memories via the creation of new neurons. Recent publications have indicated that, contrary to previous doctrine, new neurons are created after birth (Gould, Tanapat, Rydel & Hastings, 2000; Shankle, Rafii, Landing, & Fallon, 1999) and may also be involved in the formation of memories (Shors, et al., 2001), but these conclusions have been challenged (Kornack & Rakic, 2001).

The interpretation of short term memory as dynamic biochemical and electrophysiological processes is hypothetical, but research has shown that transmission between nerve cells is separable into different chronological processes (Greengard, 2001). The concept of working memory extending over longer periods of time was introduced by Ericsson and Kintsch (1995) and is support by neuroscience research showing dendritic changes occurring on a continuum ranging from seconds to days (Antonova, et al., 2001; Wong & Wong, 2001)

The existence of large scale modularity in the brain has been recognized since the 1800s when it was discovered by Broca and Wernicke (Gazzaniga, Ivry & Mangun, 1998). The high degree of specificity in local networks has been a more recent finding. Spatially limited networks, characterized by the current model as task networks, have been found to subserve functions such as: the detection and orientation of lines (Hubel & Wiesel,1968), the motion of patterns (Movshon, Adelson, Gizzi, & Newsome, 1985), the recognition of objects (Tanaka, 1997), and the recognition of faces (O'Craven & Kanwisher, 2000). The existence of global networks is widely accepted in the neuroscience and cognitive communities (Stuss & Alexander, 2000; Varela, Lachaux, Rodriguez, & Martinerie, 2001); as is the recognition of networks that modulate attention, although agreement on the specific mechanisms of modulation may differ (Driver & Frith, 2000; Posner & Rothbart, 1998; Rogers, Andrews, Grasby, Brooks & Robbins, 2000). The concept of integrative networks can be thought of as a reformulation of accepted views of the functioning of working memory in the frontal lobes (Duncan & Owen, 2000; Levy & Goldman-Rakic, 2000), but once again, the opinions on the exact organization and mechanisms differs.

Conclusions

Given the advances in the neurosciences it is now both necessary and advantageous to formulate cognitive models based on neurological processes rather than on metaphors derived from other disciplines. Inevitably, we will see many new comprehensive models linking cognition to its neurological foundations. This paper represents an initial attempt to formulate such a model in the hope that it can be used as a pedagogical tool for students and teachers.

Acknowledgements

I would like to express gratitude to Hedy Amiri, Patty Costello, Laird Edmans, Christine Ng, Maria Sera, Kelly Snyder, Al Yonas, Jazmin Yomha-Cevasco, the members of the Text and Discourse Comprehension Group at the University of Minnesota, and the reviewers assigned to my paper by the Cognitive Science Society for their comments, suggestions, and corrections.

References

Antonova, I., Arancio, O., Trillat, A. C., Wang, H. G., Zablow, L., Udo, H., Kandel, E. R., & Hawkins, R. D. (2001). Rapid increase in clusters of presynaptic proteins at onset of long-lasting potentiation. *Science, 294,* 1547-1550.

Atherton, M. & Bart, W. M. (2001, April). Education and fMRI: Promise and Cautions. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Barinaga, M. (1999). Neurobiology - New clues to how neurons strengthen their connections. *Science, 284,* 1755-1757.

Berman, D. E. & Dudai, Y. (2001). Memory extinction, learning anew, and learning the new: Dissociations in the molecular machinery of learning in cortex. *Science, 291,* 2417-2419.

Bokde, A. L. W., Tagamets, M. A., Friedman, R. B., & Horwitz, B. (2001). Functional interactions of the inferior frontal cortex during the processing of words and word-like stimuli. *Neuron, 30,* 609-617.

Bruer, J. T. (1998). Let's put brain science on the back burner. *NASSP Bulletin*, 82(598), 9-19.

Bruer, J. T. (1999a). In search of...brain-based education. *Phi Delta Kappan, 80,* 648-654.

Bruer, J. T. (1999b). *The myth of the first three years: A new understanding of early brain development and lifelong learning.* New York: The Free Press.

Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain, 123,* 291-307.

Collingridge, G. L., & Bliss, T. V. (1995). Memories of NMDA receptors and LTP. *Trends in Neurosciences, 18*, 54-56.

Driver, J., & Frith, C. (2000). Shifting baselines in attention research. *Nature Reviews Neuroscience, 1,* 147-148.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23,* 475-483.

Education Commission of the States. (1996). Bridging the gap between neuroscience and education. Denver, CO: Author.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211-245.

Gazzaniga, M. (1998). *The mind's past.* Berkeley, CA: University of California Press.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (1998). *Cognitive neuroscience: The biology of the mind.* New York: W.W. Norton & Company.

Gould, E., Tanapat, P., Rydel T., & Hastings, N. (2000). Regulation of hippocampal neurogenesis in adulthood. *Biological Psychiatry, 48,* 715-720.

Hubel, D. H., & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology, 195,* 215-243.

Klintsova, A. Y., & Greenough, W. T. (1999). Synaptic plasticity in cortical systems. Current Opinion in *Neurobiology, 9,* 203-208.

Lagemann, E. C. (2000). *An elusive science: The troubling history of educational research.* Chicago: The University of Chicago Press.

Larkman, A. U., & Jack, J. J. (1995). Synaptic plasticity: hippocampal LTP. *Current Opinion in Neurobiology, 5*, 324-334.

Levy, R., & Goldman-Rakic, P. S. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. *Experimental Brain Research, 133,* 23-32.

Kennedy, M. B. (2000). Signal-processing machines at the postsynaptic density. *Science, 290*, 750-754.

Kornack, D. R., & Rakic, P. (2001). Cell proliferation without neurogenesis in adult primate neocortex. *Science, 294*, 2127-2130.

Matus, A. (2000). Actin-based plasticity in dendritic spines. *Science, 290,* 754-757.

Miyake, A., & Shah, P. (1999). Toward unified theories of working memory: Emerging general consensus, unresolved theoretical issues, and future research directions. In A. Miyake & P. Shah (Eds.) *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 442-481). New York: Cambridge University Press.

Movshon, J. A., Adelson, E. H., Gizzi, M. & Newsome, W. T. (1985). The analysis of moving visual patterns. In C. Chagas, R. Gattass & C. G. Gross (Eds.), *Pattern recognition mechanisms* (pp. 117-151). Rome: Vatican Press.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12,* 1013-1023.

Poldrack, R. A., Wagner, A. D., Prull, M. W., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage, 10,* 15-35.

Posner, M. I., & Rothbart, M. K. (1998). Attention, self-regulation and consciousness. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences, 353,* 1915-1927, 1998.

Pugh, K. R., Mencl, W. E., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., Shaywitz, S. E., & Shaywitz, B. A. (2000). Functional neuroimaging studies of reading and reading disability (developmental dyslexia). *Mental Retardation and Developmental Disabilities Research Reviews, 6*, 207-213.

Robertson, D. A., Gernsbacher, M. A., Guidotti, S. J., Robertson, R. R., Irwin, W., Mock, B. J., & Campana, M. E. (2000). Functional neuroanatomy of the cognitive process of mapping during discourse comprehension. *Psychological Science, 11,* 255-260.

Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience, 12,* 142-162.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Exploration in the microstructure of cognition, Vol. 1.* Cambridge, MA: MIT Press.

Shankle, W. R., Rafii, M. S., Landing, B. H., & Fallon, J. H. (1999). Approximate doubling of numbers of neurons in postnatal human cerebral cortex and in 35 specific cytoarchitectural areas from birth to 72 months. *Pediatric & Developmental Pathology, 2,* 244-259.

Shors, T. J., Miesegaes, G., Beylin, A., Zhao, M., Rydel, T., & Gould E. (2001). Neurogenesis in the adult is involved in the formation of trace memories. *Nature, 410,* 372-376.

Sternberg, R. J., & Grigorenko, E. L. (2001). Unified psychology. *American Psychologist, 56,* 1069-1079.

St George, M., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic integration in reading: engagement of the right hemisphere during discourse processing. *Brain, 122,* 1317-1325.

Stuss, D. T., & Alexander, M. P. (2000). Executive functions and the frontal lobes: a conceptual view. *Psychological Research, 63*, 289-298.

Varela, F., Lachaux, J. P., Rodriguez ,E. & Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience, 2,* 229-239.

Wong, W. T., & Wong, R. O. L. (2001). Changing specificity of neurotransmitter regulation of rapid dendritic remodeling during synaptogenesis. Nature *Neuroscience, 4,* 351-352.

Do People Update Spatial Relations Described in Texts?

Marios N. Avraamides (Marios@Psu.Edu)
Department of Psychology,
The Pennsylvania State University
604 Moore Building
University Park, PA 16801 USA

Abstract

Previous studies (e.g., Rieser, 1989) have established that physical rotations result in effortless updating of spatial information contained in visually perceived scenes. The present experiment provided evidence that this is not the case for scenes that are encoded through texts. Performance was better under the perspective that the scenes were learned than under novel perspectives, regardless of whether rotations were physical or imagined. In addition, the experiment suggested that the orientation of the ecological frame affects spatial performance even when people operate in a purely represented framework.

Introduction

Various studies (e.g., Chance, Gaunet, Beall, & Loomis, 1998; Klatzky, Loomis, Beall, Chance, & College, 1998; Rieser, Guth, and Hill, 1986; Simons & Wang, 1998; Wang & Simons, 1998) have established a link between active locomotion and the successful spatial updating of previously viewed scenes. Researchers have proposed that spatial updating is carried out by an internal mechanism that constantly computes the relative locations of objects as people move in the environment (Wang & Spelke, 2000). Indeed, when people change their position within an environment they seem to experience little difficulty with keeping track of how the locations of objects change, even when the objects are no longer in view (Attneave & Farrar, 1977; Fukusima, Loomis, & Da Silva, 1997; Loomis, Klatzky, Colledge, Cicinelli, and Pellegrino, 1993). In fact, responding to locations of objects after locomoting to a new position is no more difficult than doing so from the original standpoint (Rieser, 1989; Rieser et al., 1986). However, difficulties arise when judgments are made from perspectives that are imagined (Presson & Montello, 1994). For example, Rieser has shown that pointing to objects from a novel standpoint is slower and less accurate when the observer mentally rotates to the new standpoint than when she is physically rotated to it.

In short, the evidence suggests that moving organisms update their representations of their surroundings in an efficient and effortless manner. However, non-moving organisms need to engage in additional mental processing in order to reason about their surroundings from imagined standpoints.

The majority of the previous studies have examined spatial updating with paradigms that involved visually presented stimuli. Visual perception is not, however, the only way we form mental representations of the world. Very often, we learn about space by reading or listening to verbal descriptions. A number of studies (e.g., Denis & Cocude, 1989) have shown that at least in terms of geometrical properties, mental representations constructed from language are equivalent to those created through perception. In general, mental representations of space seem to preserve many of the characteristics of real environments (Zwaan & Radvanksy, 1998).

An interesting issue that arises is whether mental representations derived from text are updated when changes in the spatial relations come about. A study by De Vega and Rodrigo (2001) attempted to provide an answer to this question by contrasting how easily people locate objects after physical and imagined rotations. In that study, participants first read sentences that described spatial layouts in which objects were located at each of the four canonical horizontal directions from a protagonist. The protagonist was then described to rotate to novel perspectives. The subjects were probed with the names of the objects and were asked to determine their locations from the perspective of the protagonist. One group of participants performed the task by physically rotating along with the protagonist, while a different group performed the task with no physical rotation. De Vega and Rodrigo contrasted the judgment latencies of the two groups to assess whether spatial updating took place. In one experiment that required that subjects use spatial labels (i.e., front, left etc) to provide their answers, performance was equally fast for the two rotation modes. In another experiment in which subjects pointed to objects, performance was faster when rotation was physical instead of imagined. De Vega and Rodrigo concluded that physical rotations led to effortless updating only in the pointing experiment. Furthermore, they suggested that the actual body position of the participant is not important for performing the task when responses are made by using spatial labels. That is because subjects performed the task in a represented framework from which their actual self was disengaged.

Certain limitations of the study by De Vega and Rodrigo (2001) create concerns regarding the

interpretation of the results. First, the absence of a difference between physical and imagined rotation latencies in the labeling task does not necessarily mean that participants failed to effortlessly update their representations. It could very well be the case that they were successful at updating their representations under both modes of rotation. This is quite possible given the simplicity of the scene and its relatively low working memory demands (only four objects needed to be tracked). Second, although no differences were observed in the latency measures, subjects were significantly more accurate in their judgments with physical rotations than with imagined rotations.

The present study uses a different measure to examine whether spatial updating takes place under either imagined or physical rotations. Performance under the original perspective (i.e. the perspective from which the scene is encoded) and novel perspectives is contrasted in the two modes of rotation. If subjects fail to update their original representation when they rotate – either physically or imaginary – to a novel orientation, then performance should be better when the task is performed from the original than from the novel perspectives. If any of the two modes of rotation results in successful updating, this should be indicated by equal performance between the original and novel perspectives.

A second goal of the study is to examine more closely the role – if any – of the participants' ecological frame in tasks that require judgments using deictic terms. A hypothesis of the present study is that the orientation of ecological frame affects how easily people discriminate the two poles of an axis and map spatial terms to the appropriate regions of space. If this is true then the orientation of the participants' bodies will affect the ease of using deictic terms in the task. This should be particularly true for judgments within the left/right axis because our bodies provide the only source of asymmetry for this axis. Manual dexterity is probably the least subtle cue that people can use to discriminate left from right (Corballis & Beale, 1976). Therefore, the prediction is that left Vs right judgments will be less difficult when the ecological reference frame of the participant is aligned with the reference frame imposed on the protagonist.

Experimental Task

The present study uses a paradigm that has been used widely in the past to study the accessibility of directions in spatial memory. The task, developed by Franklin and Tversky (1990), involves the presentation of a narrative which describes a naturalistic scene in the second person. Objects are described occupying positions at canonical directions from the protagonist (i.e., above, in the front, on the left etc). Participants are given unlimited time to study the narrative. Then, the protagonist is described to rotate to a new orientation. The narrative continues on a sentence by sentence fashion with participants pressing a key to request a new sentence. Occasionally, instead of a new sentence, the name of an object appears. Participants are asked to report where the object is with respect to the current perspective of the protagonist. The task continues until all objects are probed from various perspectives.

Studies that used this paradigm (e.g., Bryant & Tverksy, 1992) were primarily focused on the accessibility pattern for the various self-object directions. Therefore, latencies from the various protagonist perspectives were collapsed to provide a single average value for each direction. The major result from these studies is that objects on the above/below axis are located faster than objects on the front/back axis, which are in turn located faster than objects on the left/right axis. Furthermore, objects in the front are retrieved faster than objects at the back of the protagonist. An account for this accessibility pattern has been provided by the Spatial Frameworks model (Franklin & Tversky, 1990) which posits that the scene is represented on the basis of the three orthogonal body axes.

The current study uses the Spatial Frameworks paradigm to examine latency differences – if any- between the original and the novel perspectives of the protagonist. As in De Vega and Rodrigo (2001), a mode of rotation manipulation is introduced. A group of participants perform the task by physically rotating to novel perspectives while a different group of participants perform the task by only imagining rotations to novel perspectives. In contrast to the narratives of De Vega and Rodrigo, the present study involves narratives that are somewhat more complex. In addition to the four objects in the horizontal plane, another two objects are described to occupy the poles of the above/below axis. Because all protagonist reorientations occur in the horizontal plane, the positions of these two objects do not need to be updated at all throughout the task,. Nevertheless, they add an extra load on working memory, thus making the task more cognitively demanding.

In order to examine the role of the ecological frame, another variable is introduced. Half of the narratives require that subjects respond with a direction judgment (e.g., left, front etc) while in the other half they simply need to respond with an axis judgment. In the latter case, if for example the object is located on the left of the protagonist, the participant only need to respond with "left/right". This manipulation has been previously used with this paradigm in a study by Bryant and Wright (1999). The rationale for including the judgment type manipulation in the present study is that if the ecological reference frame helps to make easier direction judgments, then simply removing the need for such judgments will produce the same effect. Furthermore, if real rotations and axis judgments affect

performance in the same way, then having one of them should be sufficient; that is, no better performance will be observed when both of them apply instead of just one.

Method

Participants Forty students (20 females) from psychology classes at the Pennsylvania State University participated in the experiment in exchange for course credit. Ten male and 10 female participants were randomly assigned to the physical and the imagined rotation conditions.

Materials Four narratives, taken from those used by Franklin and Tversky (1990) and two taken from Bryant & Wright (1999) were used in the present study. The first portion of the narratives described in the second person a naturalistic scene -- a barn, a construction site, a hotel lobby, a space museum, a lagoon, and a navy ship –in which objects were located at the six canonical axes of the reader-protagonist. Narratives were modified to include 6 instead of 5 objects and in contrast with Franklin and Tversky, the initial portions were also presented on computers. The six critical objects in each narrative appeared in blue upper-case characters, in contrast to the rest of the narrative which appeared in lower-case black characters. The order in which objects were introduced in the initial portions of the narratives was determined with the use of a 6 x 6Latin Square so that each object direction appeared at a different serial position in each narrative. For each participant, half of the narratives were randomly assigned to the direction judgment condition and the other half to the axis judgment condition. Narratives were presented to participants in a random order with the constraint that no more than 2 narratives of the same judgment type were presented in sequence.

Design The experiment was a 2 (mode of rotation: imagined, real) x 2 (type of judgment: direction, axis) x 2 (perspective: original, novel) x 6 (self-object directions: above, below, front, back, left, right) mixed factorial design. The type of rotation was manipulated between subjects while both the type of judgment and self-object directions varied within subjects. Each self-object direction was tested four times in each narrative, once under each of the four possible perspectives. Self-object directions were tested in a random order within each perspective. Also, the original perspective was never the first perspective that subjects were tested on.

Procedure The procedure was similar to the one used by Franklin and Tversky (1990). Two main differences from De Vega and Rodrigo (2001) were that a voice key was not used to collect responses and that the narratives included objects located on the above/below axis. The narratives were presented on a laptop computer that participants held onto their laps while sitting in a swivel chair. Participants were given unlimited time to study the first portion of the narratives. The narrative then continued in a sentence by sentence fashion with subjects pressing the space bar to request a new sentence. The second sentence after the first portion described the protagonist rotating to a new orientation in the horizontal plane. Depending on the condition they were assigned to, participants were instructed to either turn their selves by swiveling the chair to produce the reorientations that were described in the text, or to imagine their selves rotating without changing their actual orientation at any point throughout the experiment. After two filler sentences were presented, the name of one of the objects in the scene appeared. Subjects were instructed to press the space bar as soon as they were ready to report the relative position of the object. The time between the appearance of the object probe and the space bar press was the critical latency measure and will hereinafter be referred to as *response latency*. For direction judgments, when subjects pressed the space bar, a list with the terms “above”, “below”, “front”, “back”, “left”, and “right”, presented in a random order each time, appeared. Participants were instructed to press the key that corresponded to the integer (1 to 6) that appeared next to the direction they chose. For axis judgments, the list contained the choices “above/below”, “front/back”, and “left/right”, and participants entered their answer using the 1 to 3 keys. The time between the appearance of the list and the key press for selecting the direction will be referred to as *answer latency*. Each narrative continued in the same fashion until all six objects were probed in each protagonist perspective. Each narrative involved 24 probes, 6 for each protagonist perspective. For each participant, three of the narratives were randomly assigned to the direction judgment type, and the rest to the axis judgment type. Below each object probe the words “direction” or “axis” were presented to remind participants of the type of response they needed to make.

Results

Participants responded correctly to 96.3% of the probes. Data from 2 subjects -- one from each rotation condition -- were discarded because accuracy did not exceed 60%.

Latency Incorrect responses were discarded from the latency analyses. Outliers in latency data were defined as reaction times deviating more than 3 standard deviations from each participant’s type of judgment mean. Outliers in response RT resulted in an additional 6% of data and were also discarded from the analyses. Latency data for each self-object direction were collapsed to form three dimension means (i.e., a

separate mean of each body axis). Moreover, latencies for the three novel perspectives did not differ from each other so they were collapsed to form a single novel perspective mean.

Answer Latency Analyzing answer RT provides a check on whether subjects successfully followed instructions on how to respond. Frequently, spillover effects of response RT on answer RT are observed. When this happens, however, answer RT patterns tend to mimic the ones observed in the response RT data and are either omitted from analyses (e.g., Bryant, Tversky, & Franklin, 1992) or combined with response RT (Franklin, Tversky, & Coon, 1992). No spill-over effects were observed in the present study. The only significant effect was a main effect for the type of judgment, $F(1,36)=90.28$, $MSE=105394$, $p<.001$. The average answer RT for direction judgments was greater than the corresponding RT for axis judgments (1944 ms and 1655 ms respectively). This is expected because the response choices were 3 for axis judgments and 6 for direction judgments.

Response Latency Several Analyses of Variance (ANOVA's) were performed. Interactions that are not discussed were not significant at $a=.05$.

As predicted, a significant mode of rotation x perspective x dimension interaction was obtained, $F(2,72)=4.23$, $MSE=75530$, $p<.05$ (figure 1).

Separate ANOVAS were performed for each mode of rotation.

When rotation was imagined, a significant perspective x dimension interaction was obtained, $F(2,36)=8.05$, $MSE=114939$, $p<.01$. Significant main effects for both perspective and dimension were obtained, $F(1,18)=22.93$, $MSE=264189$, $p<.001$ and $F(2,36)=19.57$, $MSE=212008$, $p<.001$ respectively. Performance was faster with the original than the novel perspective for above/below and left/right. For front/back the difference was not statistically significant, $F(1,18)=2.73$, $MSE=193450$, $p=.12$. Nevertheless, even in this dimension the average for the original perspective was smaller than that for novel perspective (2127 ms and 2363 ms respectively). Overall, the patterns of accessibility of the three axes conformed to the predictions of the Spatial Frameworks model for both the original and novel perspectives. Judgments on the left/right dimension were particularly slow when performed under a novel perspective (3025 ms) than under the original perspective (2208 ms), $F(1,18)=32.83$, $MSE=193290$, $p<.001$.

When rotation was physical, latency for the original perspective was shorter than the average latency for the novel perspective (1821 ms and 2149 ms respectively), $F(1,18)=57.26$, $MSE=53699$, $p<.001$. The perspective x dimension interaction was not significant, $F(2,36)=1.99$, $MSE=36120$, $p=.15$. However, there was a main effect of dimension, $F(2,36)=21.61$, $MSE=55533$, $p<.001$. The pattern of latencies for the three body axes corresponded with the pattern predicted by the Spatial Frameworks model.

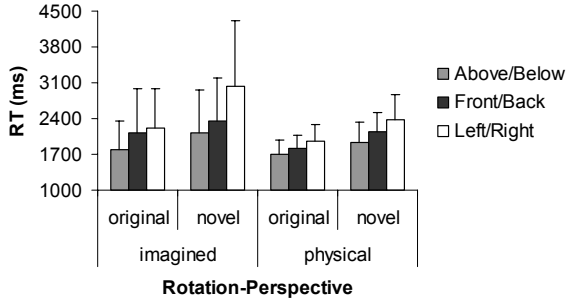


Figure 1: Latency for body axes as a function of the mode of rotation and the perspective of the protagonist.

Additionally, the analysis revealed a mode of rotation x judgment type x dimension interaction, $F(2,70)=3.49$, $MSE=60334$, $p<.05$.

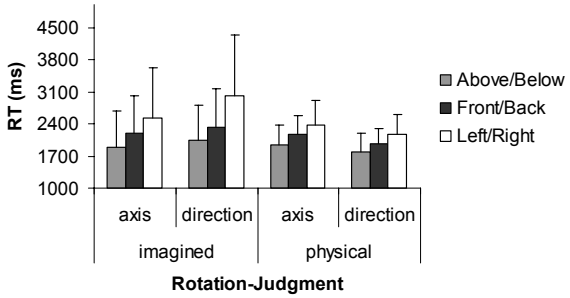


Figure 2: Latency for body axes as a function of the mode of rotation and the judgment type.

Further analyses were performed separately for each judgment type. When judgment was direction, a significant rotation x dimension interaction was obtained, $F(2,72)=6.61$, $MSE=134329$, $p<.01$. Overall with direction judgments, subjects were faster when the rotation was physical than imagined (2507 ms and 1979 ms respectively), $F(1,36)=5.48$, $MSE=1448481$, $p<.05$. An effect for dimension was also present, $F(2,72)=31.49$, $MSE=134329$, $p<.001$. Latency patterns conformed to the Spatial Frameworks model for both rotation modes. However, times were longer for all dimensions when rotation was imagined. The interaction was caused by a very long latency average for the left/right dimension under imagined rotations. When the judgment was axis, latencies for the various axes were similar for both types of rotation. The rotation x dimension interaction did not approach significance, $F(2,7)=1.23$, $MSE=84641$, $p=.30$. Similarly, no main effect for rotation mode was obtained, $F(1,35)=.07$, $MSE=1353673$, $p=.80$. However, a significant main effect for dimension was

obtained, F(2,70)=30.96, MSE=84641, p<.001. The Spatial Frameworks pattern was observed in the dimension latencies. As seen in figure 2, the average latency for the left/right dimension (2452 ms) was shorter than the respective average of imagined rotation with a direction decision (3045 ms) but longer than that of real rotation with a direction decision (2170 ms).

An interesting result was the pattern observed for the rotation x judgment interaction, F(1,35)=11.66, MSE=216901, p<.01. When rotation was imagined, axis judgments (2212 ms) were faster than directional judgments (2463 ms), F(1,17)=8, MSE=212424, p<.05. However, when rotation was physical, the pattern of judgment times were actually opposite than the one obtained with imagined rotations. Axis judgments took longer than direction judgments (2155ms and 1979ms respectively). This difference was marginally significant, F(1,18)=4, MSE=221130, p=.06. A further analysis revealed that a significant type of judgment x perspective interaction for imagined rotations, F(1,17)=5, MSE=24966, p<.05. For novel perspective, direction judgments were performed significantly faster than axis judgments, F(1,17)=13.6, MSE=56705, p<.01. However, for the original perspective the two judgment types were not statistically different, F(1,17)=0.95, MSE=150568, p=.34.

Accuracy In contrast to the latency analysis, the ANOVA on accuracy revealed a significant main effect for the mode of rotation, F(1,36)=4.41, MSE=.017, p<.05. Accuracy was higher when rotations were physical instead of imagined (98% and 94% respectively). An advantage for physical rotation is also reported by De Vega and Rodrigo (2001). A significant mode of rotation x type of judgment interaction, F(1,36)=5.32, MSE=.017, p<.05, revealed that the effect of rotation was confined to direction judgments.

As in the latency analysis, a significant mode of rotation x perspective x dimension interaction was obtained, F(2,72)=3.84, MSE=.0008, p<.05. The mode of rotation x judgment x dimension interaction that was obtained in the latency analysis, was marginally significant, F(2,72)=3.07, MSE=.002, p=.05. For both interactions, the pattern of results resembled the patterns obtained in the latency analysis to a great extent. The only deviation from the latency data was that sometimes the judgments for front/back were no less accurate than judgments for above/below.

Discussion

The present experiment produced results that deviated from those reported by De Vega and Rodrigo (2001). When participants made direction judgments – the condition tested by De Vega and Rodrigo – they were both faster and more accurate with physical than imagined rotations. De Vega and Rodrigo reported that their subjects were more accurate but not faster with physical rotations. Perhaps, the use of only four objects in their narratives made it easy for participants to keep track of how their relative locations of objects changed with the protagonist reorientations.

Although the performance advantage for physical rotation would qualify as evidence for successful updating under De Vega and Rodrigo's (2001) methodology, other results from the present study suggest otherwise. Results suggest that the fastest performance under physical rotations was due to the relative ease with which direction judgments, especially left Vs right, were made when rotation was physical. The same result was obtained for imagined rotation when axis judgments were introduced. Indeed, when participants were not required to make a direction judgment, their performance was similar to that of participants with physical rotations. This suggests that the orientation of the ecological reference frame is important for making difficult discriminations between the poles of body-centric axes. The differences between the three dimensions were smaller when the ecological frame was aligned with the frame of the protagonist. This was true both when participants performed the task from the original perspective, under which their ecological frame is aligned with the protagonist frame, in either mode of rotation.

Despite the performance advantage with physical rotations, the present results provide clear evidence that participants did not update the self-object direction with either imagined nor physical rotations. In both cases, performance was both faster and more accurate when the task was performed from the original than a novel perspective. This result is not obtained in studies that use visually presented scenes (e.g., Rieser et al., 1986), and suggests one aspect that mental representations derived from texts might differ from those derived from perception. A possible account for this dissociation is provided by De Vega and Rodrigo (2001). While visually perceived scenes are anchored in a sensorimotor framework based on the ecological reference frame of the observer, mental representations of described scenes are grounded into a mental framework from which the self is detached. While physical movements can provide proprioceptive feedback that helps updating the mental representation in the sensorimotor framework, this feedback is not very helpful for representations that are anchored in a mental framework. However, as shown in the present study, although the ecological self might be disengaged from the mental framework, it is still important for making spatial decisions in it.

Finally, the accessibility patterns for the three dimensions conformed to the predictions of the Spatial Frameworks model (Franklin & Tversky, 1990). Objects on the above/below axis were located faster

than objects on the front/back axis, which were in turn located faster than objects on the left/right axis. The differences among the dimensions were greater when rotations were imagined and judgments were for direction. While this result might be due to the higher difficulty of performing the task under these circumstances, it could alternatively mean that a part of the Spatial Frameworks effect is due to the need for discriminating the poles of the axes. Results from Bryant and Wright (1999) suggest that the difficulty of making discriminating decisions within the axes does not fully account for the Spatial Frameworks results.

In summary, the present study provided more substantial evidence to confirm the conclusions of De Vega and Rodrigo (2001), while at the same time provided a better understanding of how the ecological self interacts with spatial reasoning about imagined spaces derived from texts. While physical rotations led to no spatial updating of the original mental representation, they produced better performance by reducing the difficulty with making discrimination between the poles of body-centric axes. This result can have important practical implications for situations where spatial reasoning occurs from imagined perspectives (e.g., teleoperating robots for rescue mission and space exploration).

Acknowledgments

I am grateful to Barbara Tversky for providing narrative material and Jessica Glick for modifying that material for use in the present study. For valuable discussions, I thank Rich Carlson, Frank Ritter, Judy Kroll, and Lael Schooler. I also thank Allison DeGrano, Tom Jolly, and Jessica Glick for their enthusiastic help with data collection.

References

Attneave, F., & Farrar, P. (1977). The visual world behind the head. *American Journal of Psychology, 90*, 549-563.

Bryant, D. J., Tversky, B., & Franklin, N. (1992). Internal and external spatial frameworks for representing described scenes. *Journal of Memory and Language, 31*, 74-98.

Bryant, D. J., & Wright, G. W. (1999). How body asymmetries determine accessibility in Spatial Frameworks. *The quarterly journal of experimental psychology, 52A*, 487-508.

Chance, S. S, Gaunet, F., Beall, A. C, & Loomis, J. M. (1998). Locomotion mode affects the updating of objects during travel: The contribution of vestibular and proprioceptive inputs to path integration. *Presence, 7*, 168-178.

Corballis, M. C., & Beale, I. L. (1976). *The psychology of left and right*. Hillsdale, NJ: Lawrence Erlbaum.

De Vega, M., & Rodrigo, M. J. (2001). Updating spatial layouts mediated by pointing and labelling under physical and imaginary rotation. *European Journal of Cognitive Psychology, 13*, 369-393.

Denis, M., & Cocude, M. (1989). Scanning visual images generated from verbal descriptions. *European journal of cognitive psychology, 1*, 293-307.

Franklin, N., & Tversky, B. (1990). Searching Imagined Environments. *Journal of Experimental Psychology: General, 119*, 63-76.

Franklin, N., Tversky, B., & Coon, V. (1992). Switching points of view in spatial mental models. *Memory & Cogntion, 20*, 507-518.

Fukusima, S. S, Loomis, J. M, & Da Silva, J. A. (1997). Visual perception of egocentric distance as assessed by triangulation. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 86-100.

Klatzky, R. L, Loomis, J. M, Beall, A. C, Chance, S. S, & Golledge, R. G. (1998). Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychological Science, 9*, 293-298.

Loomis, J. M, Klatzky, R. L, Golledge, R. G, Cicinelli, J. G, Pellegrino, J. W, & Fry, P. A. (1993). Nonvisual navigation by blind and sighted: Assessment of path integration ability. *Journal of Experimental Psychology: General, 122*, 73-91.

Presson, C. C, & Montello, D. R. (1994). Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception, 23*, 1447-1455.

Rieser, J. J, Guth, D. A, & Everett, W., & Hill, E. W. (1986). Sensitivity to perspective structure while walking without vision. *Perception, 15*, 173-188.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*, 1157-1165.

Simons, D. J, & Wang, R. F. (1998). Perceiving real world viewpoint changes. *Psychological Science, 9*, 315-320.

Wang, R. F., & Simons, D. J. (1998). Active and passive scene recognition across views. *Cognition, 70*, 191-210.

Wang, R. F., & Spelke, E. S. (2000). Updating egocentric representations in human navigation. *Cognition, 77*, 215-250.

Zwaan, R. A, & Radvansky, G. A. (1998). Situation models in Language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

An Exploration of Real-World Analogical Problem Solving in Novices

Christopher R. Bearman (c.bearman@lancaster.ac.uk)
Department of Psychology, Fylde College, Lancaster University,
LA14YD, UK

Linden J. Ball (l.ball@lancaster.ac.uk)
Department of Psychology, Fylde College, Lancaster University,
LA14YD, UK

Thomas C. Ormerod (t.ormerod@lancaster.ac.uk)
Department of Psychology, Fylde College, Lancaster University,
LA14YD, UK

Abstract

Despite the postulated importance of analogising to human cognition, the study of analogical problem solving in the laboratory has found disappointing results. Providing an analogue to a participant prior to asking them to solve a problem gives only a small benefit at best. Recently, studies outside the laboratory have suggested that experts frequently use analogies in real-world situations. It is less clear whether novices can also spontaneously invoke and use analogies to solve realistic problems. In the current investigation, undergraduates were observed solving a large-scale management problem over two weeks. It was found that many analogies were produced (on average 4.6 per one-hour session), and that 77% of these analogies reflected a structural rather than a superficial mapping between a base and a target. It was also determined that 56% of these structural analogies involved non-trivial mappings of higher-order relations. Further, it was found that analogies were drawn to serve two different purposes: problem solving and illustration. In generating illustrative analogies, participants frequently made superficial mappings, but when generating analogies to solve problems, they never made purely superficial mappings.

Analogy is defined by the Oxford English Dictionary as the "resemblance of relations or attributes as a ground of reasoning." The ability to draw analogies is considered to be fundamental to intelligent human behaviour (Holyoak & Thagard, 1995). Analogy appears to be central to both the processing of information and its retrieval from long-term memory (Shank, 1999) and underpins theories of creative thinking and intelligence (Holyoak & Thagard, 1995; Poze, 1983; Raven, 1938; Sternberg, 1977).

The postulated importance of analogising for intelligent human behaviour stands in contrast to the findings from laboratory studies of analogical problem solving. It is often found that providing a superficially dissimilar but structurally analogous source to a participant prior to them solving a target problem, without a hint about either its relevance or relatedness to the target, gives little or no gain in target solution rates (Anoli, Antonietti, Crisafulli & Cantoia, 2001; Gick & Holyoak, 1980, 1983).

Transfer of the analogous solution without a hint is not improved by: providing a static diagrammatic representation of the structural analogue (Gick & Holyoak, 1983; Podone, Hummel & Holyoak, 2001); providing an abstract verbal statement of the underlying structural analogue (Gick & Holyoak, 1983); or simply re-presenting the source to the participant while they are processing the target (Anoli et al., 2001; Gick & Holyoak, 1980). There is some evidence of analogical transfer when multiple analogous sources are provided (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983), and also when the analogous source is a general concept (Schunn & Dunbar, 1996). Generally, however, the findings from laboratory-based studies of analogy have been disappointing.

Recently, analogy has begun to be investigated in applied as well as experimental contexts. This research has particularly focused on the behaviour of experts. For example, Hargadon (1999) describes how 'knowledge brokers' in management draw comparisons between different areas in order to move ideas from where they are known to where they are not. Marchant, Robinson, Anderson and Schadewald (1993) investigated the use of analogies in the interpretation of tax statutes in graduate students and professional lawyers. They found that both groups demonstrated high rates of transfer from a structural analogue. Dunbar and Blanchette (2001; Dunbar, 2001) have documented the use of analogy by immunologists and molecular biologists, finding that structural analogising was particularly prevalent when the scientist was engaged in theory building. Dunbar and Blanchette (2001) were also able to determine the function of the analogies in their observations. When *isolated*

unexpected results occurred the scientists drew analogies to similar experiments, what Dunbar and Blanchette (2001) call local analogies. However, when a *series* of unexpected results occurred the scientists drew more distant analogies to the function of similar components in other organisms. The type of mapping appears to differ depending on the purpose for which it is drawn.

Thus, it appears that experts in applied settings are able to draw analogies between base and target problems. This is consistent with the widely held view that differences between experts and novices reflect different representational levels of information encoding. Experts are able to encode information at a deeper, structural level, while novices generally only encode information at a surface or superficial level (Chi, Feltovich & Glaser, 1981; Klein, 1999; Novick, 1988). Consistent with this explanation is a study by Thompson, Gentner and Loewenstein (2000), which found that unless management students were actively encouraged to compare source analogues in order to draw out structural relations, then rates of transfer of an underlying concept were low.

There is some evidence, however, that novices can also make use of structural analogies without being encouraged to create structural mappings. Blanchette and Dunbar (2000, 2001; Dunbar, 2001) found that novices were able to draw structural inferences when reasoning using metaphors. In a study where participants had to *explain* a concept to another person using a metaphor, it was found that the metaphor was frequently chosen from a domain outside the one being explained, suggesting that cross-structural mapping was occurring. One of the important aspects of Blanchette and Dunbar's study may be that participants were able to draw on any area of their knowledge in creating the analogies. This is in contrast to the previous studies involving novices that have examined the ability to produce a specific analogical mapping. Thus, it may be that novices are able to use analogy effectively if they are allowed to draw on memory more generally.

Metaphors are, however, slightly different from the types of analogies drawn to *solve* problems. Although metaphor relies upon the mechanism of analogy it is different from the kinds of analogy used in problem solving in two key respects. First, the relationship between source and target is different. In using analogy to solve problems the person must find a source that informs a less well understood target, whilst in drawing a metaphor a person has a target and must generate a source that explicates the underlying topic (Holyoak & Thagard, 1995). Second, metaphors often depend upon a combination with metonymy (where ideas substitute for each other) and the pattern of relations often shifts because of this in a way that problem solving analogies rarely do (Holyoak & Thagard, 1995).

The present study was, therefore, designed to extend existing research by addressing the question of whether novices can spontaneously make analogical mappings in order to solve management problems if they are allowed to draw more widely on stored knowledge. Also, in light of Blanchette and Dunbar's (2001) observations concerning the differing functions of analogies in real-world situations, we were alert to the possibility that management contexts might similarly be associated with analogy use aimed at achieving different functions.

Method

The study investigated undergraduate students in Management conducting an analysis of a 'business case' as part of a 'case method' course, a teaching method designed to simulate real-life management decision-making (Easton, 1992). Participants worked in groups and were required to specify the problems and opportunities inherent in the case, and to produce a set of solutions that might optimise the business described in the case. The task may be described as ill-defined (Van Lehn, 1989) and in some respects, un-defined. No restrictions were placed on the knowledge sources that participants might employ during the case analysis. No source analogues were presented to participants, and the concept of analogy was not mentioned to them as part of the course or the investigation.

Materials

The cases consisted of descriptions of a business or industry facing a loosely defined threat or opportunity. The business cases were chosen for their pedagogical merit by the assessing tutor and consisted of three different situations that described the position of a business or industry. The cases were 'The Champagne Industry in 1993' (Cool, Howe & Henderson, 1994), 'Petrol Retailing in Europe: The UK Market' (Levy, 1999), and 'Delta Dairies' (Easton & Dritsas, 1992). The cases were 19, 13 and 11 pages long, respectively, and are available from the European Case Clearing House Collection (http://www.ecch.cranfield.ac.uk.).

Participants

The participants were 24 final year undergraduate students from Lancaster University. The outcomes of each group's analysis was examined by a tutor as part of the student's course assessment. Participants were not paid.

Procedure

Six groups of four people were formed on the basis of who would and wouldn't work well with each other by the assessing tutor. Each group examined a single case with each case being analysed twice by two separate

groups. The groups spent two weeks conducting an analysis of their case. During this period the groups met both on their own and with the tutor. The groups met with the tutor between three and four times for approximately an hour. It is data from these sessions that were used for the present investigation. At the end of the analysis period the groups were required to make a 20 minute presentation to their peers, in the form of recommendations for the business.

Results

19 tutorials (out a possible 22) were observed and audio-taped. Instances of analogising in the tutorials were transcribed from the audio-tapes (off-task analogies were excluded). For the purposes of this study, analogy was considered to have occurred when reference was made to an episode of prior experience (a base) and was applied to a current idea in some way (a target). The extract was excluded when the base was drawn from lectures, the assessing tutor or directly from the case. The application of management knowledge or general knowledge was not considered to be an analogy in this study. A 15% sample of these tapes was re-coded by a second coder, and any discrepancies in analogy extraction were discussed until consensus was reached. Over the 19 tutorials there were 86 occasions when an analogy was made, a mean of 4.5 analogies per hour-long tutorial (standard deviation = 4.98, range = 0-20) and only one tutorial contained no discernible analogising.

Structure of the Analogies

The extracts of analogies were sorted into three categories based on a form of predicate calculus similar to that used by Gentner (1983). Extracts were defined in terms of whether the mapping between the base and the target involved merely superficial attributes, a first-order relation, or a complex systemic relation. An extract was classed as superficial if the mapping was in terms of objects only, with no discernible structural mapping. A simple, first-order relational mapping occurred when a simple relation that held in the source was also observed to hold in the target. A complex mapping of systems of relations occurred when higher-order relations were seen to be mapped, such as causal relations or plot structure (cf. Gentner, 1983).

It was found that 43% of the analogies generated that included an explicit mention of a target and base situation involved mappings of higher-order relations, such as can be seen in the following extract:

P1: "The marketing option, as well, which we'll use on the short-term; have you seen the BP advert?
CL: Yes.
P1: We're going down that line. We were saying so the other day. The amount at which it burns cleaner is negligible, but perceptions - so go for a cleaner image in the way BP are doing now. They're playing on the fact that that pollution is becoming more and more evident. Everybody is starting to understand it now, accept it, rather than just saying, 'it's rubbish that, we're not Greenpeace people'. It's becoming more of a factor in society against dirty polluting petrol. So, I mean, using the BP model, try to change people's perceptions about what your company at the moment, what the fuel is in [inaudible] it a greener alternative." [Group 4, 2nd tutorial]

The analogical mapping from base and target that is apparent in this extract can be restated using a propositional formalism, as follows:

Cause [Cause [Becoming-more-evident(Pollution), Becoming-greener(Society)], Use-as-marketing-option (BP, Green-standpoint)] ↔
Cause [Cause [Becoming-more-evident(Pollution), Becoming-greener(Society)], Use-as-marketing-option (Company-Y, Green-standpoint)]

The initial set of propositions capture the idea that the company in the base situation (British Petroleum) is taking a green standpoint because society is becoming greener in response to pollution becoming more evident. It is these systems of causal relations that this management group is mapping across to their own marketing option. It is noteworthy, too, that there are a number of higher-order relationships evident in the base situation, and the fact that these relations get mapped between base and target appears to demonstrate that novices can produce non-trivial analogies with sophisticated relational structures.

A further 34% of the analogies that involved the explicit mention of base and target situations involved first-order relational mappings. An example of such a mapping can be observed in the following extract:

P1: "That does happen though doesn't it, in supermarkets?
CL: What?
P1: That you could, it goes both - the exclusivity goes both ways but for huge brands, not the same size as ours. They state that a competitor can't be (sold in the supermarket)." [Group 2, 3rd tutorial]

This analogy can be represented propositionally as:

Has-exclusivity-deal-with(Huge-brand, Supermarket)
↔
Has-exclusivity-deal-with(Company-Y, Supermarket)

Here the group is thinking of copying the idea that they think is used by big companies of forcing a competitor off the shelves of a supermarket by signing an exclusivity deal.

Overall, then, 77% of the analogies produced by the novices observed in this study were based on structural mappings (either first-order relations or higher-order relations).

Function of the Analogies

The data were subjected to a thematic analysis in order to investigate further the type and function of the analogies employed by the management novices in this study. The emphasis in this analysis was on the *solution* or *idea* that emerged from the mapping, rather than on the nature of the mapping itself. In a thematic analysis extracts are grouped together based on their similarity, such that categories are developed based on common themes. Thematic analysis is a useful way of sorting qualitative data so that categories are allowed to emerge in a relatively atheoretical way (see Plummer, 1995; Smith, 1995). A second coder was able to recreate the themes identified with 88% accuracy (following a training session using 1/3 of the data). This analysis indicated that analogies were appearing to serve two different functions: problem solving and illustration.

Problem-Solving Analogies Of these analogies, 23% were direct base-to-target solutions, with both base and target present in the same extract. These took the form of 'x did y, so we can copy them'. For example:

"You could go on convenience a bit because there is an Esso station in Southampton, where I'm from, that is totally self-sufficient. It doesn't have anybody working there, and has, like, Coke dispensers, and all the kinds of food dispensers, and you pay at the pump and then you, you know, I've seen people pick up snacks from these machines and then they go. They're completely unmanned. That might be a possibility for convenience." [Group 3, 1st tutorial]

This type of analogy may be represented propositionally as:

Market-on-convenience(Esso-station,Fully-self-service)
↔
Market-on-convenience(Company-Y,Fully-self-service)

A further 15% of the analogies were elaborated base-to-target solutions. In these cases, a base was mapped to the target as in the direct base-to-target solutions, but the information gained from the analogy was used to form a new solution, rather than simply mapping the solution across wholesale from the base. For example:

"Another thing that we were having difficulty coming up with is an actual price, because we were thinking, 'Shall we out-price Moet et Chandon by only a small amount because it gives that exclusivity, and we didn't want to go for exactly the same price because we've got this unique selling point?' So if you just did it a tiny bit more expensive, going to that bit much it's as good as and it's got this unique selling point, and it's only a tiny bit more so that it's not too much of a stretch to buy it over Moet et Chandon. So people realise that it must be better, because it's that bit more expensive, and it's got this unique selling point." [Group 2, 2nd tutorial].

This analogy and its associated solution development may be represented in propositional form as:

Analogy: *Use-as-marketing-points (Moet-et-Chandon, Product-quality-and-product-exclusivity)* and *Indexed-by (Product-quality, High-price)* ↔
Use-as-marketing-points (Company-Y, Product-quality-and-product-exclusivity) and *Indexed-by (Product-quality, High-price)*
Solution idea: Cause *[More-expensive-than (Company-Y-one-press-champagne, Moet-et-Chandon-champagne), Gain-market-advantage-over (Company-Y, Moet-et-Chandon)]*

An additional 35% of the analogies were sources that shaped the group's decision making but which lacked a target that was explicitly referenced in the extract itself. An example of this comes from Group 2, 2nd tutorial:

"You know, like, how Safeway have got a grading system where, like, they've got bronze, silver and gold labelled wines, and things like that, you know, if they have something, I don't know what Casino have. But the supermarket's recommendation can be quite powerful. If you're looking for a wine in Safeway's and you don't particularly recognise the label, if you read the little Safeway bit on the back you know it's this level of sweetness, and it goes with this and that and the other. You're quite tempted to try some first time."

This extract shows that the group has taken the idea of the power of the supermarket's recommendation and this later guides the group's solution towards striking a deal with a French supermarket which leads to the recommendation of a dual-branding scheme, where the supermarket and the producer's name is on the bottle.

All of the analogies produced in order to serve a problem-solving purpose used either higher-order relational mappings or first-order relational mappings. There were no purely superficial mappings used to solve a problem.

Illustrative Analogies Twenty-seven percent of the analogies were designed not to facilitate directly the generation of a new solution idea, but instead for the purpose of exemplifying or illustrating an existing idea. Such analogies, therefore, appeared to be metaphorical in nature and intent rather than directed at problem solving *per se*. In such cases, the participant generated a source to explicate the target. For example, a member of Group 4 (2nd tutorial) drew parallels between the market positions of Coca Cola and Pepsi to illustrate the position faced by an oil company under consideration:

"You know, you said the other day that Coca Cola and Pepsi are within an arm's reach, - there's not much of a differentiation. It's the same here. It's just we've got more petrol stations and more people buying out of convenience."

The previous extracts concerned global target ideas of facilitating convenience at petrol stations, and the pricing for a new one-press champagne. In contrast, the illustration analogies lack these overarching ideas and extend no further than the base to target mapping. They are merely designed as a comparison of one idea with another.

In contrast to the analogies drawn to solve problems where there was no superficial mappings, when the analogy was drawn for illustrative purposes, 57% of the analogies were based on superficial mappings (with 26% based on higher-order mappings and 17% based on first-order relations).

When participants are solving problems it does not make any sense to map a superficially similar but structurally dissimilar source to a target, since this would not aid problem solving. In contrast, when an analogy is merely being used for illustrative purposes, it *is* possible simply to use a superficial mapping, since the purpose is merely to facilitate understanding rather than advance solution development.

Discussion

These results are important for two main reasons. First, by demonstrating spontaneous analogising by novice problem-solvers in a naturalistic domain, they corroborate the widely held view (e.g., Anderson, 2000; Holland, Holyoak, Nisbettt & Thagard, 1986) that analogising plays a fundamental role in human problem-solving. Second, the emphasis on structural rather than superficial mappings demonstrates a sophistication in the manipulation of domain knowledge that is not usually associated with novices. It may be that, in some domains at least, novices are quite capable of recognising and manipulating information at a conceptual rather than superficial level.

Like the experts observed by Dunbar and Blanchette (2000), it was observed in this study that novices also drew analogies to serve different purposes, and that the nature of the mapping differs depending on this purpose. When solving problems, the novices used only first-order and higher-order relational mappings. However, when they were illustrating an idea, over half of the analogies were made using superficial mappings.

The analysis also highlighted a potentially important distinction between solutions that are simply *mapped across* from a source analogue, and solutions that are developed as a result of additional idea generation subsequent to the attainment of an analogical mapping. It may be that much of the skill in management problem solving is to extend and elaborate upon initial analogical mappings.

There are some important qualifications to our results. Perhaps the key one is that the data we report here reflect group tutorial activity in which a fair proportion of the exchanges consist of the communication of ideas and outcomes among participants. On the basis of these data alone, we cannot ascertain whether the role played by analogy is one of genuine problem-solving, or whether it serves a mainly communicative role, making ideas and solutions that have been discovered and worked through using other problem-solving strategies easier to share between individuals. In a sense, it does not affect the outcomes reported here, since either role is crucial in collaborative problem-solving. However, future studies that are not tutorial-based are needed to determine the precise function of novice analogising in this domain.

This investigation differs from experimental studies of analogy in the following key ways. First, analogies could be drawn from any area of a participant's experience, and sources were not provided by the experimenter. Second, participants solved the problems in groups rather than individually. Third, the participants had two weeks of discussion-based learning to analyse and solve the presented problem. It is clear that the method of exploring analogy use employed here and the standard experimental method represent very different paradigms, such that direct comparisons between the two should be drawn with caution. However, considering that experiments are supposed to be analogues of real-world situations in a simplified form, a reconsideration as to how analogical problem solving can fruitfully be investigated experimentally may need to be undertaken in light of the mounting evidence that people frequently make cross-structural analogies in the real-world.

Acknowledgements

This research was funded by an ESRC grant to the first and third authors, with co-funding from the European Case Clearing House. We are grateful to Geoff Easton

for providing us with access to his students, and to Rosamund Ward who acted as the second coder for our coding-reliability checks.

References

Anderson, J. R. (2000). *Learning and memory: An integrated approach.* New York: Wiley.

Anoli, L., Antonietti, A., Crisafulli, L., & Cantoia, M. (2001). Accessing source information in analogical problem-solving. *Quarterly Journal of Experimental Psychology, 54A*, 237-261.

Blanchette, I., & Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. *Memory and Cognition*, *28,* 108-124.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory and Cognition, 29,* 730-735.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*, 1147-1156.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science,* 5, 121-152

Cool, K., Howe, J., & Henderson, J. (1994). *The champagne industry in 1993.* European Case Clearing House Collection.

Dunbar, K. (2001). The analogy paradox: Why analogy is so easy in naturalistic settings, yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.). *Analogy: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Dunbar, K., & Blanchette, I. (2001). The in vivo/in vitro approach to cognition: the case of analogy. *Trends in Cognitive Sciences*, *5,* 334-339.

Easton, G. (1992). Learning from case studies (2nd Edn.). New York: Prentice Hall.

Easton, G., & Dritsas, M. (1992). Delta Dairies. European Case Clearing House Collection.

Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Psychology, 7*, 155-170.

Gick, M. L., & Holyoak, K. J. (1980). Analogical Problem Solving. *Cognitive Psychology, 12,* 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Halford, G. (1992). Analogical reasoning and conceptual complexity in cognitive development. *Human Development, 35*, 193-217.

Hargadon, A. B. (1999). The theory and practice of knowledge brokering: Case studies of continuous innovation. *Dissertation Abstracts—International Section A: Humanities and Social Science, 59*, 3075.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction, Processes of inference, learning and discovery.* Cambridge, MA: MIT Press.

Klein, G. (1999). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Levy, A. (1999). *Petrol retailing in Europe: The UK market.* European Case Clearing House Collection.

Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1993). The use of analogy in legal argument: Problem similarity, precedent and expertise. *Organisational Behaviour and Human Decision Making Processes, 55*, 95-119.

Novick, L. R. (1988). Analogical transfer, problem similarity and expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 510-520

Pedone, R., Hummel, J. E., & Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. *Memory and Cognition, 29*, 214-221.

Poze, A. (1983). Analogical connections: The essence of creativity. *Journal of Creative Behaviour, 17*, 240-258.

Plummer, K. (1995). Life story research. In J. A. Smith, R. Harre, & L. V. Langenhoven, (Eds.), *Rethinking methods in psychology.* London: Sage

Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. London: Lewis.

Schank, R. C. (1999*). Dynamic memory revisited*. Cambridge, UK: Cambridge University Press.

Schunn, C. D., & Dunbar, K. (1996). Priming, analogy and awareness in complex reasoning. *Memory and Cognition, 24*, 271-284.

Smith, J. A. (1995). Semi-structured interviewing and qualitative analysis. In J. A. Smith, R. Harre, & L. V. Langenhoven, (Eds.), *Rethinking methods in psychology.* London: Sage

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84*, 353-378.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behaviour and Human Decision processes, 82,* 60-75.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of cognitive science.* Cambridge, MA: MIT Press.

Neonatal Learning of Faces: Environmental and Genetic Influences

James A. Bednar (jbednar@cs.utexas.edu)
Risto Miikkulainen (risto@cs.utexas.edu)
Department of Computer Sciences, The University of Texas at Austin
Austin, TX 78712 USA

Abstract

Newborn face perception is controversial, but the current evidence suggests that (a) newborns follow face-like schematic patterns further than similar patterns, (b) infants can learn individual faces soon after birth, and (c) full face processing abilities develop through months or years of experience with faces. Previous models have not adequately accounted for all three types of results. In prior work, we showed how a biologically based self-organizing system and spontaneous activity patterns can explain newborn face preferences. In this paper we show that this general-purpose learning system can explain both neonatal and later learning. Using computational simulations, we demonstrate that newborn learning need not be based on the external outline, as has been supposed, and that postnatal decreases in response to schematic faces need not represent a decrease in response to real faces. These simulations provide concrete predictions to guide future experiments with infants, while suggesting new techniques for designing complex adaptive systems in general.

Introduction

Specific regions in the adult visual cortex respond preferentially to human faces. How this face processing capability develops is not yet clear. Many researchers have argued that infants process only general visual properties like size and spatial frequency until after weeks or months of experience (Maurer & Barrera, 1981). Others have found a preference for faces at birth (Goren, Sarty, & Wu, 1975; Johnson, Dziurawiec, Ellis, & Morton, 1991; Simion, Valenza, & Umiltà, 1998) and that infants can learn and discriminate between specific faces even in the first few hours and days after birth (Bushnell, 2001; Pascalis, de Schonen, Morton, Deruelle, & Fabre-Grenet, 1995). Full face processing abilities clearly take several years to develop.

In this paper we show that a single learning system can account for all three types of results, i.e. face preferences at birth, face learning in the first few days after birth, and the gradual development of full face processing abilities. Using the HLISSOM self-organizing model (Hierarchical Laterally Interconnected Synergetically Self-Organizing Map), we have previously shown how prenatal learning of spontaneous neural activity can lead to newborn face preferences (Bednar & Miikkulainen, 2000). In this paper we show that the same self-organizing system can learn from faces in real images, and that the learning process can explain postnatal changes in infant face detection abilities. Together these simulations show how genetic information can be expressed within a highly adaptive system, and provide concrete predictions for future experiments with infants.

Development of face detection

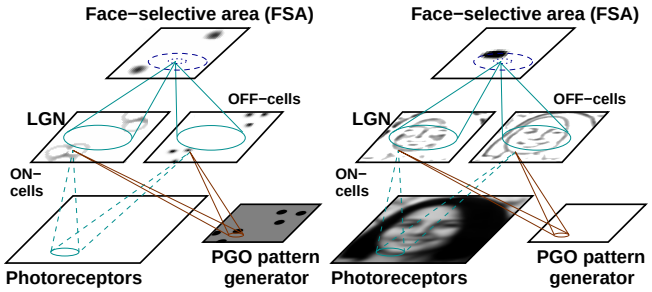
Face detection abilities change significantly between birth and two months of age. When shown moving schematic faces in the visual periphery, newborns and one month olds will follow them further than other similar patterns (Goren et al., 1975; Johnson et al., 1991; see example schematics in figure 5*a-d*). Older infants do not show a peripheral schematic face preference (Johnson et al., 1991) but between one and two months they begin to respond to facial features in central vision (Maurer & Barrera, 1981).

Previous models invoke separate visual processing mechanisms for these newborn and later face preferences. For instance, Johnson and Morton (1991) proposed that infants are born with a simple subcortical system they termed CONSPEC. CONSPEC serves only to detect and direct attention to face-like patterns in the periphery, perhaps using a simple three-dot template (two dots for the eyes and a third for the nose and mouth). A separate cortical system CONLERN would begin to control behavior after one month, and would gradually develop more sophisticated face processing through learning in central vision.

However, the CONSPEC/CONLERN model does not account for neonatal face learning. For instance, an infant only a few days old will prefer to look at its mother's face, relative to the face of a stranger (Pascalis et al., 1995). This mother preference has been thought to involve the external outline of the face only, in contrast to the internal facial feature learning of CONLERN, because the preference disappears when internal features are masked (Pascalis et al., 1995).

Accordingly, Johnson and Morton (1991) and subsequent authors have proposed extending the CONSPEC/CONLERN model to include face-outline learning in CONSPEC, or a third, separate subsystem for learning face outlines at birth (de Schonen, Mancini, & Leigeois, 1998; Simion et al., 1998). However, recent studies suggest that newborns can also learn internal features (Slater, Bremner, Johnson, Sherwood, Hayes, & Brown, 2000). Such learning could require a fourth subsystem, like CONLERN but for the periphery and operational at birth. (CONLERN itself cannot explain newborn learning of internal features, because were it present at birth, it would no longer explain the shift from peripheral to central face preferences after one month.)

We will show that such increasingly complex models are unnecessary. A single, CONLERN-like system processing the entire visual field is sufficient to explain the experimental data, if CONSPEC is replaced by a system that generates training patterns before birth. We have previously shown that a system trained on such spontaneous activity can account for the measured face preferences of newborns (Bednar, 2002; Bednar & Miikkulainen, 2000). The 3 hypotheses of the present paper are that: (1) networks trained on spontaneous activity learn more robustly after birth, compared to systems exposed only to environmental stimuli, (2) the decline in re-



(*a*) Prenatal learning (*b*) Postnatal learning

Figure 1: **HLISSOM model of face detection.** The model is a hierarchy of sheets of neural units, modeling the infant visual pathway. The model learns in two phases, each driven by input from a different source. During prenatal learning (*a*), activity originates internally in the PGO pattern generator. In postnatal learning (*b*), visual images are drawn on the photoreceptor sheet. For either type of input, the LGN is strongly activated by high-contrast edges and borders. The FSA is activated by the strongly responding units in the LGN, and activity then spreads laterally within the FSA. FSA weights are adapted when the activity settles. The FSA weights are initially uniform and unselective, but through self-organization they become selective for facelike patterns at the corresponding location on the retina. Their response is then information that an organism can use directly to control behaviors like visual fixation.

sponse to schematic patterns after one month results from learning of real faces and their outlines, not from a shift to a separate system, and (3) learning of both facial features and outlines can explain the development of a mother preference, including why it disappears when the outline is masked. Each of these hypotheses will be tested in a computational experiment with the HLISSOM model. Together the experiments will show that infant face learning and face preferences can be explained by a single, general-purpose learning system, which learns from both internally generated patterns of activity and from the visual environment.

HLISSOM Model

The architecture for the HLISSOM model is shown in figure 1, and will be briefly reviewed below. (For more details, see Bednar, 2002.) The model consists of a hierarchy of two-dimensional sheets of neural units modeling different areas of the nervous system: two sheets of input units (the retinal photoreceptors and the ponto-geniculo-occipital (PGO) pattern generator, described under *Prenatal learning* below), two sheets of LGN units (ON-center and OFF-center), and a sheet of cortical units ("neurons") representing a high-level area, the face-selective area (FSA)[1]. Each FSA neuron corresponds to a vertical column of cells through the six anatomical layers of the cortex.

[1]The FSA represents the first region in the ventral processing pathway that has receptive fields spanning approximately 45° of visual arc, i.e. large enough to span a human face at close range. Areas V4v and LO are likely FSA candidates based on adult patterns of connectivity, but the infant connectivity patterns are not known (Rolls, 1990). The generic term "face-selective area" is used rather than V4v or LO to emphasize that the model results do not depend on the region's precise location or architecture, only on the fact that the region has receptive fields large enough to allow face-selective responses. Cortical areas between the LGN and the FSA have been bypassed for simplicity; see Bednar (2002) for a more complex model including the primary visual cortex (V1).

The input to the model is an activity pattern on a sheet of photoreceptors or the PGO generator (see examples in figure 1). Each cell (i, j) in the ON- and OFF-center layers of the LGN computes its response η_{ij} as a scalar product of a fixed weight vector and its receptive fields on each input sheet:

$$\eta_{ij} = \sigma\left(\sum_{\rho ab} \gamma_\rho X_{\rho ab} w_{ij,\rho ab}\right), \quad (1)$$

where σ is a piecewise linear sigmoid activation function, ρ specifies the input sheet (either photoreceptors or PGO), γ_ρ is a constant scaling factor, $X_{\rho ab}$ is the activation of input unit (a, b) on sheet ρ, and $w_{ij,\rho ab}$ is the corresponding weight value. The lower bound δ of the sigmoid acts as an activation threshold; there is no response for activation below δ. Each FSA neuron computes its initial response like that of an LGN cell, except that ρ is either the ON or OFF LGN layer. After the initial response, the FSA activity evolves through short-range excitatory and long-range inhibitory lateral interaction:

$$\eta_{ij}(t) = \sigma\left(\sum_{\rho ab} \gamma_\rho X_{\rho ab}(t-1) w_{ij,\rho ab}\right), \quad (2)$$

where ρ specifies the weight type (either ON channel afferent, OFF channel afferent, lateral excitatory, or lateral inhibitory), γ_ρ is a constant scaling factor for each weight type (negative for inhibitory lateral weights), and $X_{\rho ab}(t-1)$ is the activation of target unit (a, b) during the previous time step. The FSA activity pattern starts out diffuse, but within a few iterations of equation 2, converges into a small number of stable focused patches of activity, or activity bubbles (as in figure 1). After the activity has settled, the connection weights of each FSA neuron are modified. All FSA weights adapt according to the Hebb rule, normalized so that the sum of the weights of each type ρ is constant for each neuron (i, j):

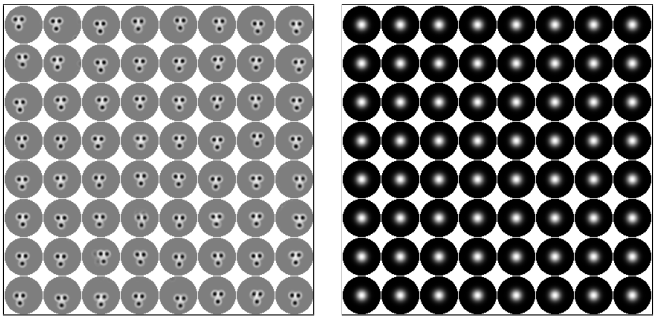
$$w_{ij,\rho ab}(t + \Delta t) = \frac{w_{ij,\rho ab}(t) + \alpha_\rho \eta_{ij} X_{\rho ab}}{\sum_{ab}\left[w_{ij,\rho ab}(t) + \alpha_\rho \eta_{ij} X_{\rho ab}\right]}, \quad (3)$$

where η_{ij} stands for the activity of neuron (i, j) in the final activity bubble, $w_{ij,\rho ab}$ is the connection weight, α is the learning rate for each type of connection, and $X_{\rho ab}$ is the presynaptic activity. The larger the product of the pre- and post-synaptic activity $\eta_{ij} X_{\rho ab}$, the larger the weight change.

For these experiments, a pair of 74×74 ON-center and OFF-center cell layers received input from a 170×170 photoreceptor sheet and an 85×85 PGO sheet. Each ON/OFF cell had a fixed Difference of Gaussians receptive field (RF) within the photoreceptor array (center $\sigma = 0.75$, surround $\sigma = 1.2$). The 24×24 FSA neurons each had a circular afferent receptive field of size 25, centered on the location in the central 24×24 portion of the ON/OFF cell layer corresponding the neuron's location in the FSA. This mapping ensures that every neuron has a complete set of circular afferent connections. Initially, the afferent and lateral weights in the FSA had a smooth circular Gaussian profile, and all weights of each type were identical. Other parameters were from Bednar and Miikkulainen (2000), scaled for this cortex size using the equations from Bednar (2002).

Prenatal learning

The simulations in this paper focus on postnatal learning, but they continue from our earlier results on prenatal learning (Bednar & Miikkulainen, 2000), which we summarize



(*a*) Prenatally trained network (*b*) Naïve network

Figure 2: **Starting points for postnatal learning.** These plots show the RFs for every third neuron from the 24×24 array of neurons in the FSA. For the prenatally trained network (*a*), the RFs were visualized by subtracting the OFF weights from the ON. The result is a plot of the retinal stimulus that would most excite that neuron. Like CONSPEC, the prenatally trained network consists of an array of roughly facelike RFs. In contrast, the neurons in the naïve network are initially uniformly Gaussian. The ON and OFF weights were identical, so only the ON weights are shown in (*b*). Later figures will compare the postnatal learning of each network.

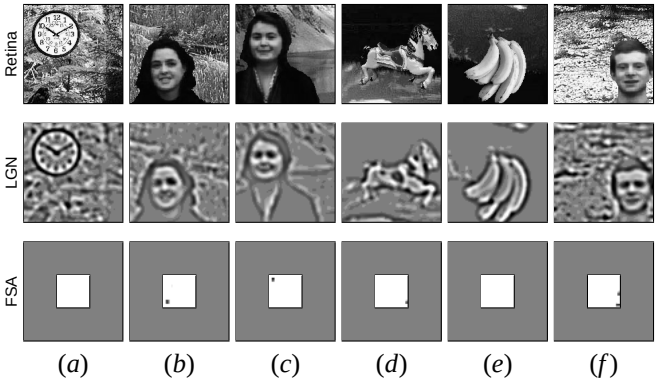


Figure 3: **Sample postnatal learning iterations.** The top row shows six randomly generated images drawn on the retinal photoreceptors at different iterations. Each image contains a foreground item chosen randomly from a set of three men and three women (adapted from Rowley et al., 1998) and six object images (from public domain clip art collections). Each foreground item was overlaid onto a random portion of an image from a database of 58 natural scenes (National Park Service, 1995), at a random location and at a nearly vertical orientation (drawn from a normal distribution around vertical, with $\sigma = \pi/36$). The second row shows the LGN response to each of these sample patterns, visualized by subtracting the OFF cell responses from the ON cell responses. Dark areas indicate high OFF cell response, light indicate high ON cell response, and medium gray indicates no response. The bottom row shows the prenatally trained FSA response to each pattern, at the start of postnatal training. For the FSA, only neurons with complete receptive fields (those in the unshaded inner box) were simulated, because those in the gray area would have RFs cut off by the edge of the retina. The gray area shows the FSA area that corresponds to the same portion of the visual field as in the LGN and retina plots, to facilitate comparison. The FSA responds to groups of dark spots on the retina, such as the eyes and mouths in (*b-c,f*) and the horse's markings in (*d*); the location of the FSA activity corresponds to the position of the group of retinal patterns that caused the response.

here. We hypothesize that before birth, training patterns arise from ponto-geniculo-occipital (PGO) waves generated during rapid-eye-movement (REM) sleep. Developing embryos spend a large percentage of their time in a precursor of REM sleep, which suggests that this state has a major role in development (Roffwarg, Muzio, & Dement, 1966). During and just before REM sleep, PGO waves originate in the brain stem and travel to the LGN, visual cortex, and many other brain areas (see Callaway, Lydic, Baghdoyan, & Hobson, 1987 for a review). PGO waves are strongly correlated with eye movements and with vivid visual imagery in dreams, suggesting that they activate the visual system as if they were visual inputs (Marks, Shaffery, Oksenberg, Speciale, & Roffwarg, 1995). PGO waves elicit different distributions of activity in different species, and interrupting them has been shown to increase the influence of the environment on development (Marks et al., 1995).

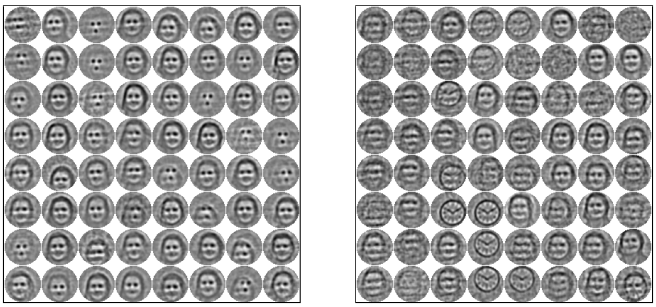
All of these characteristics suggest that PGO waves may be providing species-specific training patterns for development (see Bednar, 2002 for more details). However, due to limitations in experimental imaging equipment and techniques, the spatial shape of the PGO wave activity patterns has not yet been measured. Based on the CONSPEC model, we chose the three-dot patterns illustrated in the PGO area of figure 1*a*. Other patterns are also possible, and will provide greater or lesser face selectivity (Bednar, 2002).

As described in previous work (Bednar & Miikkulainen, 2000), FSA neurons exposed to prenatal patterns developed receptive fields (RFs) preferring upright, triangular arrangements of three dots (figure 2*a*). The resulting map responds to most frontal face images, and not to most objects or backgrounds. At this stage, the trained map can be considered an implementation of CONSPEC, except that it was constructed by learning and will continue to learn after birth.

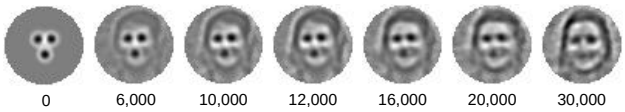
To determine whether the prenatal training biases subsequent learning (hypothesis 1 above), we also simulated a control condition called the *naïve* network. The naïve network is so called because it models neurons that have not had experience with coherent activity patterns until after birth. So that the naïve and prenatally organized networks would match on as many parameters as possible, we constructed the naïve network from the prenatally trained network *post hoc* by explicitly resetting afferent receptive fields to their uniform-Gaussian starting point (figure 2*b*). This procedure removed the prenatally developed face selectivity, but kept the lateral weights and all of the associated parameters the same. The activation threshold δ for the naïve FSA network was then adjusted so that for a given training pattern both networks would have similar activation levels; otherwise the parameters were the same for each network. This procedure ensures that the comparison between the two networks will be as fair as possible, because besides the thresholds the networks differ only by whether the neurons have face-selective weights at birth.

Postnatal testing and learning

The postnatal learning experiments reported in this paper simulate gradual learning of specific individuals and objects seen against a variety of different backgrounds. Figure 3 shows samples of the images we used and describes how they were generated. The prenatally trained and naïve networks were each exposed to the same pseudorandom sequence of 30,000 of these images.



(*a*) Prenatally trained network (*b*) Naïve network



(*c*) Snapshots of typical prenatally trained RF

(*d*) Snapshots of naïve network RF

Figure 4: **Prenatal patterns bias postnatal learning in the FSA.** Plots (*a*) and (*b*) show the final RFs for every third neuron from the 24×24 array of neurons in the FSA, visualized as in figure 2*a*. As the prenatally trained network learns from real images, the RFs morph smoothly into prototypes, i.e. representations of average facial features and hair outlines (*c*). By postnatal iteration 30,000, nearly all neurons have learned face-like RFs, with very little effect from the background patterns or non-face objects (*a*). Postnatal learning is less uniform for the naïve network, as can be seen in the RF snapshots in (*d*). In the end, many of the naïve neurons do learn face-like RFs, but others become selective for general texture patterns, and some become selective for objects like the clock (*b*). Overall, the prenatally trained network is biased towards learning faces, while the initially uniform network more faithfully represents the environment. Thus prenatal learning can allow the genome to guide development in a biologically relevant direction.

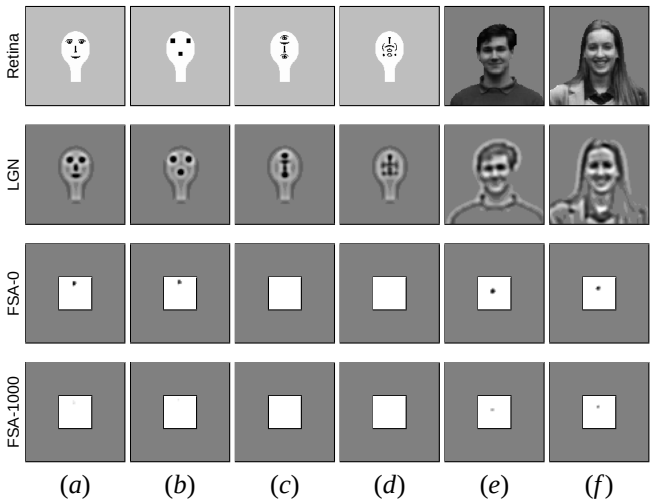


Figure 5: **Decline in response to schematic faces.** Before postnatal training, the prenatally trained FSA (third row from top) responds significantly more to the facelike stimulus (*a*) than to the three-dot stimulus (*b*; $p = 0.05$) or the scrambled faces (*c*-*d*; $p = 10^{-8}$). Assuming that infants attend most strongly to the stimuli that cause the greatest neural response, these responses replicate the schematic face preferences found by Johnson and Morton (1991) in infants up to one month of age. Some of the Johnson & Morton, 1991 experiments found no significant difference between (*a*) and (*b*), which is unsurprising given that they are only barely significantly different here. As the FSA neurons learn from real faces postnatally, they respond less and less to schematic faces. The bottom row shows the FSA response after 1000 postnatal iterations. The FSA now rarely responds to (*a*) and (*b*), and the average difference between them is no longer significant ($p = 0.25$). Thus no preference would be expected for the facelike schematic after postnatal learning, which is what Johnson and Morton (1991) found for older infants, i.e. 6 weeks to 5 months old. The response to real faces also decreases slightly through learning, but to a much lesser extent (*e*-*f*). The response to real faces declines because the newly learned average face and hair outline RFs are a weaker match to any particular face than were the original three dot RFs. That is, the external features vary more between individuals than do the internal features, and thus their average is not a close match to any particular face. Even so, there is only a comparatively small decrease in response to real faces, because real faces are still more similar to each other than to the schematic faces. Thus HLISSOM predicts that older infants will still show a face preference if tested with more-realistic stimuli, such as photographs.

At the beginning of the postnatal phase, and at intervals throughout, we tested the network using schematic images previously tested with newborns, and with photographs of faces. In order to compare the neural activity in the model to babies' attentional preferences, we assume that newborns pay more attention to the stimuli that are most effective at activating their visual processing system, focusing on the highest level activated. Patterns activating the FSA will be preferred over those activating only lower areas, and patterns that both activate the FSA will be ranked by their FSA activity. We quantify these comparisons by presenting each stimulus 25 times at different retinal locations, and averaging the sum of the FSA activity. As in the psychological studies we are modeling, differences between patterns will be tested with the two-tailed Student's t-test, treating p values below 0.05 as significant.

Results

Experiment 1: Bias from prenatal learning

Figure 4 shows that with postnatal exposure to real images, both the naïve and prenatally trained networks develop RFs that are averages (i.e. prototypes) of faces and hair outlines. RFs in the prenatally trained network smoothly increase in face selectivity, and eventually nearly all become highly selective for faces (figure 4*b*). Postnatal self-organization in the naïve network is less regular, and the final result is less face selective. Thus prenatal training biases postnatal learning towards biologically relevant stimuli, i.e. faces (hypothesis 1).

Experiment 2: Decline in response to schematics

Figure 5 shows that the HLISSOM model replicates the disappearance of peripheral schematic face preferences after one month (hypothesis 2; Johnson et al., 1991). In HLISSOM, the decrease results from the afferent weight normalization (equation 3). As the FSA neurons in HLISSOM learn the hair and face outlines typically associated with real faces, the connections to the internal features necessarily become weaker. Unlike real faces, the facelike schematic patterns match only on these internal features, not the outlines. As a result, the response to schematic facelike patterns decreases as real faces

are learned. Eventually, the response to the schematic patterns approaches and drops below the fixed activation threshold δ. At that point, the model response is no longer higher for schematic faces (because there is no FSA response, and V1 responses are similar). In a sense, the FSA has learned that real faces typically have both inner *and* outer features, and does not respond when either type of feature is absent or a poor match to real faces.

Yet the FSA neurons continue to respond to real faces (as opposed to schematics) throughout postnatal learning (figure 5*e-f*). Thus the model provides a clear prediction that the decline in peripheral face preferences is limited to schematics, and that if infants are tested with sufficiently realistic face stimuli, no decline in preferences will be found.

Experiment 3: Mother preferences

Figure 6*a-b* shows that when one face (i.e. the mother) appears most often, the FSA response to that face becomes stronger than to a similar stranger. This result replicates the mother preference found in infants a few days old (hypothesis 3; Bushnell, 2001; Pascalis et al., 1995). Interestingly, figure 6*c-d* shows that the mother preference disappears when the hair outline is masked, which is consistent with Pascalis et al.'s claim that newborns learn outlines only. However, Pascalis et al. (1995) did not test the crucial converse condition, i.e. whether newborns respond when the facial features are masked, leaving only the outlines. Figure 6(*e-f*) shows that there is no response to the head and hair outline alone either, and thus that this face learning is clearly *not* outline-only.

In the model, the decreased response with either type of masking results from holistic learning of *all* of the features typically present in real faces. As real faces are learned, the afferent weight normalization ensures that neurons respond only to patterns that are a good overall match to all of the weights, not simply matching on a few features. Many authors have argued that adults also learn faces holistically (e.g. Farah et al., 1998). These results suggest that newborns may learn faces in the same way, and predict that newborns will no prefer their mother when her hair outline is visible but her facial features are masked.

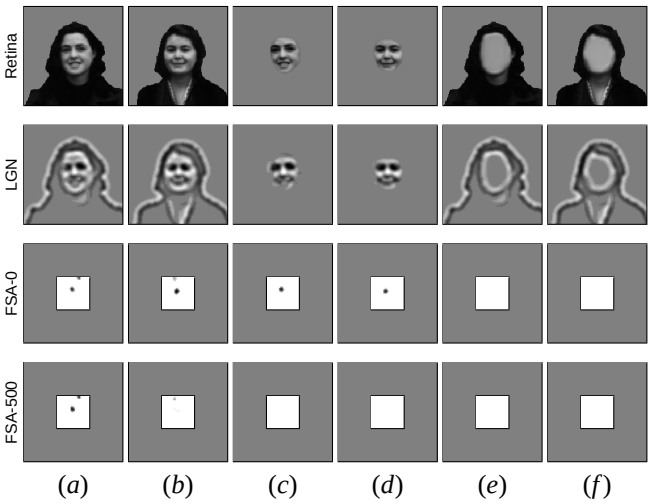


Figure 6: **Mother preferences depend on both internal and external features.** Initially, the prenatally trained FSA responds well to both women above (*a-b*; FSA-0), with no significant difference ($p = 0.28$). The response is primarily due to the internal facial features (*c-d*; FSA-0), although there are some spurious three-dot responses due to alignment of the hair with the eyes (*a-b*; top of FSA-0). Designating image (*a*) as the mother, we presented it in 25% of the postnatal learning iterations. (This ratio is taken from Bushnell, 2001, who found that newborns look at their mother's face for an average of about one-fourth their time awake over the first few days.) Image (*b*), the stranger, was not presented at all during training. After 500 postnatal iterations, the response to the mother is significantly greater than to face (*b*) ($p = 0.001$). This result replicates the mother preference found by Pascalis et al. (1995) in infants 3–9 days old. The same results are found in the counterbalancing condition — when trained on face (*b*) as the mother, (*b*) becomes preferred ($p = 0.002$; not shown). After training with real faces, there is no longer any FSA response to the facial features alone (*c-d*), which replicates Pascalis et al.'s (1995) finding that newborns no longer preferred their mother when her face outline was covered. Yet contra Pascalis et al. (1995), we cannot conclude that what has been learned "has to do with the outer rather than the inner features of the face", because no preference is found for the face outline alone either (*e-f*). Thus face learning in HLISSOM is holistic. Face learning in adults is also thought to be holistic (Farah et al., 1998), and these results show that we do not need to assume that newborns are using a different type of face learning than adults.

Discussion and future work

The HLISSOM simulations show that internally generated patterns and a self-organizing system can together account for newborn face preferences, neonatal face learning, and longer term development of face detection. The results suggest simple but novel explanations for why newborn learning appears to depend on the face outline, and why the response to schematic faces decreases over time. These explanations lead to concrete predictions for future infant experiments. Over the first two months the response to real faces in the periphery should continue even as response to schematics diminishes, and the mother preference of newborns should disappear when the facial features are masked. The results also show that internally generated patterns allow the genome to steer development towards biologically relevant processing, making learning of more sophisticated abilities quicker and more robust.

The results above do not address one interesting phenomenon: in central vision, preference for schematic faces is not measurable until 2 months of age (Maurer & Barrera, 1981), and is gone by 5 months (Johnson et al., 1991). This time course is delayed relative to peripheral vision, where preferences are present at birth but disappear by 2 months.

Johnson and Morton (1991) originally proposed that in the periphery the preferences disappear because CONLERN matures and inhibits CONSPEC, while in central vision they disappear because CONLERN learns properties of real faces. HLISSOM provides a unified explanation for both phenomena: a single learning system stops responding to schematic faces because it has learned from real faces.

Why, then, would the time course differ between peripheral and central vision? As Johnson and Morton acknowledged, the retina changes significantly over the first few months. In particular, at birth the fovea is much less mature than the periphery, and may not even be functional yet (Abramov, Gordon, Hendrickson, Hainline, Dobson, & LaBossiere, 1982; Kiorpes & Kiper, 1996). Thus schematic face preferences

in central vision may be delayed relative to those in peripheral vision simply because the fovea matures later. A single cortical learning system like HLISSOM is thus sufficient to account for the time course of both central and peripheral schematic face preferences.

The development of the fovea may also affect mother preferences. Consistent with our results, Bartrip, Morton, and de Schonen (2001) found that infants 19–25 days old do not significantly prefer their mothers when either her internal features or external features are covered. Interestingly, Bartrip et al. found that older infants, 35–40 days old, do prefer their mothers even when the external features are covered. The gradual maturation of the fovea may again explain these later-developing capabilities. Unlike the periphery, the fovea contains many ganglia with small RFs, and which connect to cortical cells with small RFs. These neurons can learn smaller regions of the mother's face, and their responses will allow the infant to recognize the mother even when other regions of the face are covered. Thus simple, documented changes in the retina can explain why mother preferences would differ over time.

In general, the idea that artificially generated training patterns can influence the development of learning systems is powerful, and could be used to construct artificial systems as well. Simple, engineered training patterns can provide an initial or ongoing bias, while learning algorithms incorporate the full complexity of the environment. This approach can allow more complex adaptive systems to be designed and implemented.

Conclusion

A single learning system can explain the seemingly complex postnatal time course of face processing, if that system is exposed to internally generated patterns. Initial face selectivity develops from these non-visual inputs, and postnatal experience interacts with these genetic factors to develop full face processing abilities. These results provide clear predictions for future infant experiments, and provide new tools for constructing complex artificial systems.

Acknowledgments

This research was supported in part by the National Science Foundation under grants IRI-9309273 and IIS-9811478.

References

Abramov, I., Gordon, J., Hendrickson, A., Hainline, L., Dobson, V., & LaBossiere, E. (1982). The retina of the newborn human infant. *Science, 217* (4556), 265–267.

Bartrip, J., Morton, J., & de Schonen, S. (2001). Responses to mother's face in 3-week to 5-month-old infants. *British Journal of Developmental Psychology, 19*, 219–232.

Bednar, J. A. (2002). *Learning to See: Genetic and Environmental Influences on Visual Development*. Doctoral Dissertation, Department of Computer Sciences, The University of Texas at Austin, Austin, TX.

Bednar, J. A., & Miikkulainen, R. (2000). Self-organization of innate face preferences: Could genetics be expressed through learning?. In *Proceedings of the 17th National Conference on Artificial Intelligence* (pp. 117–122). Cambridge, MA: MIT Press.

Bushnell, I. W. R. (2001). Mother's face recognition in newborn infants: Learning and memory. *Infant and Child Development, 10* (1/2), 67–74.

Callaway, C. W., Lydic, R., Baghdoyan, H. A., & Hobson, J. A. (1987). Pontogeniculooccipital waves: Spontaneous visual system activity during rapid eye movement sleep. *Cellular and Molecular Neurobiology, 7* (2), 105–49.

de Schonen, S., Mancini, J., & Leigeois, F. (1998). About functional cortical specialization: The development of face recognition. In (Simion & Butterworth, 1998), pp. 103–120.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105* (3), 482–498.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56* (4), 544–549.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*, 1–19.

Johnson, M. H., & Morton, J. (1991). *Biology and Cognitive Development: The Case of Face Recognition*. Oxford, UK; New York: Blackwell.

Kiorpes, L., & Kiper, D. C. (1996). Development of contrast sensitivity across the visual field in macaque monkeys (*Macaca nemestrina*). *Vision Research, 36* (2), 239–247.

Marks, G. A., Shaffery, J. P., Oksenberg, A., Speciale, S. G., & Roffwarg, H. P. (1995). A functional role for REM sleep in brain maturation. *Behavioural Brain Research, 69*, 1–11.

Maurer, D., & Barrera, M. (1981). Infants' perception of natural and distorted arrangements of a schematic face. *Child Development, 52* (1), 196–202.

National Park Service (1995). Image database. http://www.freestockphotos.com/NPS.

Pascalis, O., de Schonen, S., Morton, J., Deruelle, C., & Fabre-Grenet, M. (1995). Mother's face recognition by neonates: A replication and an extension. *Infant Behavior and Development, 18*, 79–85.

Roffwarg, H. P., Muzio, J. N., & Dement, W. C. (1966). Ontogenetic development of the human sleep-dream cycle. *Science, 152*, 604–619.

Rolls, E. T. (1990). The representation of information in the temporal lobe visual cortical areas of macaques. In Eckmiller, R. (Ed.), *Advanced Neural Computers* (pp. 69–78). New York: Elsevier.

Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20* (1), 23–38.

Simion, F., & Butterworth, G. (Eds.) (1998). *The Development of Sensory, Motor and Cognitive Capacities in Early Infancy: From Perception to Cognition*. East Sussex, UK: Psychology Press.

Simion, F., Valenza, E., & Umiltà, C. (1998). Mechanisms underlying face preference at birth. In (Simion & Butterworth, 1998), pp. 87–102.

Slater, A., Bremner, G., Johnson, S. P., Sherwood, P., Hayes, R., & Brown, E. (2000). Newborn infants' preference for attractive faces: The role of internal and external facial features. *Infancy, 1* (2), 265–274.

Conditional Promises and Threats – Cognition and Emotion

Sieghard Beller (beller@psychologie.uni-freiburg.de)
Department of Psychology, University of Freiburg
D-79085 Freiburg, Germany

Abstract

Conditional promises and threats are speech acts that can be used to manipulate the behavior of other persons. Although reasoning studies have been able to reveal some peculiarities of these concepts, the *explanation* has remained fragmentary. To fill in this gap, a theoretical analysis of conditional inducements is proposed, which integrates cognitive as well as emotional aspects. An experiment – focussing on linguistic, pragmatic, emotional, and deontic consequences – corroborates the analysis and shows that persons have a clear understanding of conditional inducements.

Introduction

Conditional promises and threats are speech acts (Searle, 1971) uttered by a person to manipulate an addressee's behavior by setting up consequences of his behavior (cf. Conison, 1997; von Wright, 1962). Walking home from school, Henry may make the following proposal to his classmate Bob:

(1) "If you lend me your bike,
then I will help you with your homework".

Bob can infer from this statement that Henry would like to borrow his bike, and that Henry believes that he (Bob) needs Henry's help.

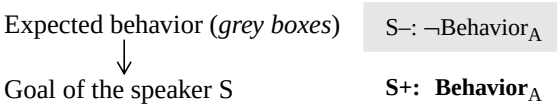
Research on human reasoning usually focusses on the *inferential* use of conditional promises/threats within the normative framework of propositional logic: Which inferences do people draw from conditional arguments? Suppose, (2) Bob *lends* Henry his bike. What follows then from (1) and (2)? Will Henry help Bob? – Although we cannot know for sure which action a person will actually take, we would expect that Henry *will* help him, or at least we think that he *will have to*. This answer is logically warranted by *Modus Ponens* (MP). But what should be concluded from the promise (1), if (3) Bob *does not* lend out his bike? Usually, persons answer that Henry *will not* help Bob either. This inference, however, corresponds to a logically invalid pattern known as *Denial of the Antecedent* (DA).

In general, the valid MP and MT (*Modus Tollens*) as well as the invalid DA and AC (*Affirmation of the Consequent*) are drawn more frequently from conditional promises and threats than from universal conditionals (e.g., Fillenbaum, 1978; Markovits & Lesage, 1990; Newstead, Ellis, Evans, & Dennis, 1997; but see Evans & Twyman-Musgrove, 1998, for a diverging result). Most people accept the complementary conditional '*If not*-P, *then not*-Q' as following from inducements of the form '*If* P, *then* Q' (Fillenbaum, 1978). Geis and Zwicky (1971) speak of "invited inferences" in normal linguistic usage. Accordingly, the associated truth tables often reflect an equivalence relation instead of an implication (Newstead et al., 1997).

Treating conditional inducements in this way enables one to *detect* effects of different propositional contents on reasoning, but it is not sufficient to *explain* the underlying causes. What is the reason for the DA inference in the introductory example? Furthermore, a simple truth-table analysis is much too restricted to capture conceptual aspects of conditional inducements beyond the if-then relation. Why, for example, is Henry *obliged* to help Bob under certain circumstances? How will Bob react *emotionally* if Henry does not help him? A detailed theoretical analysis is presented to overcome these limitations. It integrates several aspects on different levels: goals and incentives on the motivational level, formulations and inferences on the linguistic level, obligation and permission on the deontic level, action sequences on the pragmatic level, and finally, affective reactions on the emotional level. The multi-level analysis explains the phenomena observed in reasoning studies; new phenomena are predicted and experimentally confirmed.

Levels of Conditional Inducements

(1) Motivational level: The basic level of analysis concerns the motivational situation in which a person utters an inducement. It is determined by *expectations*, *goals* and *incentives*. The speaker (S) wants an addressee (A) to show a certain *goal*-behavior (i.e., to perform a certain action or to refrain from performing an action) with a positive value for himself, the speaker (S+: Behavior_A). In the introductory example, it was Henry who wanted Bob to lend him his bike. Henry must expect that the addressee is not willing to show this behavior voluntarily, otherwise an inducement would not be necessary. Thus, the speaker has to induce a behavioral change:



This change can be motivated in two ways: First, the speaker may promise to *reward* the *desired* goal behavior S+ with a positive consequence for the addressee

(➥ A+: Reward_S). Believing that Bob needs help with his homework, Henry may promise to help him (A+) if Bob lends him his bike (S+). The reward should be under the speaker's control and should *not* occur for any other reason, as otherwise it cannot develop its motivational effect (e.g., Evans & Twyman-Musgrove, 1998). The whole motivational schema may be represented as:

Promise

S–: ¬Behavior_A	A–: ¬Reward_S
S+: Behavior$_A$	➥ **A+: Reward**$_S$

Instead of rewarding the desired behavior, the speaker may *punish* the *undesired* behavior he fears (S–) with a negative consequence (A–). If he usually helps Bob with his homework, Henry can use Bob's expectation and threaten to withdraw his help (A–) if Bob does not lend him his bike (S–). The corresponding schema is:

Threat

S–: ¬Behavior_A	➥ **A–: Punishment**$_S$
S+: Behavior$_A$	A+: ¬Punishment_S

In both cases, the speaker announces (explicitly or implicitly) that he will react positively (A+) if the addressee shows the desired behavior (S+), and negatively otherwise. There is an essential difference, however: If the addressee cooperates (S+), then in the first case he *gets* something he cannot expect without the promise (the reward A+), whereas in the second case he only *avoids* the punishment (A–) without getting anything positive in return.

(2) Linguistic level: The motivational schemas directly determine which formulations are appropriate to express the intended speech act. Conditionals '*If* P, *then* Q' can be used equally well with both schemas. Conditionals point out a necessary consequence 'Q' of an antecedent condition 'P', and that is exactly what the speaker intends on the motivational level: to establish a new, definite consequence for one of the addressee's behavioral options. The canonical formulations are:

"If you do P [S+], then I will reward you with Q [A+]" vs.
"If you do P [S–], then I will punish you by Q [A–]".

Looking at the underlying motivational schemas, it becomes clear why the complementary form '*If not*-P, *then not*-Q' is inferred, and why conditional inferences (MP, MT, DA, AC) seemingly correspond to an equivalence relation. The motivational level suggests that there are two action sequences: a cooperative one and a not-cooperative one. The complementary conditional reflects that part of the motivational background that is not expressed explicitly by the canonical statement. Together, the canonical and the complementary conditional yield the equivalence interpretation. Different from a logical equivalence, however, the reversed form (e.g., "If I reward you with Q [A+], then you will do P [S+]") is not really equivalent to the canonical one. By reversing antecedent and consequent, the temporal order changes as well, so that the speaker can no longer guarantee that action 'P' is a necessary consequence of the antecedent event 'Q' ('P' is not under his control).

The differences between the motivational schemas also explain why only threats are formulated disjunctively, whereas both promises and threats can be formulated conjunctively (Fillenbaum, 1978). The *conjunctive* formulation expresses the connection between the new consequence set by the speaker and the addressee's behavior:

"Do P [S+] *and* I will reward you with Q [A+]" vs.
"Do P [S–] *and* I will punish you by Q [A–]".

A *disjunction* points out alternatives. In the case of a threat, it enables the speaker to express both his goal S+ and the punishment A–, which are part of alternative action sequences: "*Refrain from* doing P [S+] *or* I will punish you by Q [A–]". If a promise were to be reformulated disjunctively, then either the speaker's goal or his reward could no longer be expressed.

(3) The deontic level deals with the question of which action a person *may* or *must* perform with respect to a social rule (e.g., Beller, 2001). Conditional promises and threats establish such a rule and determine which actions persons are obliged to perform. Since the addressee can freely decide whether or not he cooperates, there is *no* deontic constraint on his behavior. He *may* cooperate, but he *need not*. The speaker's situation is different. Consider the promise

"If you do P [S+], then I will reward you with Q [A+]".

Once the addressee cooperates and fulfills the speaker's goal 'P', the promisor is *obliged* to cooperate and to give reward 'Q'. The promisor himself declared 'Q' to be a necessary consequence of condition 'P', so he *must* guarantee the reward. If the addressee does not cooperate, then there is *no* deontic constraint; the speaker *need not* give the reward, but he is *permitted* to do it voluntarily. Which obligation, however, results from a threat

"If you do P [S–], then I will punish you by Q [A–]"?

Two lines of argumentation are possible here: First, arguing analogously to the promise, the speaker is *obliged to punish* the addressee ('Q') if A does *not* cooperate ('P'). The speaker declared punishment 'Q' to be a necessary consequence of condition 'P', so he *must* react consequently (and indeed perhaps he should, in order to keep his credibility). What is the case if the addressee *cooperates*? By analogy, there is *no* constraint on the speaker's action, so he *need not* punish the addressee, but he actually *may* punish him. An implicit social rule, however, intuitively contradicts this interpretation: A person *must not* be punished without reason, whereas one *may* well give a reward without reason.

Second, it can be argued that the threat implies a *complementary promise* that determines the deontic interpretation: "If you *refrain from* doing 'P' [S+], then I will *not* punish you by Q [A+]". Associated with a promise is an obligation for the speaker to cooperate (A+) once

the addressee has fulfilled his goal (S+). If the addressee refrains from doing 'P' (S+), then the speaker *must* refrain from the punishment. If the addressee does not cooperate, then there is no deontic constraint and the speaker is *permitted* to punish the addressee. Since the punishment is now justified, this interpretation is in line with the implicit social rule mentioned above.

(4) The pragmatic level deals with the question of which actions are actually taken after an inducement has been uttered. Since both persons are assumed to have full freedom of action, four action sequences are possible: If the addressee fulfills the speaker's goal (S+), then subsequently the speaker may also cooperate (A+) or may not (A–). If the addressee does *not* show the goal behavior (S–), then the speaker may not cooperate either (A–) or may cooperate (A+). Common to all four sequences is a particular *temporal order*: The addressee decides first whether he wants to cooperate, whereas the speaker has to react to the addressee's behavior.

(5) Emotional level: Eventually, one of the four action sequences follows a conditional inducement. Each sequence is characterized by goals, expectations and incentives. These factors are directly relevant for the elicitation of emotions (e.g., Lazarus, 1991; Roseman, Antoniou & Jose, 1996). Goal-relevance is a necessary requirement for emotional reactions in general; goal-congruent events elicit positive emotions, while goal-incongruent events elicit negative emotions. Applied to conditional promises and threats, addressee and speaker should feel a positive emotion if the partner cooperates (and fulfills their goal or expectation), while a negative emotion should result if the partner does not cooperate. Which specific emotion arises in a given situation depends on further appraisal dimensions that cannot be described in detail here (for extensive analyses see, e.g., Roseman et al., 1996). Summarizing those studies though, *joy* can be expected when a person gets something positive, *relief* when an expected negative event does not occur, and *anger* when the partner does not cooperate even though he or she is obliged to.

Experiment

The proposed multi-level analysis integrates motivation, linguistics, pragmatics, deontic considerations, and emotions, thereby overcoming the limitations of a purely truth functional analysis of conditional promises and threats. In order to test hypotheses regarding particular facets of the analysis, an experiment was conducted which consisted of two parts.

The starting point of part I was an influential finding of Leda Cosmides (1989). Cosmides showed with domain-specific versions of Wason's (1966) selection task that persons are sensitive to which of the partners involved in a reciprocal exchange is accused of breaking his promise, but *not* to the conditional formulation. It did not make a difference whether the promise was formulated canonically or reversed. The multi-level analysis makes just the opposite prediction; namely that persons should be sensitive to the formulation of inducements. A reversed inducement is *not* equivalent to the original one, since it also implies a reversal of the temporal order and of the roles (*speaker-addressee-asymmetry*): Given a particular role allocation, the canonical conditional should be preferred to the reversed one, the complementary conditional (and not the reversed one) should be preferred as implication, and the action sequence should be "addressee first". In addition, it was assessed which *emotional reactions* persons attribute to the addressee if the speaker keeps or breaks "the rule".

Part II of the experiment focusses on an aspect that has not been explored until now: the *deontic inferences* people draw from conditional promises and threats. The prediction for promises is clear: If the addressee has fulfilled the speaker's goal, then an obligation arises for the speaker to give the promised reward; otherwise there is no such obligation. The deontic interpretation of conditional threats, however, is not equally clear. Do people infer (from the conditional form) an obligation for the speaker to punish the addressee A if A does not cooperate, or do they rather infer (from the *complementary* promise) an obligation to cooperate and refrain from the punishment if the addressee cooperates? It was expected that the second interpretation would predominate since it is not in conflict with general moral rules.

Method

Both parts of the experiment were integrated into one questionnaire. They used different basic scenarios from which four context stories each were constructed. The stories in part I dealt with the exchange situation mentioned in the introductory example (help with homework in exchange for borrowing a bike) but varied with regard to role allocation and speech act. Four tasks had to be solved, which focussed on the linguistic, pragmatic and emotional level. The context stories in part II were constructed from two different scenarios; one dealt with mutual lending of things and the other with mutual destruction of toys. Again, the speech acts varied, but this time only one role allocation was used. Four tasks asked for deontic inferences from the inducements. To facilitate the discussion of the results, stories and tasks for both parts are described later in separate sections.

Participants: 40 students from two introductory cognitive psychology courses (at the University of Freiburg) participated in the experiment. 18 students were male and 22 female, with a mean age of $M = 23.8$ years (*range*: 20-39 years).

Design: Participants were randomly assigned to one of four groups ($n = 10$). The four context stories of each part varied between groups. The speech acts were balanced within groups: If part I was about a promise then part II dealt with a threat (and vice versa).

Procedure: The questionnaire was administered at the beginning of the first course session. After a general instruction on the first page, the questionnaire began with the tasks of part I, followed by the tasks of part II.

The tasks were ordered as described below and each was written on a new page. Participants were instructed to work on the tasks in the given order, and to take as much time as needed. All materials were presented in German.

Part I: Assessing the Speaker-Addressee-Asymmetry and Emotional Reactions

In order to assess the speaker-addressee-asymmetry, four context stories were designed, which were similar to the introductory exchange scenario. The stories described the person's goals and their usual behavior; they varied with regard to the intended speech act (promise vs. threat) and the roles (speaker vs. addressee). In two stories, Henry wants to borrow Bob's bike. He tries to achieve this goal either by a promise or by a complementary threat (canonical conditionals: "If you lend me your bike, then I will help you with your homework" vs. "If you do *not* lend me your bike, then I will *not* help you with your homework"). In the other two stories, the roles were interchanged: This time, Bob wants Henry to help him with his homework and he tries to achieve this goal by the reversed promise or threat ("If you help me with my homework, then I will lend you my bike" vs. "If you do *not* help me with my homework, then I will *not* lend you my bike"). Each story was followed by four tasks.

(1) Formulation task: In the first task, the motivational background was given together with the type of speech act to be used. The participants were then instructed to choose from four given conditionals the one that was most appropriate for the speaker's intended inducement. The conditionals were derived from the canonical one by reversing and negating 'P' and 'Q'. Table 1 shows the results. As predicted, the canonical conditionals were clearly preferred (90% aggregated over all context stories). If the speech act changed then the complementary conditional was chosen, whereas if the role allocation changed then the reversed form was preferred.

(2) Inference task: From this task onwards, the context stories had been supplemented by the canonical conditional. The instruction called for participants to choose an adequate implication of the given conditional from the three others known from the formulation task. The results are presented in Table 2. As predicted, participants preferred the complementary conditional, which leaves the order of the actions the same, over the reversed one (85.0% vs. 12.5%, aggregated over all tasks, $\chi^2(1, n = 39) = 21.6$; $p < 0.001$).

Table 1: Frequency of choosing each conditional as the speaker's adequate promise or threat ($n = 10$ in each condition; canonical conditionals are **bold**-faced).

	Henry's		Bob's	
Conditional	Promise	Threat	Promise	Threat
If bike *then* help	**10**	1	-	-
If no bike *then* no help	-	**9**	-	-
If help *then* bike	-	-	**10**	3
If no help *then* no bike	-	-	-	**7**

Table 2: Frequency of choosing the most adequate implication of a given promise or threat ($n = 10$ in each condition; complementary inducements are **bold**-faced).

	Henry's		Bob's	
Conditional	Promise	Threat	Promise	Threat
If bike *then* help	given	**9**	-	-
If no bike *then* no help	**8**	given	1	2
If help *then* bike	2	-	given	**8**
If no help *then* no bike	-	1	**9**	given

(3) Sequence task: In the third task, the participants had to decide on the order of the actions once the conditional inducement had been made. It was expected that the addressee would decide first whether he is willing to cooperate. Without exception, all participants (100%) answered the sequence question according to this prediction. If Henry made the inducement then Bob decides first whether he lends out his bike, and vice versa. Thus, changing roles reversed the typical action sequence.

Altogether, the results of the first three tasks corroborate the predicted speaker-addressee asymmetry.

(4) Emotion task: After the introduction of the conditional inducement in the context story, the emotion task mentioned that the addressee *cooperated* and fulfilled the speaker's goal (S+). Participants had then to decide (i) what the *speaker* has to do in order to keep versus not to keep 'the rule', and (ii) which feeling the *addressee* will have afterwards. Three critical emotions (relief, joy, and anger) were given together with four distractors in a multiple-choice format and participants were instructed to choose the most appropriate one.

(i) To keep the rule means that, given that the addressee cooperated before (S+), the speaker will also cooperate (A+). Cooperation corresponds to the MP inference in the case of a conditional promise '*If* P [S+], *then* Q [A+]', but to the NA inference in the case of a threat '*If* P [S–], *then* Q [A–]'. Not keeping the rule means reacting defectively towards the addressee (A–) *even though* the addressee fulfilled the speaker's goal (S+). In the case of a promise, defection violates the conditional statement itself ('P *and not*-Q'). In the case of a conditional threat, however, defection corresponds to '*not*-P *and* Q' and violates the complementary conditional. The results show that the participants had a clear understanding of these regularities: Asked what the speaker has to do in order to "keep the rule", all persons (100.0%) choose the MP-option given a promise, but the NA-option (95.0%) given a threat (aggregated over both

Table 3: Attributed emotional reactions of the addressee on keeping vs. not keeping a promise/threat.

		Keeping		Not Keeping	
Emotion		Promise	Threat	Promise	Threat
Relief	(+)	5	10	-	-
Joy	(+)	15	5	-	1
Anger	(–)	-	3	19	19
Others	(–)	-	3	3	1

The frequencies do not add up to 20 in each column because four persons marked two emotions.

role versions). Correspondingly, all persons (100.0%) answered that not keeping a promise corresponds to 'P *and not*-Q', while again 95.0% chose the complementary category '*not*-P *and* Q' in the case of a threat.

(ii) How does the addressee react emotionally in these cases? Table 3 lists positive and negative emotions (+/–) aggregated over both role versions. The answers are in line with the predictions from appraisal theories: If the speaker keeps the rule then the addressee is said to feel a *positive* emotion (85.4% positive vs. 14.6% negative); otherwise a *negative* emotion results (2.3% positive vs. 97.7% negative; $\chi^2(1, n = 84) = 59.1$; $p < 0.001$). In the latter case, the addressee was uniformly said to feel *angry*, whereas in the former case different emotions were associated with the two speech acts: Keeping the promise mostly resulted in *joy* (75% joy vs. 25% relief) while in the case of a threat *relief* predominates (23.8% joy vs. 47.6% relief; $\chi^2(1, n = 35) = 6.08$; $p = 0.014$); on three occasions the addressee was even said to feel *angry*. This may be the result of having been forced to cooperate by a threat. Whether these differences reflect differences between the speech acts or between the incentives (lending out one's bike vs. giving help with the other's homework) is open to further analyses.

Part II: Assessing Deontic Inferences

Part II aimed to test the hypothesis that the deontic interpretation of a conditional threat follows the interpretation of the corresponding complementary promise: If the addressee cooperates, then the speaker is obliged to cooperate; otherwise he is not. Thus, the deontic inferences from complementary promises and threats need to be compared. This was done in two content versions (mutual lending vs. mutual destruction).

The scenarios of mutual lending stated that Peter would like to borrow Corinna's comic book. He tries to achieve this goal either by a promise ("If you lend me your comic book, then I will lend you my computer game") or by a complementary threat ("If you do *not* lend me your comic book, then I will *not* lend you my computer game"). The mutual destruction scenario concerned two quarreling children. Sarah is about to smash George's Lego car. George would like to prevent Sarah from smashing his car. George knows that Sarah has set up her Playmobil farm. Again, George tries to achieve this goal either by a threat ("If you smash my car, then I will smash your farm") or by a complementary promise ("If you do *not* smash my car, then I will *not* smash your farm"). Altogether, four context stories were used. Each was followed by four tasks that asked for deontic inferences about the speaker's action after the addressee had already cooperated versus the speaker's action after he had not.

Task 1+2: The addressee cooperated: The first two tasks supplemented the context story with the information that the addressee fulfilled the speaker's goal (i.e., Corinna lent her comic book to Peter, and in the other scenario, Sarah did not smash George's Lego car). The first task required participants to decide whether the addressee's cooperation implies an obligation for the speaker to cooperate also. The second task asked whether the speaker is permitted to cooperate. It was expected that the deontic interpretation of the threat would follow the one of the complementary promise in both content versions equally: The speaker is *obliged* to cooperate (i.e., Peter must lend out his computer game while, in the other scenario, George must refrain from smashing Sarah's Playmobil farm), and the speaker is *permitted* to do so.

Task 3+4: The addressee did not cooperate: The other two tasks stated that the addressee did not fulfill the speaker's goal (i.e., Corinna did not lend out her comic book, while Sarah smashed George's Lego car). Again, the participants had to decide whether the speaker is obliged to cooperate and whether he is permitted to do so. This time it was predicted that – independent from the speech act and the content – *no obligation* would arise for the speaker (i.e., Peter need not lend out his computer game and George need not refrain from smashing Sarah's Playmobil farm), but again the speaker is *permitted* to cooperate.

To test the hypothesis that the deontic interpretation of conditional threats is equivalent to the interpretation of the complementary promises, a log-linear analysis (Kennedy, 1992) with two independent variables (speech act and content) was performed for each task. The analyses corroborated the hypotheses: neither the factor *speech act* nor the factor *content* significantly contributed to the data. Both factors could be removed from the analyses without loosing the fit of the resulting log-linear model (for each analysis: $G^2 < 10.5$, $df = 6$, $p > 0.10$). It is thus justifiable to aggregate the data of each task over the four groups.

The aggregated results are shown in Table 4. Most participants drew the deontic inferences that were predicted from the explicit (or implicit) conditional promise: An obligation arises for the speaker only if the addressee A cooperates (67.5% *obligation* vs. 0% *no obligation*), but not if A does not cooperate (10.0% *obligation* vs. 77.5% *no obligation*; $\chi^2(1, n = 62) = 47.8$; $p < 0.001$). Independent from the fact whether the

Table 4: Percentages of deontic inferences aggregated over content versions and speech acts ($N = 40$ in each condition; expected inferences are **bold**-faced).

	The addressee A ...	
	cooperated	did not cooperate
Obligation: Must the speaker S cooperate?		
obligation	**67.5**	10.0
no obligation	0.0	**77.5**
undecidable	32.5	12.5
Permitted: May the speaker S cooperate?		
permitted	**67.5**	**50.0**
not permitted	5.0	22.5
undecidable	27.5	27.5

addressee cooperated or not, the speaker was said to be permitted to cooperate (58.8% *permitted* compared to 20.6% *not permitted* and *undecidable* answers on average; $\chi^2(1, n = 80) = 23.3$; $p < 0.001$).

Summary and Discussion

The results of both experimental parts show a clear and consistent picture that strongly corroborates the predictions from the multi-level analysis.

Conditional inducements are specifically formulated depending on the motivational background and the intended speech act. Thus, conditional promises and threats cannot simply be reversed. This is due to the speaker-addressee-asymmetry: The canonical conditional and its reversal correspond to speech acts of different persons, they have different implications and are associated with complementary action sequences.

It could further be shown that the deontic interpretation of conditional threats is not derived from the conditional formulation, but from the implicit complementary promise. No matter whether a person uses a promise or a threat to pursue his or her goal, there is an obligation to cooperate if the addressee fulfills this goal.

Finally, conditional inducements concern individual goals, actions, and incentives, and are thus highly emotional speech acts. The addressee was said to feel *joy* or *relief* when the speaker kept "the rule" and cooperated, whereas the addressee reacted *angrily* when the speaker broke the rule. This is in line with predictions from appraisal theories of emotion (e.g., Lazarus, 1991; Roseman et al., 1996). Several questions, however, are open to further analyses: Which emotional reactions are associated with other possible action sequences? How does the content of the inducements (e.g., reciprocal exchange or mutual destruction) affect the emotional reactions? Further experiments are needed to answer these questions.

In short, the multi-level analysis of conditional inducements brings together motivation, linguistics, pragmatics, deontic considerations, and emotions. It thereby overcomes the limitations of a purely truth functional analysis often found in reasoning studies.

Acknowledgements I am grateful to Andrea Bender, Stefan Kleinbeck, Gregory Kuhnmünch, and Josef Nerb (Freiburg) who helped to develop the materials and/or gave valuable comments on earlier versions of this paper.

References

Beller, S. (2001). A model theory of deontic reasoning about social norms. In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 63-68). Mahwah, NJ: Lawrence Erlbaum.

Conison, J. (1997). The pragmatics of promise. *Canadian Journal of Law and Jurisprudence, 10*, 273-322.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Evans, J. St. B. T., & Twyman-Musgrove, J. (1998). Conditional reasoning with inducements and advice. *Cognition, 69*, B11-B16.

Fillenbaum, S. (1978). How to do some things with IF. In J. W. Cotton, & R. L. Klatzky (Eds.), *Semantic factors in cognition* (pp. 169-231). Hillsdale, NJ: Lawrence Erlbaum.

Geis, M. L., & Zwicky, A. M. (1971). On invited inferences. *Linguistic Inquiry, 2*, 561-566.

Kennedy, J. J. (1992). *Analyzing qualitative data*. New York: Praeger.

Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Markovits, H., & Lesage, C. (1990). Pragmatic reasoning schemas for conditional promises: Context and representation. In J.-P. Caverni, J.-M. Fabre, & M. Conzalez (Eds.), *Cognitive Biases* (pp. 183-192). North Holland: Elsevier.

Newstead, S. E., Ellis, M. C., Evans, J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3*, 49-76.

Roseman, I. J., Antoniou, A. A., & Jose, P. E. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion, 10*, 241-277.

Searle (1971). What is a speech act? In J. R. Searle (Ed.), *The philosophy of language* (pp. 39-53). London: Oxford University Press.

von Wright, G. H. (1962). On promises. *Theoria, 28*, 276-297.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology.* Harmondsworth: Penguin.

Combining Simplicity and Likelihood in Language and Music

Rens Bod (rens@science.uva.nl)
Cognitive Science Center Amsterdam, University of Amsterdam
Nieuwe Achtergracht 166, Amsterdam, The Netherlands

Abstract

It is widely accepted that the human cognitive system organizes perceptual input into complex hierarchical descriptions which can be represented by tree structures. Tree structures have been used to describe linguistic, musical and visual perception. In this paper, we will investigate whether there exists an underlying model that governs perceptual organization in general. Our key idea is that the cognitive system strives for the simplest structure (the "simplicity principle"), but in doing so it is biased by the likelihood of previous experiences (the "likelihood principle"). We will present a model which combines these two principles by balancing the notion of most likely tree with the notion of shortest derivation. Experiments with linguistic and musical benchmarks (Penn Treebank and Essen Folksong Collection) show that such a combination outperforms models that are based on either simplicity or likelihood alone.

Introduction

It is widely accepted that the human cognitive system organizes perceptual input into complex, hierarchical descriptions which can be represented by tree structures. Tree structures have been used to describe linguistic perception (e.g. Chomsky 1965), musical perception (e.g. Lerdahl & Jackendoff 1983) and visual perception (e.g. Marr 1982). Yet, there seems to be little or no work which emphasizes the commonalities between these different forms of perception and which searches for a general, underlying mechanism which governs all perceptual organization (cf. Leyton 2001). This paper aims to study exactly that question: acknowledging the differences between linguistic, musical and visual information, is there a general, unifying model which can predict the perceived tree structure for sensory input? In studying this question, we will use a strongly empirical methodology: any model that we might hypothesize will be tested against benchmarks such as the linguistically annotated Penn Treebank (Marcus et al. 1993) and the musically annotated Essen Folksong Collection (Schaffrath 1995). While we will argue for a unified model of language, music and vision, we will carry out experiments only with linguistic and musical benchmarks, since no benchmarks of visual tree structures are currently available, to the best of our knowledge.

Figure 1 gives three simple examples of linguistic, musical and visual input with their corresponding tree structures given below.

Thus a tree structure describes how parts of the input combine into constituents and how these constituents combine into a representation for the whole input. Note that the linguistic tree structure is labeled with syntactic categories, whereas the musical and visual tree structures are unlabeled. This is because in language there are syntactic constraints on how words can be combined into larger constituents, while in music (and to a lesser extent in vision) there are no such restrictions: in principle any note may be combined with any other note.

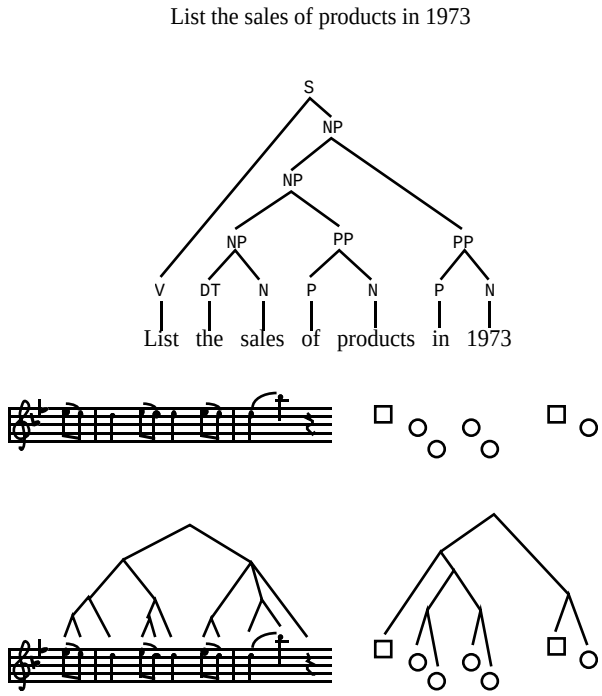
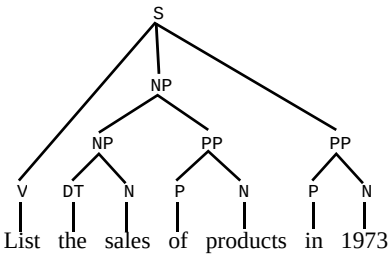


Figure 1: Examples of tree structures.

Apart from these differences, there is also a fundamental commonality: the perceptual input undergoes a process of hierarchical structuring which is not found in the input itself. The main problem is thus: how can we derive the perceived tree structure for a given input? That this problem is not trivial may be illustrated by the fact that the inputs above can also be assigned the following, alternative tree structures in figure 2.



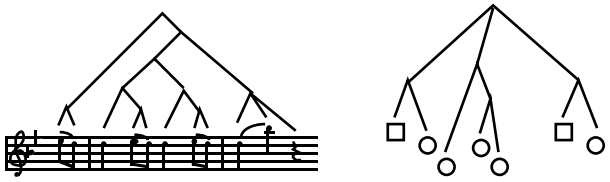


Figure 2: Alternative tree structures for figure 1.

These alternative structures are possible in that they *can* be perceived. But while the alternative tree structures are all possible, they are not plausible: they do not correspond to the structures that are actually perceived by the human perceptual system.

The phenomenon that the same input may be assigned different structural organizations is known as the *ambiguity problem*. This problem is one of the hardest problems in modeling human perception. Even in language, where a phrase-structure grammar may specify which words can be combined into constituents, the ambiguity problem is notoriously hard. Charniak (1997: 37) argues that almost every sentence from the Wall Street Journal has many, often more than one million different parse trees. The ambiguity problem for musical and visual input is even harder. Talking about rhythm perception in music, Longuet-Higgins and Lee (1987) note that "Any given sequence of note values is in principle infinitely ambiguous, but this ambiguity is seldom apparent to the listener."

Two principles: likelihood and simplicity

How can we predict from the set of all possible tree structures the tree that is actually perceived by the human cognitive system? In the field of visual perception, two competing principles have traditionally been proposed to govern perceptual organization. The first, initiated by Helmholtz (1910), advocates the *likelihood principle*: sensory input will be organized into the most probable organization. The second, initiated by Wertheimer (1923) and developed by other Gestalt psychologists, advocates the *simplicity principle*: the perceptual system is viewed as finding the simplest perceptual organization (see Chater 1999 or Van der Helm 2000 for an overview). These two principles are not only relevant for visual perception, but also for linguistic and musical perception. In the following, we briefly discuss these principles for each modality, after which we go into the question of how the two principles can be integrated.

Likelihood

The likelihood principle is particularly influential in the field of natural language processing (see Manning and Schütze 1999, for a review). In this field, the most appropriate tree structure of a sentence is assumed to be its most likely structure. The likelihood of a tree is usually computed from the probabilities of its parts (e.g. phrase-structure rules) taken from a large annotated language corpus (a *treebank*). A widely used treebank for testing and comparing probabilistic natural language parsers is the Penn Wall Street Journal Treebank (Marcus et al. 1993). State-of-the-art probabilistic parsers such as Collins (2000), Charniak (2000) and Bod (2001a) obtain around 90% precision and recall on the Wall Street Journal. Also in the field of psycholinguistics, the likelihood principle is widely used: Jurafsky (1996), Crocker and Brantz (2000) and Hale (2001) are examples of psycholinguistically inspired probabilistic parsers.

The likelihood principle has also been applied to musical perception, e.g. in Raphael (1999) and Bod (2001b/c). As in probabilistic natural language processing, the most probable musical tree structure can be computed from the probabilities of rules or fragments taken from a large annotated musical corpus, for instance from the Essen Folksong Collection (Bod 2001b).

In visual perception psychology and vision science, there has recently been a resurgence of interest in probabilistic models (e.g. Hoffman 1998; Kersten 1999). Mumford (1999) has seen fit to declare the Dawning of Stochasticity.

Simplicity

The simplicity principle has a long tradition in the field of visual perception psychology (e.g. Restle 1970; Leeuwenberg 1971; Simon 1972; Buffart et al. 1983; van der Helm 2000). In this field, a visual pattern is formalized as a constituent structure by means of a "visual coding language" based on primitive elements such as line segments and angles. Perception is described as the process of selecting the simplest structure corresponding to the "shortest encoding" of a visual pattern.

The notion of simplicity has also been applied to musical perception. Collard et al. (1981) use the coding language of Leeuwenberg (1971) to predict the metrical structure for four preludes from Bach's *Well-Tempered Clavier*. More well-known in musical perception is the theory proposed by Lerdahl and Jackendoff (1983) who use a system of preference rules based on the Gestalt-preferences identified by Wertheimer (1923), and which can therefore also be seen as an embodiment of the simplicity principle.

Notions of simplicity also exist in language processing (e.g. Frazier 1978; Gorrell 1995; Osborne 2000). Bod (2000a) defines the simplest tree structure of a sentence as the structure generated by the smallest number of subtrees from a given treebank.

Combining the two principles

The key idea of the current paper is that both principles play a role in perceptual organization: the simplicity principle as a general cognitive preference for economy, and the likelihood principle as a probabilistic bias due to previous perceptual experiences. Informally stated, our working hypothesis is that the human cognitive system strives for maximal economy (the simplest structure), but that in doing so it is biased by the likelihood of previous experiences (in the last section we will discuss some other combinations of simplicity and likelihood that have been proposed). To formally instantiate our working hypothesis, we need a parsing model to start with which can incorporate these principles. In principle any parsing model might do, as long as it can assign tree structures to

perceptual input according to some criterion. For the current paper, we have chosen to start with the Data-Oriented Parsing model (Bod 1998) because (1) it has several other models as special cases, such as context-free parsing models and lexicalized models, and (2) it has been quite successful in predicting tree structures for both linguistic input (Bod 2001a) and musical input (Bod 2001b).

The basic idea of DOP is that it learns a grammar by extracting subtrees from a given treebank and uses these subtrees to analyze fresh input. Suppose we are given the following linguistic treebank of only two trees (we will come back to musical treebanks in the next section),

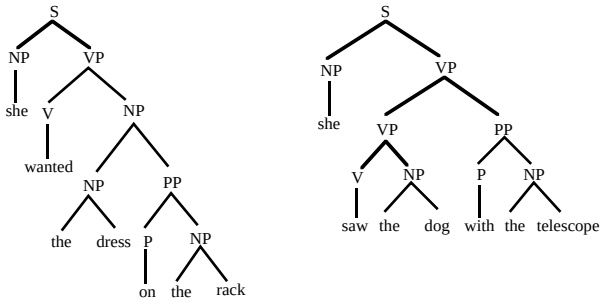


Figure 3: An example treebank

then the DOP model can parse a new sentence, e.g. *She saw the dress with the telescope*, by combining subtrees from this treebank by means of a *node-substitution operation* (indicated as ◦):

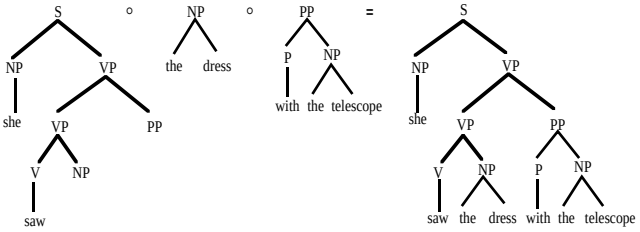


Figure 4: Parsing a sentence by combining subtrees

Thus the node-substitution operation combines two subtrees by substituting the second subtree on the leftmost nonterminal leaf node of the first subtree. Since DOP uses subtrees of arbitrary size, there are typically several derivations, involving different subtrees, that produce the *same* parse tree; for instance:

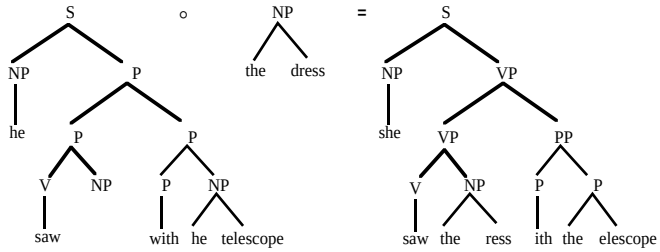


Figure 5: Different derivation producing same tree.

The more interesting case occurs when there are different derivations that produce *different* parse trees. This happens when a sentence is structurally ambiguous; for example, DOP also produces the following alternative parse tree for *She saw the dress with the telescope*:

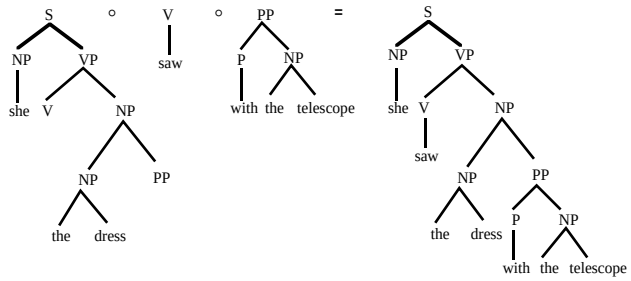


Figure 6: Different derivation producing different tree.

The original DOP model in Bod (1993) uses the likelihood principle to predict the perceived tree structure. We will refer to this model as *Likelihood-DOP*. Likelihood-DOP selects the most likely tree structure from among all possible tree structures on the basis of the probabilities of its subtrees. The probability of a subtree *t* is estimated as the number of occurrences of *t* seen in the corpus, divided by the total number of occurrences of corpus-subtrees that have the same root label as *t*. The probability of a derivation is computed as the product of the probabilities of the subtrees involved in it. Finally, the probability of a parse tree is equal to the sum of the probabilities of all distinct derivations that produce that tree. In Bod (2001a) and Goodman (2002), efficient algorithms are given that compute for an input string the most probable parse tree.

Likelihood-DOP does not do justice to the preference humans display for the simplest structure, e.g. the one that is generated by the shortest derivation consisting of the fewest subtrees. This is what we will call *Simplicity-DOP*. Instead of producing the most probable parse tree for an input, Simplicity-DOP produces the parse tree generated by the fewest corpus-subtrees, independent of the probabilities of these subtrees. For example, given the corpus in Figure 3, the simplest parse tree for *She saw the dress with the telescope* according to Simplicity-DOP is given in Figure 5, since that parse tree can be generated by a derivation of only two corpus-subtrees, while the parse tree in Figure 6 (and any other parse tree) needs at least three corpus-subtrees to be generated. In Bod (2000a) it is shown how the shortest derivation can be efficiently computed by means of a best-first bottom-up chart parsing algorithm. Simplicity-DOP obtains quite impressive results on the WSJ, though its results are lower than Likelihood-DOP (Bod 2000a). Yet, the set of correctly predicted parse trees of Simplicity-DOP is *not* a subset of the set of correctly predicted parse trees of Likelihood-DOP. This suggests that we may expect an accuracy improvement if simplicity and likelihood are combined into a new model, which we will call *Combined-DOP*.

The underlying idea of Combined-DOP is that the human perceptual system searches for the shortest derivation (i.e. the simplest tree structure), but that in doing so it is biased by the "weights" of the subtrees. The length of a derivation is then not defined simply as the sum of the derivation steps (as in Simplicity-DOP), but as the sum of the weights of these steps, where a low weight should be seen as an easy step and a heavy weight as a

difficult step. As a measure for weight of a subtree, we have worked out various proposals that were experimentally tested over the last few years. Most of these proposals have been reported in the literature (e.g. Bod 2000a; Cormons 1999). The best measure for subtree weight so far is based on the *rank* of a subtree. First, all subtrees are grouped with respect to their root label. Next, for each root label the weight of a subtree is defined as its rank in the frequency ordering in the corpus. Thus, the most frequent subtree in each ordering gets a weight of 1, the second most frequent subtree gets a weight of 2, etc. The weight of a derivation is then defined as the sum of the weights of the subtrees in the derivation. The derivation with the lowest weight is taken as the "best" derivation producing the perceived parse tree. Thus, the best derivation is not determined by the smallest sum of the subtrees (as in Simplicity-DOP), but by the smallest sum of the *weights* of the subtrees.

We performed one additional adjustment to the weight of a subtree. This adjustment consists in a smoothing technique which averages the weight of a subtree by the weights of its own *sub*-subtrees. That is, instead of taking only the rank of a subtree as its weight, we compute the weight of a subtree as the (arithmetic) mean of the weights of all its sub-subtrees (including the subtree itself). The effect of this smoothing technique is that it redresses a very low-frequency subtree if it contains high-frequency sub-subtrees.

The Test Domains

Our linguistic test domain consists of sections 02-21 of the Wall Street Journal portion of the Penn Treebank, which contains approx. 40,000 phrase-structure trees. Since the Penn Treebank has been extensively described in the literature (e.g. Marcus et al. 1993; Manning & Schütze 1999), we will not go into it any further here.

The musical test domain consists of the European folksongs in the Essen Folksong Collection (Schaffrath 1995), which correspond to approx. 6,200 musical grouping structures. The Essen Folksong Collection has been previously used by Bod (2001b) and Temperley (2001) to test their musical parsers. The musical coding language used in the Essen Folksong Collection is based on the Essen Associative Code (ESAC). The pitch encodings in ESAC resemble "solfege": scale degree numbers are used to replace the movable syllables "do", "re", "mi", etc. Thus 1 corresponds to "do", 2 corresponds to "re", etc. Chromatic alterations are represented by adding either a "#" or a "b" after the number. The plus ("+") and minus ("-") signs are added before the number if a note falls resp. above or below the principle octave (thus -1, 1 and +1 refer al to "do", but on different octaves). Duration is represented by adding a period or an underscore after the number. A period (".") increases duration by 50% and an underscore ("_") increases duration by 100%; more than one underscore may be added after each number. If a number has no duration indicator, its duration corresponds to the smallest value. A pause is represented by 0, possibly followed by duration indicators. No loudness or timbre indicators are used in ESAC. The only extra information we (automatically) added to the grouping structures in the Essen Folksong Collection consists of the label "S" for each top node of each whole song and the label "P" for each underlying phrase. In this way, we obtained conventional parse trees that can directly be used by our DOP models to parse new input strings (see also Bod 2001b). The Essen Folksong Collection is freely available via http://www.esac-data.org.

As mentioned in the introduction, no visual treebank is currently available, to the best of our knowledge. We are currently developing a treebank of analyzed architectural plans, and will report on experiments with that treebank in due time.

Experimental Evaluation

To evaluate our DOP models, we used the *blind testing* method which dictates that a treebank be randomly divided into a training set and a test set, where the strings from the test set are parsed by means of the subtrees from the training set. We applied the PARSEVAL metrics of *precision* and *recall* to compare a proposed parse tree *P* with the corresponding correct test set parse tree *T* (see Black et al. 1991):

$$\text{Precision} = \frac{\#\text{ correct constituents in } P}{\#\text{ constituents in } P} \qquad \text{Recall} = \frac{\#\text{ correct constituents in } P}{\#\text{ constituents in } T}$$

A constituent in *P* is "correct" if there exists a constituent in *T* of the same label that spans the same elements (i.e. words or notes). To balance precision and recall into a single measure, we will employ the widely used F-score: *F-score = 2*Precision*Recall / (Precision+Recall)*.

We will use this F-score to quantitatively evaluate our models on the Wall Street Journal and the Essen Folksong treebanks. We divided both treebanks into 10 training/test set splits, where 90% of the trees was used for training and 10% for testing. These splits were random, except for one constraint: that all the primitive elements (i.e. words and notes) in the test set also occurred in the training set. In this way, we did not have to worry about unknown words or unknown notes (the latter being actually inexistent for our musical treebank). Although there are various statistical ways to cope with unknown words, we wanted to rule out this problem as it might obscure our comparison.

In our experiments we were first of all interested in comparing the three DOP models (Likelihood-DOP, Simplicity-DOP and Combined-DOP) on the two domains. For computational reasons, we limited the maximum size of the subtrees to depth 14, as in Bod (2001a). Table 1 shows the average F-scores for each of the models.

Table 1: F-scores obtained by the three DOP models

	Likelihood-DOP	Simplicity-DOP	Combined-DOP
Language	90.4%	88.1%	91.7%
Music	86.0%	84.3%	86.9%

The table shows that Likelihood-DOP outperforms Simplicity-DOP, but that Combined-DOP outperforms Likelihood-DOP. According to paired *t*-testing, the

improvement of Combined-DOP over Likelihood-DOP was statistically significant both for language ($p<.0001$) and for music ($p<.04$).

We also performed a series of experiments where we restricted the size of the subtrees. Recall that by restricting the subtrees to depth 1, Likelihood-DOP becomes equivalent to a probabilistic context-free grammar, while Simplicity-DOP would just return the smallest possible tree structure. While Likelihood-DOP still obtained relatively good results at depth 1 for both language and music (resp. 75.1% and 76.6%), Simplicity-DOP scored very badly for language (22.5%) though still reasonably for music (70.0%). Interestingly, Combined-DOP scored worse than Likelihood-DOP at depth 1 (resp. 68.2% vs. 74.6%). Only after subtree depth 6 for language and subtree depth 2 for music, Combined-DOP outperformed Likelihood-DOP. The highest F-scores were obtained with the "unrestricted" subtrees (in table 1).

Elsewhere we have shown that virtually *any* constraint on the subtrees results in an accuracy decrease (Bod 2001a/b). This is because in language, almost any relation between words (including between so-called non-headwords) can be important for predicting the perceived parse tree of a sentence. The same counts for music, where there is a continuity between "jump-phrases" and "non-jump-phrases", which can only be captured by large subtrees (see Bod 2001b/c for an extensive discussion).

Discussion: Other Combinations of Simplicity and Likelihood

We have seen that our combination of simplicity and likelihood is quite rewarding for linguistic and musical perception, suggesting a deep parallel between the two modalities. Yet, we should raise the question whether a model which massively stores and re-uses previously perceived structures has any cognitive plausibility. Interestingly, there is quite some evidence that people store various kinds of previously heard fragments, both in music (Saffran et al. 2000) and language (Jurafsky 2002). But do people store fragments of *arbitrary* size, as proposed by DOP? In his overview article, Jurafsky (2002) reports on a large body of psycholinguistic evidence showing that people not only store lexical items and bigrams, but also frequent phrases and even whole sentences. For the case of sentences, people not only store idiomatic sentences, but also "regular" high-frequency sentences. Thus, at least for language it seems that humans store fragments of arbitrary size provided that these fragments have a certain minimal frequency. However, there seems to be no evidence that people store *all* fragments they hear, as suggested by DOP. Only high-frequency fragments seem to be memorized. However, if the human perceptual faculty needs to *learn* which fragments will be stored, it will initially need to keep track of all fragments (with the possibility of forgetting them) otherwise frequencies can never accumulate. This results in a model which continuously and incrementally updates its fragment memory given new input, which is in correspondence with the DOP approach.

There have been other proposals for integrating or reconciling the principles of simplicity and likelihood. Chater (1999) argues that the principles are identical in the context of Kolmogorov's complexity theory (Kolmogorov 1965). And in the context of Information Theory the simplicity principle can be defined in terms of bit length, such that maximizing likelihood corresponds to minimizing bit length (cf. Rissanen 1978). First note that the likelihood principle aims at maximizing the probability of a structure given an input, $p(structure \mid input)$. Next, define the simplicity principle as minimizing the informatic-theoretical notion of bit length, which is the (negative) logarithm of the probability of a structure given an input: $-log\ p(structure \mid input)$. Now it is easy to see that maximizing $p(structure \mid input)$ leads to the same *structure* as minimizing $-log\ p(structure \mid input)$. Thus the two principles lead to the same result.

However, in the context of DOP we defined the simplest structure as the one generated by the shortest derivation consisting of the smallest number of subtrees (reflecting the smallest number of steps needed to parse an input). And this notion of simplest structure is provably different from the most probable structure given an input. Although it is possible to redefine our notion of simplest structure in terms of bit length, it would not lead to any new model, and to no improved result. By conceptually separating between simplicity and likelihood in DOP and by combining them in a novel way, we have shown that an improved model can be obtained.

What we have not done in this paper is to isolate the perceptual properties for which *no* prior expectations are needed. Even Simplicity-DOP, albeit non-probabilistic, is heavily based on previously perceived data. It is very likely that there are perceptual grouping properties for which no prior expectations are necessary. DOP does not contribute to the discovery of such properties, but it does neither neglect them, as they are implicit in the treebank. Bod (2001b) shows that Wertheimer's Gestalt principles are reflected in about 85% of the phrases in the Essen Folksong Collection (where phrases have boundaries that fall on large time or pitch intervals). DOP automatically takes these principles into account by subtrees that contain such phrases, but DOP also takes into account phrases whose boundaries do *not* fall on large intervals (so-called "jump-phrases"). By using all subtrees, DOP mimics the preferences humans have used in analyzing the perceptual data, whatever these preferences may have been.

References

Black, E. et al. (1991). A Procedure for Quantitatively Comparing the Syntactic Coverage of English, *Proceedings DARPA Speech and Natural Language Workshop*, Pacific Grove, Morgan Kaufmann.

Bod, R. (1993). Using an Annotated Language Corpus as a Virtual Stochastic Grammar, *Proceedings AAAI-93*, Menlo Park, Ca.

Bod, R. (1998). *Beyond Grammar: An Experience-Based Theory of Language*, Stanford: CSLI Publications.

Bod, R. (2000a). Parsing with the Shortest Derivation. *Proceedings COLING-2000*, Saarbrücken, Germany.

Bod, R. (2001a). What is the Minimal Set of Fragments that Achieves Maximal Parse Accuracy? *Proceedings ACL'2001*, Toulouse, France.

Bod, R. (2001b). Memory-Based Models of Music Analysis. *Proceedings International Computer Music Conference* (ICMC'2001), Havana, Cuba.

Bod, R. (2001c). Memory-Based Models of Melodic Analysis: Challenging the Gestalt Principles. *Journal of New Music Research*, 30(3), in press.

Bod, R., J. Hay and S. Jannedy (eds.) (2002a). *Probabilistic Linguistics*. Cambridge, The MIT Press. (in press)

Bod, R., R. Scha and K. Sima'an (eds.) (2002b). *Data-Oriented Parsing*. Stanford, CSLI Publications. (in press)

Buffart, H., E. Leeuwenberg and F. Restle (1983). Analysis of Ambiguity in Visual Pattern Completion. *Journal of Experimental Psychology: Human Perception and Performance*. 9, 980-1000.

Charniak, E. (1993). *Statistical Language Learning*, Cambridge, The MIT Press.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing, *AI Magazine*, Winter 1997, 32-43.

Charniak, E. (2000). A Maximum-Entropy-Inspired Parser. *Proceedings ANLP-NAACL'2000*, Seattle, Washington.

Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *The Quarterly Journal of Experimental Psychology*, 52A(2), 273-302.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*, Cambridge, The MIT Press.

Collard, R., P. Vos and E. Leeuwenberg, (1981). What Melody Tells about Metre in Music. *Zeitschrift für Psychologie*. 189, 25-33.

Collins, M. (2000). Discriminative Reranking for Natural Language Parsing, *Proceedings ICML-2000*, Stanford, Ca.

Cormons, B. (1999). *Analyse et désambiguisation: Une approche à base de corpus (Data-Oriented Parsing) pour les répresentations lexicales fonctionnelles*. PhD thesis, Université de Rennes, France.

Crocker, M. and T. Brants (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29, 647-669.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. PhD. Thesis, University of Connecticut.

Goodman, J. (2002). Efficient Parsing of DOP with PCFG-Reductions. In R. Bod et al. 2002b.

Gorrell, P. (1995). *Syntax and Parsing*. Cambridge University Press.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings NAACL'01*, Pittsburgh, PA.

van der Helm, P. (2000). Simplicity versus Likelihood in Visual Perception: From Surprisals to Precisals. *Psychological Bulletin*, 126(5), 770-799.

von Helmholtz, H. (1910). *Treatise on Physiological Optics* (Vol. 3), Dover, New York.

Hoffman, D. (1998). *Visual Intelligence*. New York, Norton & Company, Inc.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation, *Cognitive Science*, 20, 137-194.

Jurafsky, D. (2002). Probabilistic Modeling in Psycholinguistics: Comprehension and Production. In R. Bod et al. 2002a.

Kersten, D. (1999). High-level vision as statistical inference. In S. Gazzaniga (ed.), *The New Cognitive Neurosciences*, Cambridge, The MIT Press.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission* 1, 1-7.

Leeuwenberg, E. (1971). A Perceptual Coding Language for Perceptual and Auditory Patterns. *American Journal of Psychology*. 84, 307-349.

Lerdahl, F. and R. Jackendoff (1983). *A Generative Theory of Tonal Music*. Cambridge, The MIT Press.

Leyton, M. (2001). *A Generative Theory of Shape*. Heidelberg, Springer-Verlag.

Longuet-Higgins, H. and C. Lee, (1987). The Rhythmic Interpretation of Monophonic Music. In: *Mental Processes: Studies in Cognitive Science*, Cambridge, The MIT Press.

Manning, C. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, The MIT Press.

Marcus, M., B. Santorini and M. Marcinkiewicz (1993). Building a Large Annotated Corpus of English: the Penn Treebank, *Computational Linguistics* 19(2).

Marr, D. (1982). *Vision*. San Francisco, Freeman.

Mumford, D. (1999). The dawning of the age of stochasticity. Based on a lecture at the Accademia Nazionale dei Lincei. (http://www.dam.brown.edu/people/mumford/Papers/Dawning.ps)

Osborne, M. (1999). Minimal Description Length-Based Induction of Definite Clause Grammars for Noun Phrase Identification. *Proceedings EACL Workshop on Computational Natural Language Learning*, Bergen, Norway.

Raphael, C. (1999). Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4), 360-370.

Restle, F. (1970). Theory of Serial Pattern Learning: Structural Trees. *Psychological Review*. 86, 1-24.

Rissanen, J. 1978. Modeling by the shortest data description. *Automatica*, 14, 465-471.

Saffran, J., M. Loman and R. Robertson (2000). Infant Memory for Musical Experiences. *Cognition* 77, B16-23.

Schaffrath, H. (1995). The Essen Folksong Collection in the Humdrum Kern Format. D. Huron (ed.). Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Simon, H. (1972). Complexity and the Representation of Patterned Sequences os Symbols. *Psychological Review*. 79, 369-382.

Temperley, D. (2001). *The Cognition of Basic Musical Structures*. Cambridge, The MIT Press.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung* 4, 301-350.

Mental Models Theory and Anaphora

Guido Boella and **Leonardo Lesmo**
Dipartimento di Informatica and Centro di Scienza Cognitiva
Università di Torino
email: {guido, lesmo}@di.unito.it

Abstract

We argue that anaphora cannot be resolved at the level of the formal language representing meaning, but, rather, by making direct reference to the *extension* of the sentences. Johnson-Laird's mental models theory provide the tool for coping with extensional representations in a cognitively plausible way.

Introduction

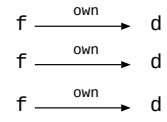
Anaphoric expressions are traditionally viewed as substitutes for more complex linguistic expressions which have already occurred earlier in the text. Anaphora has proven difficult to analyze at a purely syntactic level, so that structural approaches like DRT [10] or semantic ones like Dynamic Semantics [4] cope with this problem by enriching the formal language used to build or to represent the meaning of sentences.

We believe that the limit of these approaches is that they have chosen the wrong level of representation for dealing with anaphora: we will show that it is necessary to make direct reference to *extensional* representations of meaning. In particular, the representation of the context should put at disposal the elements of the situation, which anaphors can refer to, instead of hiding them behind quantified expressions.

However, extensions can possibly be infinite or too large to be dealt with directly. But there is a proposal which uses extensional representations of finite and limited size, and which has been shown to be cognitively plausible, i.e., the *mental models theory* of [9]. Johnson-Laird has used mental models in order to explain how people reason without having to resort to formal logic. Inferences are performed by manipulating extensional representations of sentences which are composed of a finite number of elements and relations: *"a mental model represents the* extension *of an assertion, i.e., the situation it describes, and the recursive machinery for revising the model represents the* intension *of the assertion, i.e., the set of all possible situations it describes.*" (p.100)

In [8]'s words: "*mental models theory is a psychological theory of language processing and reasoning. The theory provides a framework within which more detailed accounts of the component processes of comprehension [...] such as anaphora interpretation [...] and reasoning can be developed, [...] Mental models theory assumes that comprehension results in the construction of representations of situations in the real world [...] These models are finite and computable, and they are constructed incrementally, with the model so far acting as part of the context for interpreting the current text.* (p.20)

A simple preliminary example illustrates our solution. In the following sentences, the acceptability is guaranteed just for the pair where the (intended) antecedent (*a donkey*) and the pronoun (*they*) do not agree in their syntactic number:
(1) *Every farmer owns a donkey. *It is pink.*
(2) *Every farmer owns a donkey. They are pink.*
When the second sentence in each discourse is interpreted, it produces a mental model which must be *integrated* with the preceding one: a referent must be found for the anaphoric expressions. If we examine in the Figure below how the first sentences of the two pairs are represented in mental models theory, we see that the problem is easily solved. The mental model contains a finite number of tokens (placeholders for individuals, here farmers f and donkeys d) and relations among tokens (the arrows labeled with own). Given the model above, which donkey, out of the represented ones, can we relate to the singular *it*, appearing in the second sentence of (1)? The problem of identifying the referent appears to be the same as in: (3) *I have three sisters. *She is blonde* where we have to choose one referent out of three candidates. One is given no (or not enough) information to identify the antecedent (among the three sisters) denoted by *she*. On the contrary, the *they* pronoun in (2) can be interpreted as referring to the set of donkeys appearing in the model, due to its plural syntactic number.



The mental model building algorithm

First of all, the sentence undergoes a syntactic and semantic interpretation process that produces a semantic network (see [6], [11] and [2] for details on the network representation). Then, following the proposal by [9], that "*a propositional representation can be used as the input to a procedural semantics that constructs mental mod-*

els", a mental model representing the meaning of the sentence is built.

The network representation

For the present purposes, we will describe briefly only the mechanism of Distributivity Ambiguity Spaces (DAS) which deals with the possible distributive readings of an NP (see [11] for details).
The nodes of the network can be simple or DASs. The latter correspond to plural NPs, and they were introduced to deal with the distinction between *collective* and *distributive* readings of predicates: each DAS includes two subnodes *Set* and *Indiv*.

In case of (4) *Three men lifted three tables*, if the subject NP is given a reading as a set, the men are seen as being jointly involved in the act of lifting tables. Viceversa in the individual reading of the subject, each man executed a separate lifting act. If the tables are interpreted as a set too, they were lifted all together (perhaps they were stacked). On the contrary, if they are interpreted as individuals, the men lifted them one at a time. The four combinations of *Set, Indiv* readings for the subject and the object do not cover all possibilities. In fact, it may happen that, for the *Indiv* reading of the subject, there exist just three tables, and each man lifted one of them (three individual lifting acts); or that each man lifted three tables (possibly, but not necessarily, the same three tables; 9 different tables could be involved), so that nine individual lifting acts have been executed. Or, in the *Set* reading of the object, the three men lifted three different stacks of tables (so, we have two more readings, for a total of 6) The extra readings (see Figure 1) are accounted for by means of a mechanism other than the DAS described above (but independently motivated, see [11]), i.e. by the presence of DEP-ON (dependent on) arcs. They are similar to Skolem functions in first order logic, and were introduced for representing quantifier scoping. Each node which is not universally quantified can be specified to be dependent on another 'plural' node. For instance, in *Every farmer owns a donkey*, the most natural reading is where each farmer owns a different donkey, so that the particular donkey 'depends on' the particular farmer.

Mental models

In order to use a more unambiguous version of the framework with respect to the 'diagrammatic' original version of [9], we refer to the formalization of mental models provided by [1].

According to [1], a model is triple $<$ T, R, A $>$, where T is a (non-empty) bi-dimensional matrix of tokens, R is a set of relations on T, and A is a set of annotations. For dealing with some interpretations, more than one model can be required.
A token is either a model or an element. An element is a pair $<$ S, A $>$ where S is a symbol from a given vocabulary and A is a set of annotations; the vocabulary consists of named individual entities (john for the proper name *John*) and generic entities belonging to some category (c_i for cars, f_i for farmers, etc.).

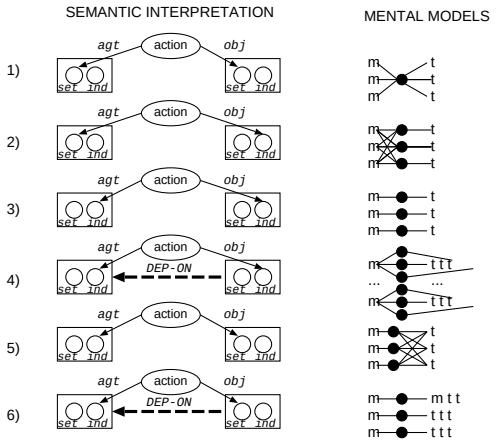
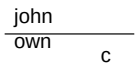


Figure 1: Six readings of (4).

A relation is an ordered sequence $<$ r, x_1, ..., x_n, A $>$ where r is a relation symbol, x_1, ..., x_n are tokens in T and A is a set of annotations. Annotations are "*the propositional enrichment of the analogical structure of the model*" [1]. In particular, the "not" annotation applies to any feature of the models. For models and relations, a negation means that they are not the case; for entities, that they are absent in a model. The "..." annotation means that the model can be further extended.
[1] consider relations such as 'above', 'faster' and two special relations "connected with" (CW) and "never connected with" (NCW). The CW relation forms an individual by connecting two of its properties. The NCW one states that two properties cannot hold for the same individual. Usually the two relations are used to represent the meaning of, respectively, *all humans are mortals* and *most lawyers are not poor*. With respect to [1]'s framework, we introduce an extension for what concerns the NCW relation. In fact, NCW is originally meant to apply only to unary predicates such as being humans or mortal. We introduce a version of the NCW relation relativized to a predicate rel, NCW(rel). In fact, the "not" annotation of a relation means that the relation is not true of the given entities involved in the relation. In $<$ faster, john, bill, $\{$not$\}$ $>$, the negation does not concern the existence or not of the two individuals John and Bill, which are introduced as existing entities. But this is not sufficient to represent the meaning of a sentence like *John does not have a car*: since the phrase *a car* inside a negation does not introduce or refer to an entity in the model, the meaning of the sentence cannot be represented by the negation of the 'have' relation: in fact, a relation as $<$ have, john, c_1, $\{$not$\}$ $>$ does not express the fact that from the model it is not possible to infer that there is a car. Rather, this annotated relation expresses the fact that there is a car in the model and John is not its owner.
What we need is something similar to the interpretation of the sentence *no lawyer is a crook* from which is not possible to infer that there is some crook in the model. The model of this sentence in [1] is not represented by a negation of some predicate 'is' but with the NCW rela-

tion discussed above: $<$ NCW, lawyer$_i$, crook$_i$ $>$. Analogously, for interpreting *John does not have a car* we introduce a NCW(have) predicate, which not only expresses the negation of the "have" predicate, but which also does not assert the existence of any car (see the Figure below).

In $<$NCW(have), john, c_1, $\emptyset$ $>$, cars, as in $<$ NCW, bicycle$_i$, c_i, $\emptyset$ $>$ (*no bicycle is a car*), are kept separate from the other entities in the model: they cannot play the role of antecedents of pronouns.



For what concerns the treatment of logical connectives, we stick to the proposal of [1].

From the network to the mental model

The model constructing procedure takes as input an existing mental model (representing the context) and the network representation of the new sentence (still associated with the syntactic tree): the newly constructed model is *integrated* with the existing ones by overlapping identical tokens and finding referent tokens for anaphoric expressions.

The process starts from the non-dependent entity nodes of the network which derive from the interpretation of NPs (i.e. NPs without exiting DEP-ON arcs), and proceeds with the other NPs, according to the (partial) order imposed by (reversed) DEP-ON arcs. After that, all co-references are solved. For instance, in (5) *Every farmer who owns a donkey beats it*, *every farmer* is processed first, then *a donkey* and, the pronoun *it* which depend on the subject NP.

More precisely, given a context M composed by a model $<$ T, R, A $>$, we have that a network W is interpreted as a new model $<$ T$'$, R$'$, A$'$ $>$, in the following way:

1. Each non-dependent entity node in the network W deriving from the interpretation of an NP is treated separately:

   (a) If the entity node is represents an NP which is a proper noun (e.g., *John*), an individual token (e.g., john) is introduced in the matrix T of the model M; if that token is already present in the model, the two tokens are identified.

   (b) If the NP is a quantified Noun (e.g., *every farmer*), a set of distinct tokens F = $\{x_1, \ldots, x_n\}$ representing the denotation of the noun is added to the context matrix T; depending on the quantifier Q, a subset of them, Q(F), will be selected for linking to other tokens by the relation where the NP occurs as an argument (selecting the whole set in case of *every* and *all*, a proportioned subset of it in case of *most*, etc). The annotation A of the model can be augmented with a "...", since, depending on the quantifier, more tokens could be added to the matrix T or the set $Q(F)$ could be revised (e.g., if Q="some", $\mid Q(F) \mid$ could be initially 2 or 3, but it can be increased in case of necessity, as in the standard treatment of syllogism in [9]).
   A special case, as in the mental models theory of [9], is represented by the quantifier *no*: its meaning is represented by selecting all the tokens F representing the denotation of the noun it quantifies (Q(F)=F); but when the relation rel involving the NP is introduced, it is interpreted as negated either in the sense of a NCW(rel) relation or in the sense of being annotated as negated. As an example, in *no farmer owns a donkey* the owning relation, is transformed in a NCW(own) relation which keeps apart all the farmers from the set of donkeys.

   (c) If the NP is an indefinite such as *a car*, two cases are possible according to the presence of a negation and the role played by the NP in the main predicate:[1]

   - If the NP is the subject of the verb or it appears in a non-negated relation, a single new token representing a car is added to the matrix T of the model and annotated as "...", since it does not convey any uniqueness presupposition.
   - If the NP appears in a negated predicate and it is not the subject of the predicate rel, some tokens representing the denotation of the noun F = $\{x_1, \ldots, x_n\}$ are introduced in T and appear in a NCW(rel) relation to keep them separate from the other tokens of the model.[2]

   (d) If the entity in the network W is the interpretation of a definite NP or a definite pronoun, then an antecedent must be searched for in the mental model constructed so far; according to the number, one or more tokens existing in the model are sought in T to act as the potential referents: further, the set of relations R must satisfy the description provided by the NP. This kind of unification, however, cannot be accomplished with items which are linked to other ones only by a NCW(rel) relation in which they appear in a non-subject role $\{t_i \mid \exists$ rel, $x_1, \ldots, x_n$ ($<$ NCW(rel), $x_1, \ldots, t_i, \ldots, x_n, \emptyset > \in R \wedge i \neq 1)\}$, i.e., these items are implicitly assumed as 'non existing' in the model. Moreover, if the set of possible referents X = $\{t_1, \ldots, t_n\}$ is composed of a subset of tokens which occur in relations with other tokens and a subset of tokens which are unrelated:
   $\{t_i \mid \exists$ rel, A, $x_1, \ldots, x_n$ ($<$ rel, $x_1, \ldots, t_i, \ldots, x_n$, A$>$ $\in$ R) $\} \cup \{t_j \mid \neg\exists$ rel, A, $x_1, \ldots, x_n$ ($<$ rel, $x_1, \ldots, t_j$, $\ldots, x_n$, A$>$ $\in$ R) $\}$
   then only the former set can be considered by the uni-

[1] Note that *John does not love a girl in his office* where the indefinite is a *specific* one (see [10]) and the speaker could identify a unique referent for it, is not covered by this rule.

[2] This treatment of indefinites is justified also from a linguistic point of view. As [10] notice, the negation of a verb must be interpreted as having an inner scope which does not include the subject of the verb, otherwise sentences as *someone does not like a Porsche* would be true in case there is no people at all. And it finds a similarity in DRT where indefinites inside the scope of a negation are interpreted in a subordinate DRT structure which will not be accessible for the resolution of anaphoric expressions.

fication process (e.g., in the interpretation of *John has many donkeys. They are pink* where the model includes a number of donkeys but only a subset of them is related with John: the pronoun *they* refers only to this subset).
Note that the set of annotations is not constrained to be empty: in fact, it is possible to make reference to a set of entities which is involved in a negated relation as in: (6) *the soldier didn't see some of the enemies. They were hiding in the trees.*

Finally, since a definite pronoun is a *definite* reference, the found referent must be non-ambiguous: if different possibilities exist, then, for pragmatics reasons, the reference fails (see example (3)).

2. If the entity node of the NP np_1 is "dependent on" another node which is built from the NP np_2, its interpretation depends on the one of np_2: this means that, for each token built in correspondence with np_2 the interpretation of np_1 must be repeated according to the rules in 1 described above for non-dependent NPs. In particular, if np_1 is a singular *indefinite* and the corresponding relation is not negated, a new token is introduced for each token associated with np_2; if np_1 is plural, a different set of example tokens is added to the model for each token associated with np_2.
For example in the distributive interpretation of *Every farmer has a donkey. They beat it*, *they* is unified with the tokens $f_1, \ldots, f_n$ representing farmers, but the interpretation of *it* (which in this reading cannot but be dependent on *them*) is performed for each f_i ($1 \leq i \leq n$) relatively to the set of tokens $\{ t_j \mid \exists$ rel, $x_1, \ldots, x_n$ ($<$ rel, $x_1, \ldots, f_i, \ldots, t_j, \ldots, x_n$, $\emptyset> \in$ R) $\}$. In the example, for each i, *it* is unified with the d_i such that $<$beat, f_i, d_i, $\emptyset >$.
3. Finally, the tokens are linked by the relations described by the predicates. The number of relations which are introduced depends on the *set* or *individual* interpretation of the DAS of the NPs involved: if an NP is considered as a set, the tokens resulting from its interpretation are included as a whole in the role they play in the relation. Otherwise, each element of the set is introduced in different instance of the relation.
4. As we discuss in the following Section, the interpretation of a sentence which includes logical connectives can result in more than one model. The rule 1 is iterated for each of the clauses in the complex sentence.
During the interpretation process some of the possible models must be discharged as inconsistent. This is a correct move but it can lead to the refection of the sentence for pragmatic reasons (as in example (11) below). In fact, if the interpretation of a sentence results in a reduced set of models which can be better described by another sentence (that is, its interpretation does not discard any model), then by the Gricean principle of cooperation, the speaker should have used it instead of the one he chose.
5. On the other hand, if the interpretation of the sentence leads felicitously to a set of models, these models become part of the context. When a subsequent sentence is interpreted, its interpretation must be compatible with *all* the models in the context. In particular, if the interpretation of the subsequent sentence produces more than one model, for each model in the context, at least one of the newly constructed models must be compatible (even if not the same one for all the model in the context). Otherwise, the sentence will be rejected (as in example (14) below).

Logical connectives

According to [10] the interplay of anaphora and logical connectives is a fundamental testbed for any theory of language interpretation. Here, the meaning of connectives is expressed by their possible models in [9]'s style. First the implicit models are constructed and if necessary the explicit ones are fleshed out.
Let's start with a simple example involving negation: (7) **John does not own a car. He washes it.*
Since, according to the representation outlined in the previous section, cars are included in NCW(own) relations, no referent can be found in the model for the pronoun *it*: $<$ T={{$\emptyset$, c_1}, {john, $\emptyset$},}}, R= {$<$NCW(own), john, c_1, $\emptyset >$, A= $\emptyset$ } $>$
So, the sentence is not interpretable according to that reading.
An example a bit more complex is: (8) *No farmer has a car. *It is red.*
A sentence like *no farmer is rich* is represented by a NCW relation between farmers and rich people see rule 1.b. In our model, this relation is extended to arbitrary predicates. Hence, the first sentence produces a model where cars appear in the set of *never connected with* entities, so that the interpretation (and failure in integration) is exactly the same as in the previous example:
$<$ T={ {{$\emptyset$,c_1}{$\emptyset$,c_2}{$\emptyset$,c_3}}, {{f_1, $\emptyset$}, {f_2, $\emptyset$}, {f_3, $\emptyset$}} }, R= {$<$NCW(own), f_1, $c_1 >$, $\emptyset >$, $<$ NCW(own), f_2, $c_2 >$, $\emptyset >$, $<$NCW(own), f_3, $c_3 >$, $\emptyset >$}, A= {...} }
On the contrary: (9) *No farmer has a car. They prefer donkeys.* is acceptable, in spite of the negation appearing in the subject NP and of its singular number. In fact, the farmers (appearing as 'existing' entities) are available for integration.

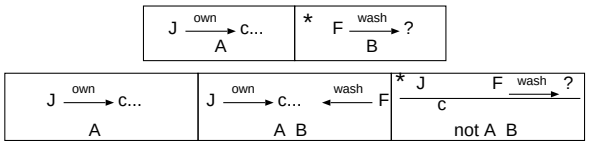
If we now consider conjunctions and disjunctions, another interesting anomaly arises:

(10) *John owns a car$_i$ and Fred washes it$_i$*
(11) **John owns a car$_i$ or Fred washes it$_i$*

The syntactic structures are identical but the acceptability is not. In order to explain this fact, [10] introduced an accessibility constraint at the structural level: "*no disjunct of a disjunctive condition is accessible from any other*".

The mental model representation of a conjunction involves the inclusion in the same model of the conjoined sentences. So, no problem arises with (10), since the referent for *it* can be found in the same model where the second conjunct must be integrated. Compare the unacceptability of **John does not have a Porsche and Fred washes it*.

On the contrary, a disjunction A ∨ B requires the construction of two *separate* models (one with A and one with B). So, in the second model of (11) there is no available referent for the pronoun (B, i.e., *Fred washes it*). But when a difficulty (such as the impossibility of understanding a sentence) occurs in a mental model, the model can be manipulated and fleshed out; in principle, when applied to the second model (B), this process could produce two alternatives. In the first one John owns a car (A and B), while in the second one he does not (not A and B, i.e., *John does not have a car and Fred washes it*). So, it seems that the first extension could solve the problem: John, in fact, owns a car and Fred washes it:



The lower part of the figure shows the three resulting models: the second one (A and B) includes the first (A) and the third (not A and B) is discharged since Fred cannot wash a car which does not exist (see the * in the third box).
However, it seems that from any disjunction at least two distinct models must be constructed, and that none of them must be included in the other: otherwise, the common part of the two models would be necessarily true and according to Grice, the speaker should not have used a disjunction to express such a meaning (see rule 4 of the interpretation algorithm).

An example that supports the previous analysis is the acceptability of the following sentence, as the reader can easily test: (12) *John does not own a car or he washes it*. The second model (the one of *he washes it*) can be fleshed out with the negation of the first disjunct (not A and B, i.e., John does own a car and Fred washes it): after this extension, the resulting model puts at disposal the required referent for the pronoun. We are left with two different models, equivalent to the [10]'s interpretation of the example, i.e *John does not own a car* or *John owns a car and he washes it*. In [10] this result is obtained by copying the negation of the first conjunct in the second DRS: such a rule, however, presupposes that disjunctions in natural language are always interpreted a exclusive disjunctions.
It is interesting to note that Dynamic Semantics [4], in order to explain this kind of examples, has to introduce a new class of anaphora, *E-type*.

The last connective to be considered is implication: (13) *If John owns a car, he washes it.*
This sentence involves two models (A and B, and not A to be further extended): in the first one, John owns a car and washes it, while in the second one he does not own any car. But, if the sentence is followed by (14) **It is a Porsche*, the pronoun of the second sentence can find a referent only in the first model in the context, while no antecedent is accessible in the second one (where the car, in fact, is connected with a NCW relation). As prescribed by rule 4 of the algorithm, a new sentence must be integrated with all the pre-existing models in the context, otherwise it is unacceptable.

An analogous reasoning explains the oddity of **If John does not own a car, Fred washes it*. Moreover, so called *examination sentences* as (15) *no students will be admitted to the exam unless they have registered four weeks in advance* can be dealt with by interpreting the conjunction *unless* as an *if not* or a stronger *exclusive or*.

Plurals

A number of interesting anaphoric phenomena are related to *plurals*. The first situation concerns a plural pronoun referring to a set of singular antecedents which occur in the model: (16) *Mary met John. They talked.*

Even if in mental models the first sentence introduces separately two tokens, the rule 1.d can cope with these case as if they were introduced simultaneously as by the NP *two people*. So it can resolve the anaphora without explicitly *summing* the antecedents to form a plural discourse referent, as DRT does.

Quantifier phrases such as *many of the farmers* do not always introduce the referents with which subsequent pronouns will be co-referential. For example, those pronouns refer to sets that have to be constructed from explicit information in the text. Here, quantifiers introduce in the model a set of tokens which pronouns can refer to in the subsequent discourse: (17) *Susan has found every book which Bill needs. They are on his desk.*
To resolve *they* we need only the set of tokens introduced in the model by the analysis of the first sentence. The right subset of books is identified in the model thanks to rule 1.d (see discussion below). In DRT, in contrast, a new discourse referent is constructed via an *abstraction* rule which copies the content of the DRSs introduced by the previous sentence.

In some approaches, definite NPs and pronouns inside the scope of a quantifier are considered like bound variables in a logical system. In (18) *every waiter wants customers to give him large tips* the pronoun does not seem to refer to any particular entity, while it does in (19) *John wants customers to give him large tips*. DRT in order to deal with both cases in a uniform way introduces the notion of discourse referent which does not correspond directly to any individual in the world, while providing antecedents for the pronouns.
In contrast, mental models theory allows unifying both cases, since quantifier phrases, in an extensional representation introduce sets of entities in the model.

Not all definite pronouns following quantifiers behave like bound variables, in particular, if they appear in a following clause, i.e., outside the quantifier scope: (20) *few congressmen admire Kennedy and they are very junior*. *They* refers to those (few) congressmen who admire Kennedy, even if there is no such an expression referring to them. If the pronoun were interpreted as a variable, the sentence would be equivalent to (21) *Few congressmen admire Kennedy and are very junior*. In [5]'s terms, there

is an *antecedent trigger*, a linguistic expression which introduces the antecedent of the pronoun but it does not have the same referent of the pronoun.
In our model, after the interpretation of the first clause the mental model contains the set of congressmen and a (small) subset of them which are in an "admire" relation with Kennedy. For rule 1.d above, the definite pronoun *they* can be resolved with this subset.

But as it might be expected, quantifiers focus on the subset of the set specified by the head noun. Hence, the unification process must be suitably constrained. In: (22) *Some farmers of this valley own a donkey. They don't like cars*, the pronoun *they* can in principle refer either to the *farmers of this valley* who own a donkey or to the complement set; according to rule 1.d in the interpretation algorithm, if the set of candidate referents can be partitioned in different sets, the pronoun is unified only with the entities which are involved in a relation (of owning a donkey).

The possibility of a plural anaphor resolved against referents described by a singular indefinite is explained by rule 1.d which deals with the interpretation of dependent NPs in distributive readings (see the Figure on first page). In (23) *Every farmer owns a donkey. They are pink* the distributive reading expresses explicitly the plurality of donkeys so that the correct referents are available for the plural pronoun. In contrast, in (24) *Every farmer owns a donkey. *He is a wise man*, the singular definite pronoun *he* cannot be resolved, since we do not have any information to choose one of the farmers (see rule 1.d).
An example slightly more complex is: (25) *Three farmers own a donkey. They beat them*. The latter sentence can be interpreted only as far the second clause is interpreted with two individual readings of the NP without DEP-ON arcs between them (case 3 of Figure 1). In this case the donkeys who form the antecedent of *them* are related each by a different relation with the farmers. Which farmer is selected for relating with the beating relation to a given donkey? as in case of rule 2 the interpretation of an anaphor is performed exactly with respect to the other tokens which are linked to it by some relation. Indeed, in the context is maintained the relation between each farmer and the donkey he owns: hence, the interpretation of the sentence leads to a situation where each farmer beats the donkey he owns, and not a different one (as it happens in some formal models of anaphora).
Finally, plural and singular pronouns can be mixed: (26) *Every farmer owns a donkey. They beat it.* Since *it* in the second sentence is dependent on the subject *they*, the interpretation of the second sentence is parallel to the interpretation of the first: the object (*it*) can be resolved against an antecedent only if it is interpreted as dependent on the subject; according to rule 1.d of the interpretation algorithm, *they* is unified with the set of farmers appearing in the model and, again, since there is an explicit relation (*own*) linking each of them to a donkey, this link is followed to determine the (singular) referent of *it*.

Similarly, the so called *donkey sentence*, (27) *Every farmer who owns a donkey beats it*, is acceptable: the procedure first interprets the subject phrase, thus obtaining, in the wide-scope reading of the universal, a representation where each farmer has at least a donkey; then it extends the representation by searching for a referent through each relation $<$own, f_i, d_i, $\emptyset >$; so, in the distributive reading the sentence, as in the example above, for each farmer, a different referent for *it* is found, i.e. the donkey owned by him. The possible antecedent must satisfy the restriction carried by the number of the singular pronoun (compare **Every farmer who has two donkeys beats it*).

Conclusion

Since mental models are a cognitively plausible theory of human reasoning, they can be also useful in finding an explanation of linguistic phenomena. In [3], we exploited mental models to provide an explanation of lexically triggered presuppositions. In [2] more complex anaphorical phenomena related to the different readings of *donkey sentences* have been coped with in the same framework.

The limit of logical approaches in explaining anaphora is that they exploit representations that are not isomorphic to our conception of the described situation. The necessity of resorting to a referential level in explaining anaphora has been highlighted also by [7]. We followed his suggestion, but going in a different direction, where mental models replace the classical model-theoretic framework to provide a cognitively plausible approach to language interpretation.

References

[1] B. Bara, M. Bucciarelli, and V. Lombardo. Model theory of deduction: A unified computational approach. *Cognitive Science*, 25(6), 2001.

[2] G. Boella, R. Damiano, and L. Lesmo. Beating a donkey: a mental model approach to complex anaphorical phenomena. In *Proc. of European Congress of Cognitive Science of ECCS'99*, Pontignano, 1999.

[3] G. Boella, R. Damiano, and L. Lesmo. Mental models and pragmatics: the case of presuppositions. CogSci99 Conference, 1999.

[4] G. Chierchia. Anaphora and dynamic binding. *Linguistics and Philosophy*, 15:111–183, 1992.

[5] F. Cornish. Antecedentless anaphors: Deixis, anaphora or what? *Journal of Linguistics*, (32):19–41, 1996.

[6] B. DiEugenio and L. Lesmo. Representation and interpretation of determiners in natural language. In *Proc. 10th IJCAI*, pages 648–653, Milano, 1987.

[7] D. A. H. Elworthy. A theory of anaphoric information. *Linguistics and Philosophy*, 18:207–332, 1995.

[8] A. Garnham. *Mental models and the interpretation of anaphora*. Psychology Press, Hove, 2001.

[9] P.N. Johnson-Laird. *Mental Models*. Cambridge University Press, Cambridge, 1983.

[10] Hans Kamp and Uwe Reyle, editors. *From Discourse to Logic*. Kluwer, Dordrecht, 1993.

[11] L. Lesmo, M. Berti, and P. Terenziani. A network formalism for representing natural language quantifiers. In *Proceedings of ECAI-88*, pages 473–478, 1988.

Comparison and the Development of Knowledge

Lera Boroditsky (lera@mit.edu)
NE20-456, MIT, 77 Mass Ave
Cambridge, MA 02139 USA

Abstract

This paper considers the role of comparison in the development of knowledge. Results show that comparing similar objects makes them appear more similar, while comparing dissimilar objects makes them appear less similar. The effect of comparison on similar items was especially striking since subjects judged items to be more similar after comparison even if the comparison task was to list differences between the two items. Further, this effect appears specific to comparison and does not appear to be simply due to a "fleshing out" of object representations (listing properties of two objects without comparing the objects themselves served to increase the objects' similarity regardless of whether the objects were similar or dissimilar to start). This suggests that comparison may play a special role in partitioning bits of experience into categories, sharpening categorical boundaries, and otherwise helping us create conceptual structure above and beyond that offered by the world.

Introduction

Are our mental representations of things in the world simply a reflection of the structure of the world, or do we create new structures and partitions in conceptual space? Further, are our representations static, or do they change over time in systematic ways as a result of the way we process and use our knowledge? This paper suggests that some common cognitive processes (in this case, comparison) can introduce systematic biases into our representations of the world. These biases may be beneficial for separating out bits of experience into categories, sharpening categorical boundaries, and otherwise helping us create conceptual structure above and beyond that offered by the world.

This paper focuses on object similarity. Similarity is a central construct in explanations of cognition. Explanations of categorization, induction, learning, and memory all rely on the construct of similarity. Things that are similar are likely to end up in the same categories, are likely to support inductive inferences for each other, will aid in the learning of other similar things, and serve as good reminders for one another in memory. But where do similarities come from? Are similarities between objects apprehended immediately and automatically, or do they develop as a function of directed processing and experience?

This question has been taken up seriously in the study of categorization in the following form: Why do categories appear to contain similar things? Is it because similar things tend to end up in the same categories, or is it that putting two things in the same category makes them appear more similar? Previous research suggests that both are true. For example, Goldstone, Lippa, and Schiffrin, (2001) showed that object representations can change as a result of category-learning, with objects assigned to the same categories becoming more similar (see also Kurtz, 1998). Further, previous research by Gentner and Namy (2000) suggests that providing children with an opportunity for comparison may help them in category learning by allowing them to discover deeper relational similarities between category members (see also Kurtz & Gentner, 1998).

This paper considers the role of comparison in the development of similarity. Results of four experiments suggest that comparison can play an important role in knowledge development. By making similar things appear more similar, and dissimilar things appear less similar comparison may help us partition bits of experience into categories and sharpen categorical boundaries.

Four experiments explore the effects of comparison on object representation. Experiments 1 and 2 examine the effects of comparison on the perceived similarity of similar and dissimilar objects. Experiment 3 contrasts the effects of comparison with those of simple "fleshing out" or elaboration of object representations. Experiment 4 extends the findings of Experiments 1and 2 to novel objects.

Experiment 1

Method

Participants 132 Stanford University undergraduates participated in the study in order to fulfill a course requirement.

Materials Materials consisted of a one-page questionnaire. The top of the page contained line-drawings of 4 named familiar animals (a deer, a horse, a goat, and a donkey) as shown in Figure 1a. The rest of the page contained three questions. For 73 participants, the first question asked them to describe three similarities between two of the animals (e.g., "Please describe 3 similarities between the goat and the donkey.") For the other 59 participants, the first question asked them to describe three differences between two of the animals (e.g., "Please describe 3 differences between the goat and the donkey.") Participants were given three blank lines for their responses. Which two animals were chosen for comparison was counterbalanced across subjects such that each pair of adjacent shapes was the focus of comparison equally often. Which animal was named first in the comparison was also counterbalanced across subjects.

The last two questions asked participants to rate the similarity of the two animals they had just compared (e.g., "How similar are the goat and the donkey?"), and of the other two animals (e.g., " How similar are the deer and the horse?"). Half of the subjects rated similarity for the previously compared pair first, and the other half rated similarity for the other pair first. As before, which animal was named first in each comparison was counterbalanced across subjects. Subjects rated similarity on a 10-point scale (1=not similar and 10=very similar).

Procedures The one page questionnaire was embedded in a larger questionnaire packet which contained many other pages unrelated to this study. Participants completed the questionnaire at home on their own time.

Results

Comparing two similar items made people think of them as more similar. This was true regardless of whether the comparison involved naming similarities between the two items (M=6.42 after naming similarities, M=5.93 without naming similarities, $t=2.01$, $p<.05$) or naming their differences (M=6.69 after naming differences, M=6.30 without naming differences, $t=1.89$, $p<.05$). There was an overall effect of comparison ($F(1,130)=7.11$, $p<.01$) and no interaction between the two comparison types ($F(1,130)=.08$, $p=.78$)).

Discussion

Experiment 1 showed that comparing two things (even when looking for their differences) can cause people to discover similarities between the two things. But why should the similarity of two objects increase after they are compared, especially if one's task is to describe their differences? One possibility is that in the process of finding and articulating differences, people are also finding similarities. As shown by Gentner & Markman (1994), the most meaningful (and easiest to name) differences are those that are attached to the structural similarities. On this view, because the process of comparison involves an alignment between two representational structures (see Markman & Gentner, 1993a, 1993b, 1996), discovering meaningful differences involves first establishing the similarities. To take a particular example, if one wanted to mention that the goat has a shorter tail than the donkey (a difference), this makes salient the fact that both animals have tails (a similarity).

But there could also be a less interesting explanation for these results. What if similarity only increases after comparison because people create a new feature for the things they compare, something like "thing I compared before." If this is the case, similarity might be increasing simply because the two things previously compared now both have this extra feature in common. One way to test this possibility, is to ask people to carry out comparisons between things that are so different, that no meaningful similarities are likely to be found. If comparison no longer serves to increase perceived similarity, then it is the ability to find meaningful similarities (and not just the creation of an extra feature) that is responsible for the findings of Experiment 1. In Experiment 2, the pictures of four similar animals used in Experiment 1 were replaced with pictures of four quite dissimilar objects: a phone, a pretzel, a hat, and a football.

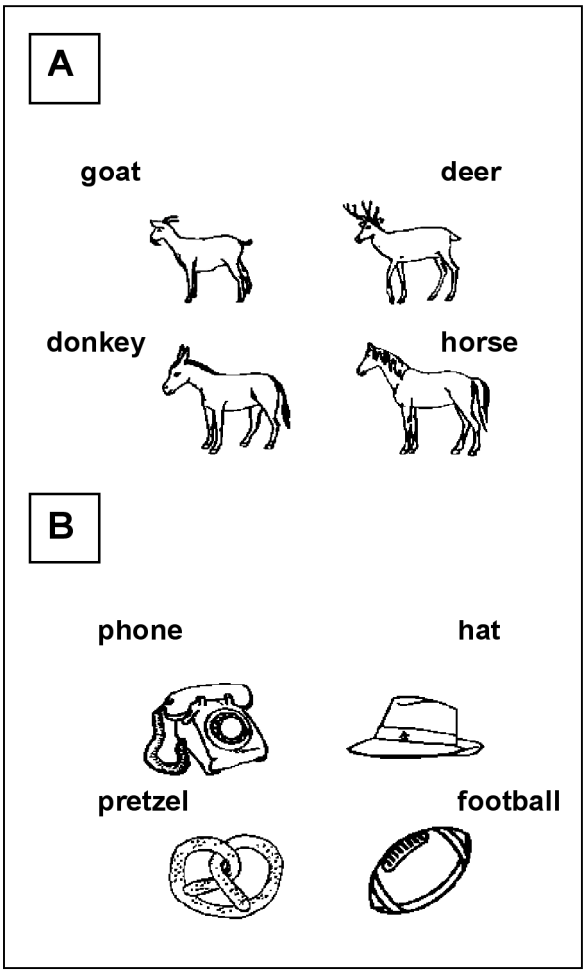


Figure 1: Stimuli used in Experiments 1 and 2 (shown in parts A and B respectively). Images were taken from Snodgrass, J. G., & Vanderwart, M. (1980).

Experiment 2

Method

Participants 61 Stanford University undergraduates participated in the study in order to fulfill a course requirement.

Materials Just as in Experiment 1, materials consisted of a one-page questionnaire. The top of the page contained pictures of 4 dissimilar objects as shown in Figure 1b. The rest of the page was constructed just as described for

Experiment 1. The comparison task for all 61 of the subjects was to name differences between two items.

Procedures The procedures were the same as in Experiment 1.

Results

First, the stimuli in this experiment were indeed perceived to be much less similar to each other (M=2.60) than those used in Experiment 1 (M=6.50), $F(1,118)=278.9$, $p<.001$. This was necessary as a manipulation check.

It turned out that comparing two dissimilar objects did not increase their similarity. In fact, naming differences between two dissimilar objects actually made participants think of the objects as less similar (M=2.41), than if they hadn't compared them before (M=2.79), $t=-1.89$, $p<.05$. This pattern was significantly different from that observed in Experiment 1 as confirmed in an interaction in a 2x2 repeated measures ANOVA (2 (named differences or not) X 2 (stimuli similar or dissimilar)), $F(1,118)=6.90$, $p=.01$.

Discussion

Comparison appears to have different effects on similar and dissimilar objects. Comparing things that are similar can lead one to discover new (or highlight old) similarities, thereby increasing the perceived similarity of the two objects. Comparing things that are dissimilar on the other hand, is less likely to lead one to discover similarities (since there are fewer similarities there to be discovered). Hence, comparing two dissimilar things may serve to make the items less similar.

The results of Experiment 2 suggest that the increase in similarity following comparison in Experiment 1 was not simply due to subjects creating an extra feature (something akin to "thing I compared before") for items they were asked to compare. A co-history of comparison does not automatically result in higher similarity. Rather, it seems that only when meaningful similarities are to be found as a result of comparison, does comparison increase similarity.

However, there is one concern. At this point, it is not clear whether the difference observed between Experiments 1 and 2 is specifically a difference brought out by the process of comparison, or a more general difference inherent in the items. It could be that performing comparisons between items is simply serving to flesh out their representations, and it could be the differences inherent in these fleshed-out representations that produce the effects, and not the way comparison (per se) interacts with the representations. To explore this possibility, instead of asking subjects to perform comparisons between items, Experiment 3 asked subjects to list properties of the items separately (without comparing the items). This property-listing task was designed to flesh-out the representations without invoking the extra step of comparison. One group of subjects performed this task with the similar items used in Experiment 1, and another group of subjects performed the task with the dissimilar items used in Experiment 2.

Experiment 3

Method

Participants 234 Stanford University undergraduates participated in the study in order to fulfill a course requirement. Of these, 119 completed the task with similar items from Experiment 1 and 115 completed the task with dissimilar items from Experiment 2.

Materials Just as in Experiments 1 and 2, materials consisted of a one-page questionnaire. The top of the page contained either pictures of the 4 similar animals shown in Figure 1a or the 4 dissimilar objects shown in Figure 1b. Instead of being asked to name differences between two of the items, participants were asked to name properties of two of the items separately (e.g., "Please describe 3 properties of the phone." followed by 3 blank lines for participants to fill in and further followed by "Please describe 3 properties of the pretzel." again followed by 3 blank lines.) All of the counterbalancing and the rest of the page was done just as described for Experiment 1.

Procedures The procedures were the same as Experiment 1.

Results

Participants judged items to be more similar if they had previously been asked to name their properties than if they hadn't. This was true for both the similar items from Experiment 1 (M=6.34 after naming properties, and M=5.92 without naming properties, $t=1.90$, $p<.05$) and the dissimilar items from Experiment 2 (M=3.37 after naming properties, and M=2.97 without naming properties, $t=2.14$, $p<.02$). This pattern for the dissimilar items was significantly different from that observed in Experiment 2 as confirmed in an interaction in a 2x2 repeated measures ANOVA (2 (items were focused or not) X 2 (comparison or property-listing)), $F(1,174)=6.817$, $p=.01$.

Discussion

Unlike comparison, listing properties of individual items did not have a different effect on similar and dissimilar items. Whereas comparison served to increase the similarity only for similar items, property-listing increased similarity for both similar and dissimilar items. The process of comparison appears to have the special effect of selectively increasing the similarity of similar items (and possibly decreasing the similarity of dissimilar items). Simply fleshing out the representations (by listing properties) was not sufficient to have this effect.

It appears that the process of comparison could play a crucial role in the development of knowledge. However, the studies so far have only tested the effects of comparison on familiar items, things that people already have representations for. Can comparison play a similar role

even when people are just learning about something new? To investigate this, novel shapes were used in Experiment 4.

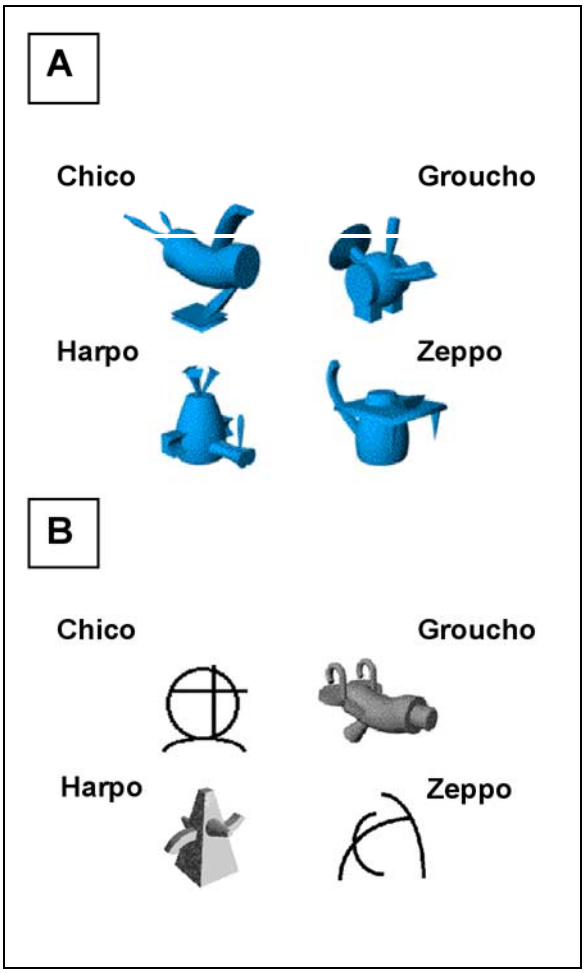


Figure 2: Stimuli used in Experiment 4. More similar items are shown in part A and less similar items are shown in part B. Images provided by Michael J. Tarr (Brown University) and Pepper Williams (University of Massachussetts).

Experiment 4

Method

Participants 188 Stanford University undergraduates participated in the study in order to fulfill a course requirement.

Materials Just as in Experiment 1, materials consisted of a one-page questionnaire. The top of the page contained color pictures of 4 named novel shapes. One set of questionnaires used the 4 similar objects shown in Figure 2a, and the other used the dissimilar objects shown in Figure 2b. The rest of the page was constructed just as in Experiment 1, with the following two differences: (1) all of the participants were asked to focus on differences (none named similarities) between the shapes, and (2) before being asked to verbally describe the differences, participants were asked to circle three differences between two of the novel shapes on the pictures themselves (e.g., "Please circle 3 differences between Chico and Harpo.") The rest of the page was constructed and counterbalanced just as for Experiment 1.

Procedures The procedures were the same as Experiment 1.

Results and Discussion

First, a manipulation check: participants indeed judged the "similar" items in Figure 2a to be more similar (M=3.84) than the "dissimilar" items in Figure 2b (M=2.37), $F(1,186)=36.6$, $p<.001$.

The effects of comparison were exactly as predicted by Experiments 1 and 2. Naming differences between two similar shapes (Figure 2a), again made people think of the two shapes as more similar (M=4.09 after naming differences, and M=3.58 without naming differences, $t=2.83$, $p<.01$). Naming differences between two dissimilar shapes (Figure 2b), on the other hand made people think of the two shapes as somewhat less similar (M=2.31 after naming differences, and M=2.43 without naming differences, $t=-.61$, $p=.27$). The patterns for similar and dissimilar items were significantly different from each other as confirmed in an interaction in a 2x2 repeated measures ANOVA (2 (named differences or not) X 2 (stimuli similar or dissimilar)), $F(1,186)=4.43$, $p<.05$.

It appears that the process of comparison had the same effect on novel items as it did on familiar items in Experiments 1 and 2. Comparing two similar novel items made them appear more similar, while comparing dissimilar novel items made them appear less similar.

General Discussion

The studies described in this paper examined the effects of comparison on perceptions of similarity. It appears that comparison can alter people's representations of objects by leading them to discover (or take note of) new similarities and differences. In future studies, it would be interesting to see how long effects of comparison last, and if these effects also extend to categorization. Previous research by Gentner and Namy (2000) suggests that this may indeed be the case. Further studies looking directly at the effects of comparison on categorization would be an interesting extension of this research.

Also worthy of further investigation are the interactions between similarity, structural alignability, and the process of comparison. In the studies reported in this paper, comparison was found to have different effects on similar versus dissimilar items (making similar items more similar, and dissimilar items less similar). However, the similar items used in these experiments were similar in several different ways: for example, both in terms of surface features and in deeper structural ways. Since several kinds of similarity were confounded, it is not clear which of these aspects contributed to the effect. In future studies it would

be interesting to investigate the separate contributions of structural and surface similarity as they interact with the comparison process. These further studies should also shed more light on why comparison has the effect it does.

Conclusions

Four studies showed that comparing similar objects makes them appear more similar, while comparing dissimilar objects makes them appear less similar. This was true for both novel and familiar objects. The effect of comparison on similar items was especially striking since subjects judged items to be more similar after comparison even if the comparison task was to list differences between the two items. Further, this effect appears specific to comparison and does not appear to be simply due to a "fleshing out" of object representations. When subjects were only asked to list properties of objects without comparing the objects themselves, the perceived similarity of the objects increased regardless of whether the items were similar or dissimilar to start. By making similar things appear more similar, and dissimilar things appear less similar, comparison may play a special role in category development. Further, it appears that even incidental conceptual experience (e.g., happening onto one comparison versus another) can play an important role in knowledge development.

These results suggest that common cognitive processes like comparison can introduce systematic biases into our representations of objects and their similarities. These biases may be beneficial for separating out bits of experience into categories, sharpening categorical boundaries, and otherwise helping us create conceptual structure above and beyond that offered by the world.

Acknowledgments

The author would like to thank Michael Ramscar, Herbert H. Clark, Dedre Gentner, the members of SLUGS, and the citizens of Cognation for many helpful discussions of this research, and Davie Yoon for invaluable help with data collection and coding.

References

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. Psychological Science, 5(3), 152-158.

Gentner, D., & Namy, L. (2000). Comparison in the development of categories. Cognitive Development, 14(4), 487-513.

Goldstone, R., Lippa, Y., & Shiffrin, R. (2001). Altering object representations through category learning. Cognition, 78 (1), 27-43.

Kurtz, K. (1998). The influence of category learning on similarity. Doctoral Dissertation, Stanford University.

Kurtz, K.J. & Gentner, D. (1998). Category learning and comparison in the evolution of similarity structure. Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, 1236.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. Memory & Cognition, 24(2), 235-249.

Markman, A. B., & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. Journal of Memory and Language, 32, 517-535.

Markman, A. B., & Gentner, D. (1993b). Structural alignment during similarity comparisons. Cognitive Psychology, 25, 431-467.

What is Universal in Event Perception? Comparing English & Indonesian Speakers

Lera Boroditsky (lera@mit.edu)
NE20-456, MIT, 77 Mass Ave
Cambridge, MA 02139 USA

Wendy Ham (wendyham@mit.edu)
NE20-456, MIT, 77 Mass Ave
Cambridge, MA 02139 USA

Michael Ramscar (michael@dai.ed.ac.uk)
University of Edinburgh, 2 Buccleuch Place
Edinburgh, EH8 9LW Scotland

Abstract

Does the language you speak shape the way you think about the world? Four studies investigate how English and Indonesian speakers encode and represent action events. Unlike English, Indonesian verbs do not include tense markers. Indonesian speakers are not required to indicate whether an event has already occurred, is happening now, or will occur in the future. Does needing to include tense to speak English grammatically change the way English speakers pay attention to, encode and remember events? We find cross-linguistic differences in memory and similarity judgments between English and Indonesian speakers, as well as between Indonesian-English bilinguals tested in English and Indonesian.

Introduction

Humans communicate with one another using a dazzling array of languages, and each language differs from the next in innumerable ways (from obvious differences in pronunciation and vocabulary to more subtle differences in grammar). For example, to say that "the elephant ate the peanuts" in English, we must include tense - the fact that the event happened in the past. In Mandarin and Indonesian, indicating when the event occurred would be optional and couldn't be included in the verb. In Russian, the verb would need to include tense and also whether the peanut-eater was male or female (though only in the past tense), and whether said peanut-eater ate all of the peanuts or just a portion of them. In Turkish, one would specify (as a suffix on the verb) whether the eating of the peanuts was witnessed or if it was hearsay. Speakers of different languages have to attend to and encode strikingly different aspects of the world in order to use their language properly (Sapir, 1921; Slobin, 1996). Do these quirks of languages affect the way their speakers think about the world? Do English, Mandarin, Russian, and Turkish speakers end up attending to, partitioning, and remembering their experiences differently simply because they speak different languages?

The relationship between language and thought is one of the most central questions in Cognitive Science. The universality of mental representations (whether or not speakers of different languages think differently about the world) has long been at the center of a controversy attracting thinkers from Plato to Chomsky, but despite much attention and debate, definitive answers have not been forthcoming. The idea that thought is shaped by language is most commonly associated with the writings of Benjamin Lee Whorf (Whorf, 1956). Whorf, impressed by linguistic diversity, proposed that the categories and distinctions of each language enshrine a way of perceiving, analyzing, and acting in the world. In so far as languages differ, their speakers too should differ in how they perceive and act in otherwise objectively similar situations. This strong Whorfian view—that thought and action are entirely determined by language—has long been abandoned in the field. However, definitively answering less deterministic versions of the "does language shape thought" question has proven a very difficult task. Some studies have claimed evidence to the affirmative (e.g., Boroditsky, 2001; Bowerman, 1996; Davidoff, Davies, & Roberson, 1999; Gentner & Imai, 1997; Levinson, 1996; Lucy, 1992; Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; Hermer-Vasquez, Spelke, & Katsnelson, 1999; Spelke & Tsivkin, 2001), while others report evidence to the contrary (e.g., Heider, 1972; Malt, Sloman, Gennari, Shi, & Wang, 1999; Li & Gleitman, in press).

One possible resolution to this debate might be that some conceptual domains are more susceptible to linguistic influence than others. For example, Gentner and Boroditsky (2001) have argued that the effect of language should be most apparent in the conceptualization of relations (typically encoded by verbs and spatial prepositions). In general, the lexicalization of actions and relations varies much more cross-linguistically than does the lexicalization of object categories, and picking out the extent and generality of a relational concept requires considerable experience with language. Recent research has supported this view (Gillette, Gleitman, Gleitman, & Lederer, 1999). For example, in one study, adults watched silent films of mothers talking to their children and tried to guess what was being said. Given only the silent film, adult participants were able to correctly guess nouns three times more often

than verbs (45% and 15% correct respectively). Further, concrete activity verbs like 'push' were much more easily guessed from silent observation than from the syntactic frames in which they were used (50% and 15% respectively), whereas verbs that denote more abstract activities like 'think' were much more easily guessed from syntax than from observation (90% and 0% respectively). It appears that acquiring representations of actions, relations, and events requires experience with language. This suggests that the eventual form of these concepts may be importantly shaped by the language experience.

This paper examines a cross-linguistic difference in verb syntax between Indonesian and English, and its effects on people's representations of action events.

Unlike English, Indonesian verbs do not include tense (they do not indicate whether the event or action took place in the past, is taking place in the present or will take place in the future). While Indonesian speakers may use other temporal words (e.g., just now, or soon) to communicate this information, these temporal markers are optional, and the tense of an action is often left to be inferred from context.

For example, in order to describe the three pictures shown in Figure 1, an English speaker might say (from left to right) "John is about to kick the ball," "John is kicking the ball," and "John has kicked the ball." In Indonesian all three pictures would likely be described by the same sentence, roughly "John kick ball." Does this difference between how English and Indonesian speakers talk about action events lead to differences in how the two groups attend to, encode, and represent the events?

We report four studies aimed at uncovering differences and similarities between Indonesian and English speakers in terms of how they encode and represent actions and events.

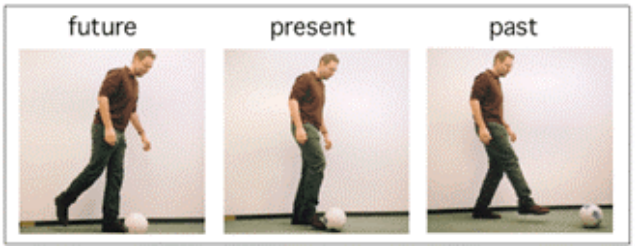


Figure 1: Example of action in three different tenses.

Experiment 1

Experiment 1 examined whether speaking a tensed language makes English speakers think of actions in different tenses as less similar, and actions in the same tense as more similar. English and monolingual Indonesian speakers were shown pairs of pictures that show either two different actors performing the same action in the same tense, or the same actor performing the same action in two different tenses (as shown in Figure 2). Subjects were asked to rate the similarity of a large set of such pairs (on a 9 point scale where 1=not similar and 9=very similar). The linguistic difference between the two languages predicts that English speakers will rate same-tense pairs more similar than will Indonesian speakers, but will rate different-tense pairs less similar than will Indonesian speakers.

Methods

Participants 14 native English and 12 monolingual Indonesian speakers participated in this study in exchange for payment. The English speakers were recruited and tested at MIT, and the Indonesian speakers were tested in Jakarta. None of the Indonesian speakers had learned English.

Materials A set of 90 pictures served as stimuli for this experiment. The pictures portrayed 10 different actions, each action performed by three different actors. For each actor performing a particular action there were 3 versions showing the actor about to perform the action, doing the action, and having done the action (as shown in Figure 1). The actions were: kicking a ball, throwing a frisbee, eating a banana, drinking orange juice, ripping a sheet of paper, cutting a rope, hula-hooping, lifting a very large ball, pouring dark liquid out of a clear container, and opening an umbrella. The actions were chosen to be sufficiently different from one another to ensure generality.

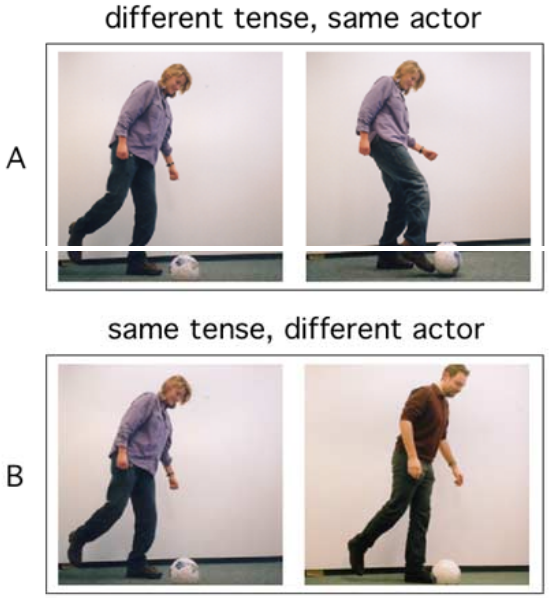


Figure 2: Examples of picture-pairs used.

Design A set of 180 picture pairs were created. Half of these showed the same actor performing the same action in different tenses (as shown in Figure 2A), and half showed two different actors performing the same action in the same tense (as shown in Figure 2B). All possible combinations of pictures that yielded such pairs were used (a total of 180 pairs).

Procedure Participants were tested individually in a quiet room. English speakers were given instructions in English, and Indonesian speakers were given instructions in Indonesian. The English instructions were: "You will see two pictures each time. Your task is to rate how similar those two pictures are. In rating them, use the numbers 1 through 9: 1 for NOT similar at all, 9 for VERY similar." The Indonesian instructions were: "Setiap kali, Anda akan melihat dua gambar. Kami minta agar Anda mengatakan seberapa mirip dua gambar itu. Gunakan angka 1 sampai dengan 9. 1 untuk SANGAT TIDAK MIRIP, 9 untuk MIRIP SEKALI."

The participants were also told that all of the pairs would be pretty similar, but they should still try to use the whole range of similarity ratings from 1 to 9.

A computer presented the 180 pairs in a new random order for each subject. Each pair stayed on the computer screen until the subject pressed a key (1 through 9) to indicate their similarity rating.

Results

Results are shown in Figure 3. As predicted, English speakers rated same-tense pictures (involving different actors) more similar than did Indonesian speakers (M=6.34 for English speakers, and M=4.82 for Indonesian speakers), $t=2.07$, $df=24$, $p<.05$. Further, as predicted, English speakers rated different-tense pictures less similar than did Indonesian speakers (M=5.56 for English speakers, and M=6.64 for Indonesian speakers), $t=1.71$, $df=24$, $p<.05$. The difference between the two language groups was confirmed as an interaction between comparison type (between-tense or within-tense) and language (English or Indonesian) in a 2x2 repeated measures ANOVA, $F(1,24)=4.41$, $p<.05$. This pattern of findings suggests that using the English tense system may change English speakers' representations of actions, making differently tensed actions appear more distinct and actions in the same tense appear more similar.

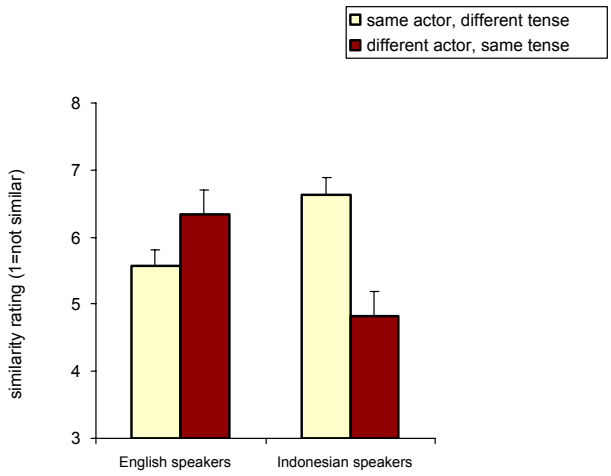


Figure 3: Results of Experiment 1.

Discussion

English and Indonesian speakers appeared to focus on different aspects of action scenes in making their similarity comparisons. English speakers seemed to hone in on tense, judging actions in the same tense but performed by different actors (as shown in Figure 2b) to be more similar than actions performed in different tenses but by the same actor (as shown in Figure 2a). The Indonesian speakers showed the opposite pattern, appearing to ignore similarity of tense in favor of similarity of actor.

This raises two further questions: (1) Do Indonesian speakers who learn the English tense system change the way they think about events? and (2) Do Indonesian-English bilinguals think differently when speaking Indonesian than when speaking English? Experiment 2 tested Indonesian-English bilinguals both in Indonesian and in English on the same task as described in Experiment 1.

Experiment 2

Methods

All of the materials, design and procedure were exactly as described for Experiment 1. Seventeen Indonesian-English bilinguals participated in this study. Seven were tested in English and ten were tested in Indonesian. All of the participants were native speakers of Indonesian and were matched on their amount of experience with English.

Results and Discussion

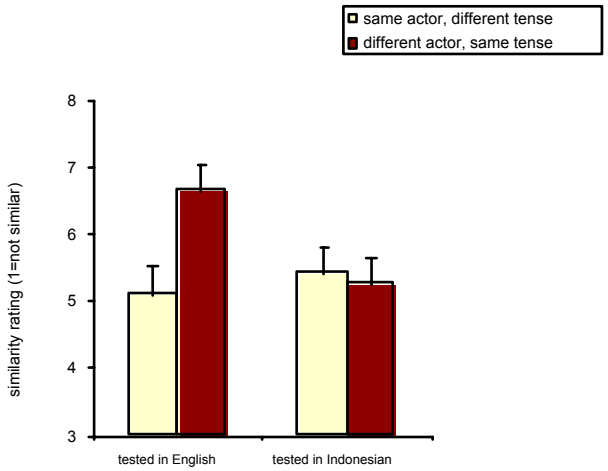


Figure 4: Results of Experiment 2.

Results are shown in Figure 4. Indonesian-English bilinguals rated same-tense pictures (involving different actors) more similar when they were tested in English than when tested in Indonesian, (M=6.64 when tested in English, and M=5.27 when tested in Indonesian). Further, Indonesian-English bilinguals rated different-tense pictures less similar when they were tested in English than when

tested in Indonesian (M=5.11 when tested in English, and M=5.42 when tested in Indonesian). The difference between the two groups was confirmed as an interaction between comparison type (between-tense or within-tense) and language of testing (English or Indonesian) in a 2x2 repeated measures ANOVA, $F(1,15)=4.40$, $p<.05$.

This pattern of findings suggests two things. First, it appears that bilinguals do think differently when speaking different languages. Even though the task was conducted in pictures, setting a linguistic context by providing instructions either in English or in Indonesian changed the way Indonesian-English bilinguals reasoned about the action events in this study.

Second, it appears that learning a new language can change the way one thinks. The Indonesian-English bilinguals who were tested in Indonesian showed a pattern that was somewhere in-between the pattern shown by monolingual Indonesian speakers and the pattern shown by English speakers. Even though they were tested entirely in Indonesian it appears that having learned English may have changed the way they think about action events. Further studies will be necessary to explore this possibility in more detail.

Experiment 3

Although the findings of Experiments 1 & 2 are very suggestive, the similarity-ratings task used is subjective and may tell us more about the participants' cognitive preferences than about their cognitive performance. Could cross-linguistic differences lead to difference in cognitive performance and not just preference?

Experiment 3 was designed to test Indonesian and English speakers' ability to remember action events. Subjects were shown pictures of people performing actions (the same pictures were used as in Experiments 1 & 2). During the learning phase, each subject saw a person performing an action in one of three tenses (e.g., they may have seen only the middle panel of Figure 1). During the test phase, subjects were shown pictures of that person performing the action in all three tenses (as shown in Figure 1) and asked to choose which one they had seen previously. We predicted that English speakers should be better than Indonesian speakers at encoding and remembering the tense in which they witnessed an action.

Methods

Participants 13 native English and 18 monolingual Indonesian speakers participated in this study in exchange for payment. The English speakers were recruited and tested at MIT, and the Indonesian speakers were tested in Jakarta. None of the Indonesian speakers had learned English.

Materials and Design All of the same pictures as described for Experiment 1 were used. During the learning phase, subjects were shown 30 of the 90 pictures (1 picture of each person doing each action in only one of the possible 3 tenses). At test, subjects were shown all three pictures of a person performing an action in all 3 tenses (all three pictures were presented simultaneously) and asked to choose which one they had seen earlier.

Procedure Participants were tested individually in a quiet room. English speakers were given instructions in English, and Indonesian speakers were given instructions in Indonesian. Participants were told to simply look at the pictures and try to remember everything they saw. Participants were not instructed to encode the pictures linguistically.

A computer presented the pictures in a new random order for each subject (both for the learning and test sets). During the learning, each picture was shown only once and stayed on the screen for 3 seconds. During the test, the pictures stayed on the screen until the subject made a response (by pressing 1, 2, or 3 on the keyboard to correspond to which picture they thought they had seen previously).

Results

Results are shown in Figure 5. As predicted, English speakers were better able to remember which of the three tense versions of a picture they had seen before. English speakers were able to pick the correct answer 41% of the time, as compared to Indonesian speakers who only succeeded 31% of the time, $t=1.72$, $df=29$, $p<.05$. Indonesian speakers were not better than chance at recognizing the picture they had seen before (chance=33.3%).

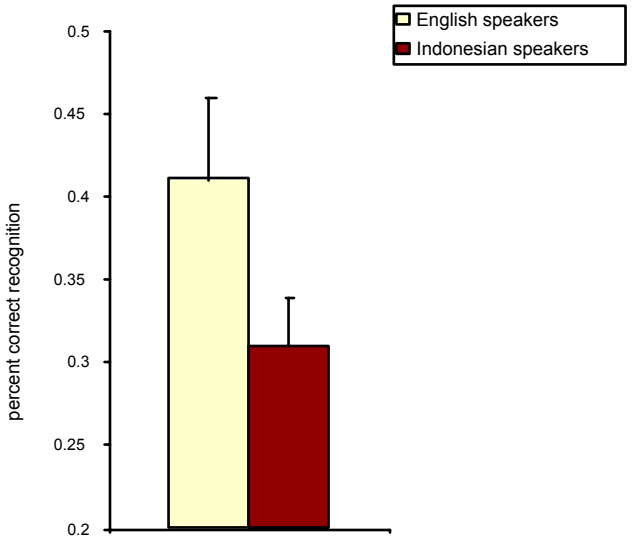


Figure 5: Results of Experiment 3.

Discussion

Results of Experiment 3 suggest that experience with particular languages may affect not only people's cognitive preferences, but also aspects of their performance on basic cognitive tasks (such as memory). However, it is possible that differences in memory performance between English and Indonesian speakers observed in this task are not due to linguistic differences, but rather to other differences in cultural upbringing or education. It is quite possible that the English speakers included in this sample (mostly MIT undergraduates) have received more training in memorization than the Indonesian speakers in our sample. In order to control for such differences, Experiment 4 tested Indonesian-English bilinguals (who were matched on educational and cultural background) either in English or in Indonesian in the same task as described for Experiment 3. By keeping constant educational and linguistic background and only varying the language of testing we can test whether speaking one language versus another can really affect aspects of one's cognitive performance.

Experiment 4

In Experiment 4, Indonesian-English bilinguals were tested either in Indonesian or in English in the same memory task as described for Experiment 3.

Methods

All of the materials, design and procedure were exactly as described for Experiment 3. Eighteen Indonesian-English bilinguals participated in this study. Seven were tested in English and eleven were tested in Indonesian. All of the participants were native speakers of Indonesian and were matched on their linguistic and educational background.

Results and Discussion

Results are shown in Figure 6. As predicted Indonesian-English bilinguals were better able to remember the tense of actions when they were tested in English (40% correct) than when they were tested in Indonesian (26% correct), $t=1.76$, $df=16$, $p<.05$. When they were tested in Indonesian, Indonesian-English bilinguals did not perform better than chance (in fact, slightly worse).

These results suggest that even something as subtle as linguistic context (whether instructions were given in English or Indonesian) can have an effect on how people encode and represent events. Even though subjects were not asked to encode the events linguistically (and the entire task was conducted in pictures), people's ability to remember the tense of events they had witnessed (whether they saw someone about to kick a ball or having already kicked a ball) depended on whether or not tense distinctions were required in the language in which they had been greeted and given instructions just prior to the task.

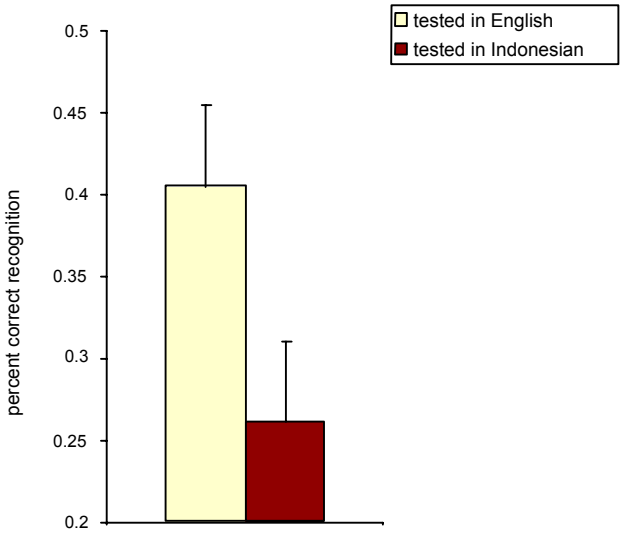


Figure 6: Results of Experiment 4.

Summary

Four studies investigated the effect of linguistic experience on how people attend to, encode, and represent action events. The results suggested that experience with the English tense system makes English speakers think of actions in the same tenses as more similar and action in different tenses as less similar. Unlike English speakers, Indonesian speakers did not appear to value tense in making their similarity judgments. Further, English speakers were better able to remember the tense of an action they had witnessed than were Indonesian speakers (who did not perform above chance on the memory task).

Studies that tested Indonesian-English bilinguals in either English or Indonesian provided further insights about how language may affect thinking. Both in memory and similarity ratings, the language that bilinguals were tested in had an effect on the subjects' performance in the task. Indonesian-English bilinguals looked just like English speakers when tested in English, and much more like Indonesian speakers when tested in Indonesian. Further, results of the similarity study suggested that learning a new language can change the way one thinks – Indonesian-English bilinguals tested in Indonesian showed a pattern of results that was somewhere in-between the English-speakers' pattern and that shown by monolingual Indonesian speakers.

Overall, it appears that representations of action events are not universal. Experience with the English tense system appears to segment actions into distinct temporal categories that are not basic or universal to human cognition. Further, even something as subtle as linguistic context (here, the language in which instructions are given for a non-linguistic task) appears to have a striking effect on how people encode and represent their experiences.

It appears that speakers of different languages do attend to, partition, and remember their experiences differently, simply due to the implementational differences of the languages they speak.

Acknowledgments

The authors would like to thank Tracy Alloway for discussion of this work and Shijun Xi and Mindy Chang for help with data collection. The authors would also like to thank Webb Phillips, Daniel Casasanto and the other citizens of Cognation for discussion and help with assembling the stimuli.

References

Boroditsky, L. (2001). Does Language Shape Thought? Mandarin and English speakers conceptions of time. Cognitive Psychology, 43(1), 1-22.

Bowerman, M. (1996). The origins of children's spatial semantic categories: cognitive versus linguistic determinants. In J. Gumperz & S. Levinson (Eds.), Rethinking linguistic relativity. Cambridge, MA: Cambridge University Press, 145-176.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. Nature, 398, 203-204.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. Science, 284, 970-974.

Gentner, D., & Boroditsky, L. (2001). Individuation, relational relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), Language acquisition and conceptual development. Cambridge, England: Cambridge University Press.

Gentner, D. & Imai, M. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. Cognition 62, 2, 169-200.

Gillette, J., Gleitman, H., Gleitman, L., Lederer, A. (1999). Human simulations of vocabulary learning. Cognition, 73(2), 135-176.

Heider, E. (1972). Universals in color naming and memory. Journal of Experimental Psychology, 93, 10-20.

Hermer-Vasquez, L., Spelke, E. S. & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. Cognitive Psychology, 39, 3-36.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? American Anthropologist, 86, 65-79.

Levinson, S. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom & M. Peterson (Eds.), Language and Space. Cambridge, MA: MIT Press, 109-169.

Li, P. & Gleitman, L. (in press). Turning the tables. Cognition.

Lucy, J. (1992). Grammatical categories and cognition: a case study of the linguistic relativity hypothesis. Cambridge, England: Cambridge University Press.

Lucy, J., & Shweder, R. (1979). Whorf and his critics: Linguistic and nonlinguistic influences on color memory. American Anthropologist, 81, 581-618.

Malt, B., Sloman, S., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40, 230-262.

Sapir, E. (1921). Language, New York, NY: Harcourt, Brace, and World.

Slobin, D. (1996). From "thought and language" to "thinking for speaking." In J. Gumperz & S. Levinson (Eds.), Rethinking linguistic relativity. Cambridge, MA: Cambridge University Press, 70-96.

Spelke, E., & Tsivkin, S. (2001). Language and number: A bilingual training study. Cognition, 78(1), 45-88.

Whorf, B. (1956). Language, Thought, and Reality: selected writings of Benjamin Lee Whorf, ed. J.B. Carroll. Cambridge, MA: MIT Press.

Atomistic and Systems Approaches to Consciousness

Andrew Brook (abrook@ccs.carleton.ca)

Luke Jerzykiewicz (ljerzyki@ccs.carleton.ca)

Cognitive Science, Carleton University, Ottawa, Ontario, Canada

Abstract

The approach to consciousness taken by most philosophers is very different from the approach taken by most cognitive psychologists, so different that one could be forgiven for wondering if they are talking about the same thing. Most philosophers focus on individual psychological states. By contrast, most psychologists focus on properties of cognitive systems as a whole such as global workspace or attention. (Some philosophers favour this approach, too, Dennett and P. M. Churchland for example.) We will expose some of the peculiarities of the dominant philosophical approach and, by looking briefly into what is needed to give an adequate account of consciousness, advance some reasons for favouring the approach dominant among psychologists.

Two Approaches to Consciousness

Reading recent writings on consciousness by philosophers and cognitive psychologists, one could be forgiven for wondering if they are talking about the same thing. Most philosophers focus on individual psychological states – individual perceptions or feelings or imaginings – or at most tiny combinations of such states (Rosenthal, 1991; Chalmers, 1996; Tye 1995). Experimental psychologists by contrast focus on properties of cognitive systems as a whole: global workspace (Baars 1988), intermediate level of processing (Jackendoff 1987), or attention. Attention has been particularly singled out for ... attention. For Posner or Mack and Rock, for example, to be conscious of something simply is to pay attention to it (Posner 1994; Mack and Rock 1998).

First we will lay out the main contours of the two dominant approaches, starting with the one favoured by most philosophers (though not all: both Dennett (1991) and P. M. Churchland (1995) favour the approach taken by psychologists). Then we will assess them. There are serious *prima facie* shortcomings in the dominant philosophical approach, but there is also a serious worry about the dominant psychological approach. To bring out these shortcomings and this worry, we will need to distinguish two kinds of consciousness and lay out some of the things that an adequate account of consciousness must explain.

The dominant approach in philosophy

Two things characterize the dominant approach to consciousness in recent philosophical work on the topic. The first is a kind of atomism. These philosophers tend to talk about conscious states one by one ('what is it like for something to look red?') or at most in tiny groups. In both cases, the cognitive system that has them is pretty much ignored. (Theorists may add, '...look red *to me'* but they do nothing with the addition.) Almost the whole of the massive literature on *qualia* exhibits this feature. (*Qualia* are the felt quality of conscious states, 'what it is like to have them', in Thomas Nagel's (1974) famous phrase.) The thing that has such states, the subject of conscious states, is ignored. Let us call this view of consciousness *atomism.*

> **Atomism** – the view that conscious states can be studied one by one or in small groups, without reference to the cognitive system that has them.

When you think about it, this atomism is remarkable. It seems obvious that consciousness does *not* come in atomically separable states in this way.

This atomism about consciousness goes with another view that we will call *local realism.* Local realism is the view that consciousness or what is distinctive about consciousness, for example that in virtue of which it is like something to have a psychological state, is a property of individual psychological states or tiny groups of psychological states.

> **Local realism** – the view that consciousness or what is distinctive about consciousness is either a non-relational property of individual psychological states or a relationship among very small numbers of psychological states.

Specifically, this approach to consciousness views it as either a nonrelational property of single psychological states or, though a relational property, one that ties only very small groups of psychological states to one another. A relationship between one state and another single state would be an example. This account is not very precise but it is precise enough for our purposes. What matters is the contrast with the kind of properties that figure centrally in theories that view consciousness as a relationship between a great many psychological states and a conscious being

whose states they are.

Local realism is a realist view because it takes the states in question to be real, not postulated abstract entities that we believe in merely because of certain concepts or explanatory strategies or something of the kind.

There are (at least) three types of local realism. In one type, appearing red to me would be a property of an experience of red, being painful would be a property of a pain, and so on. We find a second type in what have come to be called transparency theories, theories holding that we see right through conscious states (hence transparency) and are conscious only of what such states are about. In this approach, something appearing red is not a property of any experience, it is a matter of experiencing something that appears to be red, feeling pain is a matter of experiencing something painful, and so on. In the third type, a representation of red gets to be conscious by being related to another psychological state, for example by being the object of a thought about that representation (Rosenthal's 1991 higher-order thought view of consciousness).

It might seem that atomism requires local realism but in fact that is not clear. Some atomists about consciousness are simply neutral about whether qualia, for example, are local or nonlocal properties of the states they discuss. This neutrality is a bit curious because these theorists believe that they can say other important things about qualia, e.g., that when it is like something to have a representation, this quale, this being like something, is radically different from other aspects of the representation, but neutral they have been. Nonetheless, local realism would certainly promote atomism: if consciousness is a local property of certain states, it would be at least very tempting to hold that one could study such states one by one and in isolation from the system that has them.

It is important to note that *local* realism about consciousness is not necessarily the same thing as *realism* about consciousness. Even if consciousness is not a local property of individual psychological states, it could still be a real property of cognitive systems as a whole. We mention this now because there have been influential treatments of consciousness recently that back off from any form of cognitive-system realism about consciousness, for example Davidson's (1996) view. In Davidson's view, consciousness arises out of a complex triangular interaction among oneself, other purposive beings, and the world. By itself, this triangulation picture need not depart from realism; the result of the triangulation, consciousness, could still be a real property of cognitive systems. For Davidson, however, not only does consciousness arise out of triangulation, it is (roughly) nothing more than triangulation. When triangulation results in stable attributions of consciousness to self and others, that is what consciousness is. And *this* view *is* incompatible with most versions of realism about consciousness.

One central issue in this atomist, local realist literature is the relationship of consciousness to representation. At minimum, being conscious of something is *one way* of representing something. Of course, things can also be represented unconsciously. In fact, probably the vast bulk of our representations never make us conscious of anything. Certainly a representation does not need to make us conscious of anything to be cognitively active. But now ask: can the difference between conscious and nonconscious representation *be captured* by appealing to representational properties or are the properties that make a state conscious nonrepresentational properties? Here there is a major split in the atomist, local realist camp, with some researchers opting for the representational alternative (Lycan 1987, Tye 1995), others insisting on the nonrepresentational one (Block 1995, Chalmers 1996). For the anti-representationalists, the difference between a state that is conscious and one that is not is *not* a difference in how that state represents anything or a difference in the kind of representation it is or a difference in anything else representational. Since we will eventually come to raise some doubts about anti-representationalism, let us flag the view explicitly:

> **Anti-representationalism** – the view that the difference between a state that is conscious and one that is not is not a difference in how that state represents anything or a difference in the kind of representation it is or a difference in anything else representational.

Here is how the anti-representationalist view can arise.

When something appears to us to be a certain way, the representation in which it appears that way can play two roles in our cognitive economy. On the one hand, the representation (or the contents of the representation) can connect inferentially to other representations: if the stick appears to have two straight parts with a bend in the middle, this will preclude representing it as forming a circle. The representation can also connect to belief: if the stick appears straight with a bend in it, we will not form a belief that it bends in a circle. And to memory: we can compare this stick as it appears to sticks I recall from the past. And action: if I want something to poke into a hole, I might reach for the stick. In all these case, so long as I am *representing* the stick in the appropriate way, it would seem to be irrelevant whether I am *conscious* of the stick or not. My representation could do these jobs for me just as well even if I were not aware either of the stick or of my representation of it. But I am also *conscious* of the stick – it does *appear* to me in a certain way. This can easily seem to be something different from any representational properties of the representation, at any rate properties such as those we just considered.[1]

[1.] Chalmers' well-known (1995) distinction between what he calls the easy problem and the hard problem of consciousness starts from this distinction between the cognitive role of

Arguments for this conclusion often take the following form: the felt quality of a state could change while its representational properties remain the same. The arguments are usually based on thought experiments such as the inverted spectrum argument (how colours appear to us could be inverted without changing how our representations of colour function as representations) or the zombie argument (there could be creatures for whom it is not like anything to represent anything whose representations nevertheless function cognitively just as representations function in us).

Sometimes such arguments go so far as to conclude that what is distinctive to consciousness is not just not representational, it is not even physical. One way of arguing for this to make one's zombie a microphysical duplicate of conscious beings. If a zombie such as this is possible, then qualia are not a physical property of conscious beings. Another is Jackson's (1986) famous thought experiment concerning Mary, the colour scientist. Mary knows everything there is to know about the experience of colour, therefore everything *physical* there is to know about the experience of colour, but she has never experienced colour herself. Then she experiences colour. Clearly she gains something she did not have before. However, she knew everything physical about colour. Therefore, what she gains must be something nonphysical.

It is not clear that any of these thought experiments establish real possibilities, or, if they do, entail the conclusions drawn from them. For reasons of length, in this paper we will pass on that issue. Instead, we will turn to alternative approach to consciousness introduced earlier, the one found more often in the work of experimental psychologists.

The dominant approach in psychology

In sharp contrast to atomism and local realism (whether in its representational or its anti-representational form), the dominant approach to consciousness in experimental psychology holds that consciousness is a property of the cognitive system as a whole. Let us call it the *system approach to consciousness:*

> **System approach to consciousness** – approaches to consciousness that view it as a property of whole cognitive systems, not individual or small groups of representations.

There is a great diversity of opinion as to what the relevant property is. We cannot begin to explore the whole range of options but here are a few examples. Baars (1988) holds that consciousness consists in a global workspace of a certain kind. Jackendoff (1987) urges that it is an intermediate level of representation, a phonetic or similar level between acoustic or visual input and full-blown conceptual content. Many theorists link consciousness very closely to attention. For example, Mack and Rock say that, "Attention [is] the process that brings a stimulus to consciousness" (Mack and Rock 1998), "if a ... percept captures attention, it then becomes an explicit percept, that is, a conscious percept" (Mack 2001, 2). Posner (1994) captures the spirit of this line of thinking about consciousness nicely:

> an understanding of consciousness must rest on an appreciation of the brain networks that subserve attention, in much the same way as a scientific analysis of life without consideration of DNA would seem vacuous. [Posner 1994, 7398]

Nor is this approach without its philosophical allies. Dennett's (1991) multiple drafts model in which states become conscious by seizing control of cognitive resources is a similar approach. (Curiously, he says almost nothing about attention.) Paul Churchland is another example. Here is how Churchland summarized his approach recently:

> [Consider] the brain's capacity to focus attention on some aspect or subset of its teeming polymodal sensory inputs, to try out different conceptual interpretations of that selected subset, to hold the results of that selective/interpretive activity in short-term memory for long enough to update a coherent representational 'narrative' of the world-unfolding-in-time, a narrative thus fit for possible selection and imprinting in long-term memory. Any [such] representation is ... a presumptive instance of the class of *conscious* representations. [Churchland 2002, 74]

This is a different conception of consciousness from the atomist one! Are there reasons to favour one over the other? There are. First reason: the systems approach has the potential to account for two kinds of consciousness, the atomist approach only one.

Two kinds of consciousness

The variety of different things that we can have in mind when we use the word 'consciousness' is a big topic, too big a topic to explore here. However, we can make one distinction fairly briefly and when we do, something quite interesting about the two approaches becomes apparent.

The distinction we have in mind is between consciousness of the world around us and consciousness of our own psychological states. Blindsight is sometimes invoked to illustrate this distinction but what is tendentiously called 'inattentional blindness' works better. (Tendentiously because there is actually a huge debate about whether the phenomenon in question has anything to do with attention [Mack, http://psyche.cs.monash.edu.au/v7/psyche-7-16-

representations and something appearing to be like something in them.

mack.htm].) In inattentional blindness research, the subject fixates on a point and is asked to note some feature of an object introduced at or within a few degrees of fixation. After a few trials, a second object is introduced, in the same region but clearly distinct from the first object. Subjects are not told that a second object will appear. When the appearance of the two objects is followed by 1.5 seconds of masking, at least one-quarter of the subjects and sometimes almost all subjects have no awareness of having seen the second object.

Yet – and this is what makes inattentional blindness better for our purposes than blindsight – when the second object is a word, subjects clearly encode it and process its meaning. Evidence? When asked shortly after to do, for example, a 'stem completion task' (i.e., to complete a word of which they are given the first two or three letters), they complete the word in line with the word they claim not to have seen much more frequently than controls do. In short, in inattentional blindness, subjects' access to the word they are not aware of seeing is nevertheless very deep-running.

In inattentional blindness, it is important to note, objects *appear* in a certain way to the subject, as is evidenced by the subject processing semantic information provided by it.[2] What we have here is not merely Block's A-consciousness, "a state ... poised for direct control of thought and action" (Block 1995, 233). The access to the unattended object is Block's -consciousness or something very much like it: the object actually appears to the subject. (Note that if this claim is correct, it poses a considerable problem for attention theories of consciousness – something else we don't have space to go into.) In these or similar cases of access without attention, subjects can, for example, point to the items in question. The objects can increase the subject's level of alertness, especially the level of alertness concerning the objects themselves. And ensuing behaviour is often appropriate, not to the way the object actually is, but to how the objects looked to the subject (Dennett, 1978). Let us call the kind of consciousness that can be present in cases of inattentional information access and so on *consciousness of the world.* By contrast, let us call the consciousness that we have when we are *conscious* of representing items in the world *consciousness of self.*[3]

Consciousness of the world – the kind of consciousness that can be present in cases of inattentional information access and so on

Consciousness of self – the consciousness that we have when we are conscious of representing items in the world

Now a reason for favouring the systems approach: all anti-representational versions of atomism and many representational versions (e.g., higher-order thought or experience models) have anything to say only about consciousness of self, the felt quality of psychological states, what it is like to have them, and cannot say anything about consciousness of the world, i.e., the way the world appears to someone. Systems approaches not only have something to say about consciousness of the world, they generally focus on it. When theorists talk about paying attention to something, for example, they generally have in mind paying attention to something in the world, not paying attention to one's own states.

What a theory of consciousness must explain

A second reason for favouring a systems over an atomistic approach to consciousness: what a theory of consciousness actually needs to be able to explain.

If consciousness is a matter of things appearing – just appearing, in the case of consciousness of the world, consciousness that they are appearing in the case of (one kind of) consciousness of self –, then consciousness is a property of the activity of representing, not of individual representations. Consciousness is a matter of *something being conscious* of something. If so, an adequate theory of consciousness has to address the question: What is *a system* capable of consciousness like?

Here are some of the features of such a system:

- It is aware of whole groups of representations in one 'act of consciousness'
- Often when it is aware of whole groups of representations, it is also aware of itself as the common subject of these representations.
- Its consciousness can be faint, full, etc.
- It has many global cognitive faculties and some of

---

2. Change blindness, attentional blink, and visual neglect and the double dissociation between the ventral and the dorsal streams in the brain discovered by Milner and Goodale (1995) are related phenomena. In all these phenomena, information that the subject is not conscious of having nevertheless enters into cognitive tasks that use semantic information , disambiguation tasks, for example.

3. Consciousness of self needs to be broken down into consciousness of one's psychological states and consciousness of oneself, the thing having those states. Moreover, there are radically different views afoot about what consciousness of one's psychological states consists in. Some theorists even maintain that it is nothing more than consciousness of the world plus a shift of attention (Dretske 1995, Tye 1995). We have to ignore all these issues here.

them are closely linked to consciousness, memory, for example, or attention, or language.

- In particular, attention is closely linked to consciousness.
- For consciousness, a system simply having information as a result of representing this, that or the other is not enough; the system must make cognitive use of the information.
- Consciousness in a cognitive system can be independent of sensory inputs
- Its consciousness disappears in deep sleep, and . . .
- reappears in dreams.[4]

When faced with issues like these, the atomistic approach to consciousness has so far just clawed the air – and it is hard to think of circumstances under which it could do any better.

Take, for example, the unity of consciousness. Conscious subjects are aware at the same time, indeed in a single act of consciousness, of a great many items. A theory of the conscious subjects has to be able account for this unity.

The unity of consciousness comes in a number of kinds. Mental unity in general comes in at least six different kinds and four of them are kinds of unified consciousness:

1. unity of our cognitive elements (we can bring, for example, beliefs, desires, perceptions, intentions, and many other things to bear on a single situation);

2. unified consciousness of our world (we are aware of a whole host of things around us in a single, unified representation) and

3.unified consciousness of one's own representations;

4. unified consciousness of self (one is aware of oneself as the "single, common subject" of one's experience, as Kant put it),

5. unified focus of a number of cognitive resources on a single item of attention;

and,

6. unified behaviour (our behaviour is highly and multiply unified – think of a concert pianist playing a sonata).

Set items (i) and (vi) aside. Items (ii) to (v), the various kinds of unified consciousness, have a common core:

Unity of consciousness – a group of representations being related to one another such that to be conscious of any of them is to be conscious of others of them and of the group of them as a single group.

Given how central unified consciousness is to the conscious mind, it is remarkable how little attention it is has received in recent writings on consciousness, especially philosophical writings. Paul Churchland (1995, p. 209) includes it as one of his Magnificent Seven, the things that a theory of consciousness has to explain, and the notion is mentioned by a few other philosophers but in general it has received little attention (the topic and what has been done with it is reviewed in Brook 2000).

That consciousness is unified has immediate implications for atomism and local realism. If consciousness is unified, consciousness cannot be a property of single representations or tiny groups of representations (e.g., a representation and a thought directed at it on the HOT model) by themselves. Nor it is something that could fruitfully be studied by studying single representations in isolation. At present, we don't think that there is a theory of consciousness, representational or nonrepresentational, that provides an adequate account of the fact that consciousness is unified. To pay attention to it is to see the prospects for atomism about consciousness immediately plummet.

A Problem for the Systems Approach?

If the systems approach to consciousness seems more likely than atomism to be able to explain what a theory of consciousness has to explain, it also faces some problems. In particular, many theorists worry that it may leave out just the most crucial element, the consciousness itself. This worry arises in the following way. 'Surely', an objector will say, 'a cognitive faculties and capacities central to your favourite systems approach to consciousness could exist, and not only exist but function as they do, in the absence of consciousness?' This line of objection is the home of zombie thought-experiments: surely something could be just like us behaviourally, or (as in this case) functionally, or even physically, and yet not be conscious. All we can say here is that *if* zombie thought-experiments are coherent, then the systems approach is in trouble, as is every other representational theory of consciousness. But that is a big 'if'. Since the price of buying the idea that zombie thought-experiments are coherent is that consciousness has to be something deeply mysterious, maybe beyond cognitive ken altogether, we want to make very sure that zombie thought-experiments *are* coherent.

Adjudicating that issue and the background issue of the merits of anti-representational atomism vs. the systems picture of consciousness is a task for another occasion. Here we have merely tried to introduce the two approaches, lay out one reason to favour the systems approach, and look briefly at one difficulty it faces.

---

[4] This list started from but goes beyond Churchland's list of the Magnificent Seven requirements on a theory of consciousness in (1995), pp. 213-14.

References

Baars, B. 1988. *A Cognitive Theory of Consciousness* Cambridge: Cambridge University Press

Brook, A. 2000. Unity of Consciousness. *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu

Chalmers, D. 1996. *The Conscious Mind* Oxford: Oxford University Press

Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul* MIT Press

Churchland, P. M. 2002. Catching consciousness in a recurrent net. In: Andrew Brook and Don Ross, eds. *Daniel Dennett* New York: Cambridge University Press

Davidson, D. 1996. Subjective, Intersubjective, Objective. In Paul Coates, ed. *Current Issues in Idealism.* Bristol, UK: Thoemmes.

Dennett, D. 1978. Toward a cognitive theory of consciousness. In his *Brainstorms* Bradford Books, 149-73

Dennett, D. 1991. *Consciousness Explained.* Boston: Little, Brown

Dretske, F. 1995. *Naturalizing the Mind*. MIT Press

Jackendoff, R. 1987. *Consciousness and the Computational Mind* MIT Press

Jackson, F. 1986. What Mary didn't know *Journal of Philosophy* 83:5, 291-5

Lycan, Wm. 1987. *Consciousness* Cambridge, MA: MIT Press

Mack, A. http://psyche.cs.monash.edu.au/v7/psyche-7-16-mack.htm.

Mack, A. and Rock, I. 1998. *Inattentional Blindness* Cambridge, MA: MIT Press

Nagel, T. 1974. What it is like to be a bat? *Philosophical Review* 83: 435-50

Posner, M. 1994. Attention: the mechanism of consciousness *Proceedings of the National Academy of Science USA* 91: 7398-7403

Rosenthal, D. 1991. *The Nature of Mind* Oxford: Oxford University Press

Tye, M. 1995. *Ten Problems of Consciousness.* Cambridge, MA: MIT Press

Reference Resolution in the Wild: On-line Circumscription of Referential Domains in a Natural, Interactive, Problem-solving Task.

Sarah Brown-Schmidt (sschmidt@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Abstract

We examined how naïve conversational participants circumscribed referential domains during the production and comprehension of referring expressions by monitoring participants' eye movements during a referential communication task. The results replicated some well-established results, e.g., incremental reference resolution, demonstrating the feasibility of studying real-time language comprehension in interactive conversation. We also observed a high proportion of underspecified referential expressions that were easily understood by addresses because of discourse and pragmatic constraints, including constraints developed as a result of the conversation.

Background

In characterizing work in language performance, Clark (1992) pointed out that the field has been largely divided into two traditions. One tradition, the language-as-action tradition, emphasizes interactive conversation as the most basic form of language use. According this tradition the principles of language performance and language design cannot be understood without taking into account the interactive collaborative processes that are embedded in conversation. A central tenet in work within this tradition is that utterances can only be understood within a particular context, which includes the time, place and participant's conversation goals. Thus researchers within this tradition have focused primarily in investigations of interactive conversation using natural tasks, typically with real-world referents.

A second tradition, the language-as-product tradition, focuses primarily on the processes by which listeners decode (and speakers encode) linguistic utterances. Psycholinguistic work on language comprehension within in this tradition typically examines moment-by-moment processes in real-time language processing using fine-grained reaction time measures. The rationale for using these measures is that comprehension processes are closely time-locked to the linguistic input which, for spoken language, unfolds over time. Until recently, the real-time response measures in the psycholinguist's toolkit required the use of de-contextualized language, typically pre-recorded sentences presented in impoverished contexts. This constraint ruled out real-time investigations of natural, interactive conversation. Moreover, a dominant theoretical perspective within the product tradition was that initial "core" processes (e.g., lexical access and syntactic processing) were informationally encapsulated from contextual influences (e.g., Fodor, 1983).

Recently, the advent of light-weight head-mounted eye-tracking systems has made it possible to investigate real-time comprehension in more natural tasks, such as tasks where participants follow spoken instructions to manipulate objects in a task-relevant "visual world" (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Fixations to task-relevant objects are closely time-locked to the unfolding utterance, providing a continuous real-time measure of comprehension processes at a temporal grain fine enough to track the earliest moments of lexical access, parsing and reference resolution (Tanenhaus, Magnuson, Dahan & Chambers, 2000).

A growing body of research employing eye-tracking techniques demonstrates clear effects of contextual constraints. For example, syntactic ambiguity resolution is influenced by referential constraints provided by the visual context, including the number of potential referents and their properties (Tanenhaus et al., 2000). Moreover, some recent work using confederates in constrained tasks suggests that under some circumstances information provided by knowledge of the speaker's perspective and intentions can affect even the earliest moments of comprehension (Hanna, 2001).

However, a major limitation of previous work is that all of the language used has come from scripted language, ruling out spontaneous collaborative processes that are likely to underlie circumscription of referential domains,

and interpretation of referential expressions in natural interactive settings. For example, Clark & Wilkes-Gibbs (1986) investigated conversational partners' use of collaborative processes to refer to low-codability shapes in a referential communication task (Krauss and Weinheimer, 1966). Pairs of participants worked together to arrange different abstract shapes. Over the course of the conversation they converged on shared names for the shapes, dramatically increasing the efficiency of their communication over the course of the interaction. In Brennan's (1996) words, conversational partners develop 'conceptual pacts' during the course of a conversation. The mere action of participating in the development of conversational pacts is essential for the increase in efficiency- overhearers privy to the entirety of the conversation and it's context are unable to perform as well in these natural tasks (Schober and Clark, 1989). These results suggest two things: 1) The act of participating in a natural conversation contributes to efficient communication. 2) Extracting conversational interaction from language comprehension removes a central component of natural language.

The goal of the present study was to explore the feasibility of examining real-time comprehension processes during natural, unscripted, interactive conversation. We focused on the comprehension of definite referring expressions, such as "the red block" and "the cloud". Definite reference provides an ideal domain for a first investigation for several reasons. First, definite reference is one of the most ubiquitous and central components of natural language. Second, use of definite reference assumes that a referent will be uniquely identifiable within a circumscribed referential domain. Much of the strongest evidence for the collaborative model of language processing comes from demonstrations that people collaborate to define referential domains. Third, work with restricted utterances has established two clear empirical results that allow one to track the time course of reference resolution: lexical competitor (cohort) effects and 'point of disambiguation' effects.

When listeners are instructed to pick up or move an object, such as a racket, fixations to the target object begin as early as 200 ms after the onset of the noun (Allopenna, Magnuson & Tanenhaus, 1998). Eye-movements launched at this point in the speech stream are equally likely to be directed to the eventual referent and other objects with names that are also consistent with the speech signal, such as a raccoon. However, looks to these competitors, hereafter "cohort" competitors are reduced or eliminated when context makes a cohort an implausible referent. Thus we can use cohort effects to infer the degree to which conversation restricts initial referential domains.

One of the most striking sources of evidence for rapid restriction of referential domains comes from point of disambiguation effects. For example, Eberhard, Spivey-Knowlton, Sedivy and Tanenhaus (1995) presented subjects with displays containing a variety of differently colored shapes, as subjects listened to instructions such as “Click on the red triangle”. In a subset of trials the color of the target item was not shared with any other items in the referential domain. In these trials the referentially disambiguating information was the color, which was conveyed in the prenominal adjective. In the remaining trials, the target item was the same color as another item in the referential domain. For example, the display accompanying the instruction “Click on the red triangle” might contain a red circle and a red triangle. In these trials the referentially disambiguating information came at the noun. Eye movements to the target were again closely time-locked to the speech. Looks to the target increased dramatically immediately following the point of disambiguation (POD), whether it came at the adjective or the noun.

In the present experiment we monitored eye-movements as pairs of participants, separated by a curtain, worked together to arrange blocks in matching configurations and confirm those configurations. The characteristics of the blocks afforded comparison with findings from scripted experiments investigating language-driven eye-movements, specifically those demonstrating cohort effects and incremental reference resolution. We investigated: (1) whether these effects could be observed in a more complex domain during unrestricted conversation, and (2) under what conditions the effects would be eliminated, indicating that factors outside the speech itself might be operating to circumscribe the referential domain.

Method

We tested four pairs of participants from the University of Rochester, who were paid for their participation in the study. The discourse partners each had an array of blocks and a board on which to place them. The boards were partially covered, creating 5 distinct sub-areas. Initially, sub-areas contained 56 stickers representing blocks. The task was to replace each sticker with a matching block. While partners' boards were identical with respect to sub-areas, partners' stickers differed: Every place that one partner had a sticker, the other partner had an empty spot. Pairs were instructed to tell each other where to put blocks so that in the end their boards would match. No other restrictions were placed on the interaction. The entire experimental study lasted approximately 2.5 hours. For each pair we recorded the eye movements of one partner, and the speech of both partners.

The initial configuration of the stickers was such that the color, size, and orientation of the blocks would encourage the use of complex noun phrases and grounding constructions. Nineteen of the stickers (and the corresponding blocks) contained pictures similar to those used by Allopenna, Magnuson and Tanenhaus (1998), in the study described above. We used a full-color version (Rossion & Pourtois, 2001) of a large corpus of normed pictures, balanced for their linguistic codablilty (Snodgrass & Vanderwart, 1980). We selected pairs of these pictures that referred to objects which had initially acoustically

consistent names (cohort competitors). Half of the cohort competitor stickers were placed such that both cohort competitor blocks would be placed in the same sub-area of the board, and half of the cohort competitor stickers were placed such that the cohort competitor blocks would be placed in different sub-areas of the board. All of the cohort competitor pairs were separated by approximately 3.5 inches. We examined the eye movements of one discourse partner with respect to the speech generated by the other discourse partner.

Results

The conversations for each of the four pairs were transcribed. We present eye-tracking analyses for two of the pairs; we are still analyzing the data from the remaining two pairs. The non-eye-tracked partner of each pair generated approximately 100-150 definite references to blocks during the course of the conversation. While the length of the conversation prevented us from initially analyzing more than 4 pairs, the large number of 'trials' generated by each pair gave us enough statistical power to circumvent this problem.

An ISCAN eyetracking visor was used (for details see Trueswell, Sekerina, Hill & Logrip, 1999). The image of the eye-tracked partner's board, and their superimposed eye position, along with the entirety of the conversation (both participants' voices) were recorded using a frame-accurate digital video recorder (a SONY DSR-30). Eye movements were analyzed at the onset of the definite reference, and continued 2000ms after the NP was complete. There was a high degree of variability in the length of utterances, especially those to color blocks.

References to blocks which had cohort competitors (approximately 75 references per pair) were expected to reveal similar cohort effects as observed in more restricted experimental paradigms. To our surprise, we observed only one look to a cohort competitors during both 2 1/2 hour conversations we have analyzed thus far. We do not think this null result is due to poor stimulus design, as we did observe looks to cohort competitors under special circumstances. Periodically, subjects needed to remove the eye-tracker to take a break. When we put the tracker back on and re-calibrated, we tested the calibration by asking the subjects to look at different items on the board. Under these circumstances the referential domain is not restricted by conversational constraints. Here we saw clear cases of subjects initially looking at the cohort competitor before looking at the intended referent.

Each pair provided us with approximately 75 trials of data for eye-movements elicited by definite references to colored blocks. Two researchers coded the definite NPs for their point-of-disambiguation, and resolved any coding differences. The POD was the point at which the NP uniquely identified a referent, given the visual context at the time. Average POD was 858ms (26 frames) following NP onset. Eye-movement analyses for NPs with a unique linguistic point of disambiguation (50%) were analyzed separately from those which were never fully disambiguated linguistically. The eye-tracking analysis was restricted to cases where at least one competitor block was present. As a result, the number of ambiguous trials used for the analysis was close to 50, while there were only approximately 20 for the disambiguated trials.

Eye-movement analyses (planned comparisons) were performed on 800ms epochs for both ambiguous and disambiguated NPs. Eye movements elicited by disambiguated references showed clear point of disambiguation effects; within 200ms of the disambiguation point, looks converged on the target block: we found a main effect of condition $F(2,20)= 64.03$, $p<.0001$, and Bonferroni post-tests revealed significantly more looks to the 'target' than 'competitor' and 'other' unrelated blocks. (see Figure 1). Before the disambiguation point, subjects were looking equally at the target and competitor block(s), the main effect of condition, $F(2,19)= 6.77$, $p<.01$, was due to significantly more looks to target than other blocks ($p<.001$); looks to target and competitor were equivalent at this region. At the 522ms baseline region (-1122 to -600ms) there was no effect for condition. This replicates the point-of-disambiguation effect seen by Eberhard, et al. (1995), demonstrating that we were successful in using a more natural task to investigate on-line language processing. Upon inspection of the eye-movement plot, one should note that we do observe a pre-POD target bias. We will argue that this initial bias is due to additional pragmatic constraints that are operating during the task.

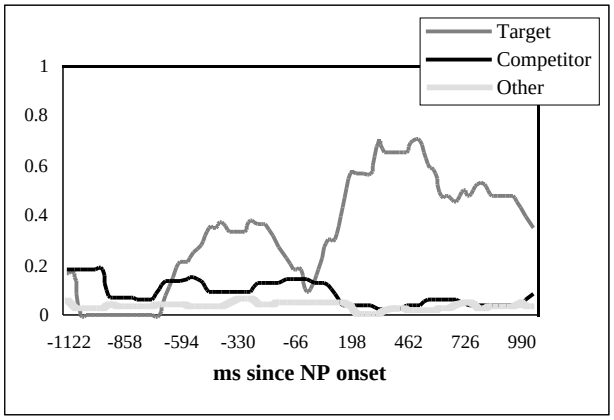


Figure 1: Proportion of fixations to Targets, Competitors, and Other blocks by time (ms) for Disambiguated NPs. Graph is centered by item with 0 ms = POD onset.

Most remarkably, ambiguous utterances elicited significantly more looks to the target than unambiguous utterances (see Figure 2). Moreover, fixations were primarily restricted to the referent shortly after onset of the definite reference; we observed significantly more looks to targets than competitors within the first 200 ms of NP onset, a significant effect of condition, $F(2,53)= 18.37$, $p<.0001$, was due to significant differences between target and both

competitor and other blocks (p's<.0001). This effect persisted in the second 800ms window.

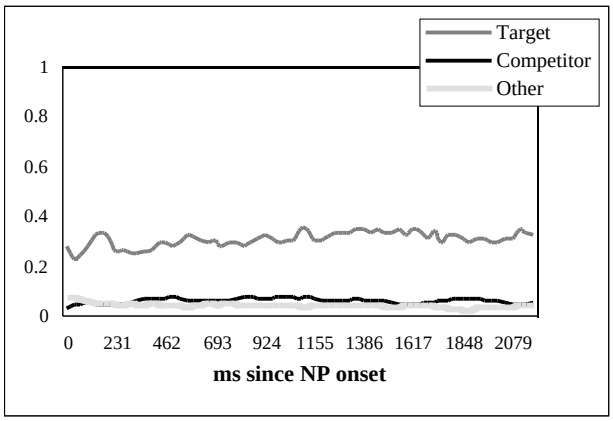


Figure 2. Proportion of fixations to Targets, Competitors, and Other blocks by time (ms) for Ambiguous NPs. Zero ms = NP onset.

These results suggest that 1) speakers systematically use less specific utterances when the referential domain has been otherwise constrained; 2) the attentional states of speakers and addresses become closely tuned; and 3) utterances are interpreted with respect to referential domains circumscribed by contextual constraints.

In order to identify the factors that led speakers to choose underspecified referring expressions, and that enabled addresses to understand them, we performed a detailed analysis of all of the definite references.

Recency is one factor that is likely to influence the form of a referring expression, with the most recently mentioned entity being more salient than other (non-focused) entities. (We are assuming that references to the most focused entity would typically result in use of a pronoun, a hypothesis that we are currently evaluating in the data set.) Thus, recency of last mention of the target block should predict the degree of specification, with references to the most recently mentioned block of a type, resulting in ambiguous referring expressions For example, if "the green block" is uttered in the context of a set of 10 blocks, 2 of which are green, 'recency' would predict that the referent should be to the green block that was most recently mentioned. Indeed, this is what we found: Ambiguous utterances were more likely to refer to the most recently mentioned block of its type, while this effect was the opposite for Disambiguated utterances. This effect was substantiated by a significant chi-square for independence, χ^2=7.389, p=.01.

While recency of mention is clearly an important factor constraining interpretation of ambiguous references, approximately 36% of these references did *not* refer to the most recently mentioned block. Additionally, the recency analysis alone does not explain why speakers sometimes chose to fully specify a reference when referring to the most recently mentioned block of its type; in fact 30% of disambiguated utterances referred to the most recently mentioned block of its type. Two questions arise from these observations: 1) Why are addresses not confused when a speaker uses an underspecified expression to refer to something other than the most recently mentioned thing; 2) What factors determine when speakers will add extra information when referring to the most recently mentioned block?

In answer to the first question, we propose that additional pragmatic and task-based factors function to constrain the referential domain, allowing speakers to underspecify, addresses to interpret ambiguous references, and may explain the early target advantage in the disambiguated trials. In addition to recency of mention, we identified three factors which contributed to the intelligibility of referring expressions: 1) Proximity of target referent to the last mentioned block; 2) Task-based constraints (such as limitations on block placement due to the size and shape of the board); 3) Spatial Terms which focus attention to a subset of the work area. We are currently performing a detailed analysis of the interaction of these constraints, and a comparison of how the predictions made by the constraints compare with the predictions of the addressee.

Finally, analysis of the eyetracking data of the final two pairs will allow us to do a sub-analysis of the eye-movements during trials influenced by these different factors. Our hypothesis is that both linguistic factors (such as recency of mention) in addition to pragmatic factors (such as proximity and task constraints) are contributing to the ease with which subjects are identifying the intended referents of these ambiguous references.

The answer to the second question, is the reverse of the answer to the first. We would like to suggest that in cases where confusion is high, conversation is inefficient, or these additional task and pragmatic constraints may select the *wrong* referent, speakers may choose to add additional disambiguating information. In order to verify this claim, we intend to compare the referential domain circumscribed by these additional constraints, with the intended referent of the speaker. We predict that a mismatch would lead to an increase in the likelihood of extra disambiguating information.

As a part of this analysis, we have also looked at the cases in which speakers overdisambiguate. In approximately 21% of the disambiguated utterances, speakers added between 1 and 3 additional elements past the POD. In the following example, the reference was disambiguated at 'long', yet the speaker chose to continue: "the long green square that was laying down". This speaker added two extra elements, a color term, and a collaboratively defined term, which means 'horizontal'. In many cases, the speaker spent a relatively large amount of time uttering extra disambiguating information, especially lengthy collaborative terms (Collaboratively-defined terms were common in this corpus; approximately 20% of ambiguous and 40% of disambiguated references contained collaboratively-defined terms). Of the overspecifying elements used, only 50% were

color terms, which are the prototypical overdisambiguating element. The other 50% included references to previous actions, the location of the object and its shape. In general, speaker overspecification may be for clear communicative purposes, rather than as a bi-product of the production system.

Conclusions

In this experiment, we investigated 1) whether it is possible to observe incremental processing effects in a complex domain during unrestricted conversation, and 2) under what conditions these effects might be absent, indicating factors outside the speech itself might be operating to circumscribe the referential domain. We were successful on both counts. We did observe incremental reference resolution in certain contexts, and in the contexts in which it was not observed we were able to identify a number of constraints that seemed to be operating to circumscribe the referential domain. These constraints included linguistic recency, pragmatic factors related to the task itself, such as physical limitations on block placement due to the size and shape of the board, and proximity of a given referent to a previously mentioned block. An example of a segment of the discourse in which recency circumscribes the referential domain is shown below:

2. ok. RIGHT directly next to the cloud?
1. mm-hmm
2. just throw in a red piece, line it up evenly
1. just a red, little *square*
2. yup
1. k, got it
2. ok
1. now, I got an easy one, so I wanna *give it* to you
2. *ok*
1. directly…ABOVE the red, grab your lamp

The first description of the target block (underlined) is unambiguous, but the second reference to the block is ambiguous given the visual context (there are 5 other red blocks in that sub-area of the board). Listeners do not have difficulty with the linguistic ambiguity in this situation because they take recency into account, unifying the referents of the two referring expressions. An example of a segment of the discourse in which task constraints circumscribe the referential domain is shown in below:

1. ok, you're gonna line it up… it's gonna go <pause> one row ABOVE the green one, directly next to it.
2. can't fit it
1. cardboard?
2. can't yup, cardboard
1. well, tell it too back
2. the only way I can do it is if I move, alright, should the green piece with the clown be directly lined up with thuuh square?

Again, the referring expression (underlined) is ambiguous given the visual context. In this case the listener does not have difficulty dealing with this ambiguity because, although there are a number of blocks one could line up with "the green piece with the clown", only one is task-relevant. Given the location of all the blocks in the relevant sub-area, the target block is the easiest block to line up with the clown. The competitor blocks are inaccessible because of the position of the other blocks or the design of the board. An example of a segment of the discourse in which proximity constrains the referential domain is the following:

2. ok, so it's four down, you're gonna go over four, and then you're gonna put the piece right there
1. ok…how many spaces do you have between this green piece and the one to the left of it, vertically up?

As before, the referring expression (underlined) is ambiguous given the visual context; there are approximately five blocks up and to the left of the previously focused block (the one referred to in the NP as "this green piece"). In this case the listener does not have difficulty dealing with the ambiguity because he considers only the block closest to the last mentioned block ("this green piece"). Finally, an example of a reference that is constrained by a spatial term is underlined in the following exchange:

2: ok, and then…alright, so then there is a dark green one? to thEE uh northeast of that green one?
1: yup
2: and, um, they're only overlapping...one...line and then there's a yellow one…**below** the dark green one that I just talked about and to the l- **to the RIGHT** of the other dark green one.

The underlined reference is constrained before the onset of the noun phrase by the spatial terms used before it (bolded). When the listener hears the target noun phrase, she is already aware that the referent is **below** 'the dark green one', and that there is a space to the left of it (as she is directed to place a yellow block **to the right** of the referent). This information narrows the interpretation of the reference down to a single block, whereas the reference was otherwise ambiguous with respect to that sub-area in general.

We are currently detailing the predictions of and the interactions between these different constraints and the degree to which they predict both speaker behavior, and the interpretation processes of the addressee. Our observations and analysis of incremental interpretation during this task suggest a view of language processing in which conversational participants coordinate a mutually aware reliance on certain discourse, pragmatic and task based constraints which facilitate efficient completion of the task at hand. Our data mark an important first step towards being able to rigorously analyze the on-line processing of interactive conversation 'in the wild'. To our knowledge, this is the first demonstration of on-line circumscription of referential domains in a natural interactive task with naïve participants. As we continue to develop more explicit models of on-line language processing, a critical part of this process should be to inform these models with observations made in these natural situations. Pairing methodologically rigorous laboratory studies with naturalistic studies such as this one is essential to an understanding of language processes that is both detailed, and ecologically valid.

To conclude, we successfully replicated a standard psycholinguistic effect, the point of disambiguation effect, in unscripted interactive conversation with naïve participants. We also obtained results suggesting that reference selection and comprehension is modulated by both discourse-based factors, such as recency, but also by task specific pragmatic constraints.

Moreover, reference resolution appeared to be affected by collaborative constraints that developed during the conversation. Our participants spontaneously created collaborative terms for troublesome words (such as 'horizontal' and 'vertical'), and tuned their utterances, and comprehension systems for such details as the recency of mention of each particular kind of block, proximity of blocks to one another, and task constraints idiosyncratic to our block-game. These observations suggest that the attentional states of the speaker and listener become closely tuned during the course of interaction. An important question for future research is how these factors differentially affect speakers and addressees. The data that we have collected is rich enough to allow us to investigate this and other questions.

Acknowledgments

This research was partially supported by NIH grant HD 27206. Thank you to Jesseca Aqui, Annie Tanenhaus, and Sanjukta Sanyal for all their help.

References

Allopenna, P. D, Magnuson, J.S. & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419-439**.**

Brennan, S. and Clark, H. (1996) Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: LMC,* 22, 482-493.

Clark, H. (1992) Arenas of Language Use. Chicago: University of Chicago.

Clark, H. & Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C. & Tanenhaus, M.K. (l995). Eye-movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409-436.

Fodor (1983). Modularity of Mind. Cambridge, MS; Bradford Books.

Hanna (2001). The effects of linguistic form, common ground, and perspective on domains of referential interpretation. Unpublished doctoral dissertation. The University of Rochester.

Krauss, R. M. and Weinheimer, S. (1966) Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4, 343-346.

Rossion, B. & Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's Object database: Color and Texture improve Object Recognition. 1st Vision Conference, Sarasota, FL.

Snodgrass, J.G., & Vanderwart, M. (1980). *JEP:Human Learning and Memory*, 6:3, 174-215.

Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21, 211-232.

Trueswell, J.C., Sekerina, I., Hill, N. & Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89-134.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, **29**, 557-580.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. & Sedivy, J.E. (l995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 632-634.

On Straight TRACS: A baseline bias from mental models

Kevin Burns (kburns@mitre.org)
The MITRE Corporation, 202 Burlington Road
Bedford, MA 01730-1420 USA

Abstract

TRACS (Tool for Research on Adaptive Cognitive Strategies) is a family of games played with a special deck of two-sided cards (see www.tracsgame.com). TRACS has the advantage of being both mathematically tractable to theoretical analysis and psychologically relevant to practical applications. The simplest game, called Straight TRACS, is a series of choices where the player must turn over one of two cards to match a third card. The object is to make the most matches on a trip through the deck. The challenge is to track the changing odds in order to make the best choices. We performed experiments and simulations to measure human performance in this probabilistic and dynamic task. We present our finding of a Baseline Bias, in which subjective odds are (incorrectly) anchored to the baseline odds. This is an interesting result because it is contrary to other well-known biases, such as Gambler's Fallacy, in which subjective odds are (incorrectly) not anchored to the baseline odds. We propose a theory of mental models to reconcile our finding with previous research on heuristics and biases.

Introduction

A dilemma of decision research is to obtain the rigors of controlled experimentation yet maintain some relevance to practical applications. Our approach is a synthetic task environment (Gray, in press) called "TRACS" (Burns, 2001a) that replicates the cognitive challenges of naturalistic decision-making (Klein, 1998), including probabilistic risk assessment and dynamic resource allocation.

TRACS is both a unique game and a useful tool (Burns, 2001b). From a mathematical perspective, TRACS is unique because it is played with a special deck of two-sided cards, and because it has extensible rules that allow the same game to be played alone or with others.

Unlike standard playing cards, the backs of the cards provide clues to the fronts, and the deck contains unequal numbers of each card type (Table 1). Compared to Poker and other games of imperfect information, the two-sided cards make TRACS more tractable to theoretical analysis of optimal solutions. Compared to Chess and other games of perfect information, the two-sided cards make TRACS more relevant to diagnoses and decisions in practical domains like business, medicine and warfare.

From a psychological perspective, TRACS is useful because it provides a naturalistic micro world for experiments and simulations. Unlike other approaches to research on probabilistic reasoning, which often employ verbal stimuli in the form of static questions, TRACS employs graphical stimuli in a game of dynamic situations. This reduces artificial framing effects (see Nickerson, 1996) and introduces realistic temporal context.

We are using TRACS to perform experiments on human subjects and to perform simulations with software agents. Our experiments are designed to elicit cognitive strategies and our simulations are designed to evaluate these strategies against normative standards. Taken together, our experiments and simulations allow us to build and test models of cognitive competence that are relevant to practical applications in command and control (Burns, 2001c).

This paper reports on our first experiment and simulations using the simplest version of the game, called Straight TRACS. We explain the game, discuss our experiment and present our finding of a Baseline Bias. We also propose a theory of mental models to reconcile our finding with previous research on heuristics and biases.

The Game

Straight TRACS is a solitaire game played with 24 two-sided cards (Table 1). The backs of the cards, called "tracks", show black shapes (triangle, circle or square). The fronts of the cards, called "treads", show colored sets (Red or Blue) of these same shapes. Table 1 shows the distribution of shape/color (track/tread) cards in the deck. This distribution defines the baseline odds. For example, at the start of the game, a triangle track is likely (6/8 = 75%) to be Red, a square track is likely (6/8 = 75%) to be Blue and a circle track is 50-50. However, during the game, the odds change as the deck is depleted.

Table 1: Distribution of cards in the deck. The backs are called "tracks" and the fronts are called "treads".

# of Cards	6	4	2	2	4	6
Front (tread)	Red	Red	Red	Blue	Blue	Blue
Back (track)	▲	●	■	▲	●	■

To play Straight TRACS, the deck is held face down and three cards are dealt to a field. Two cards are dealt face down (showing their tracks) and the third card is dealt face up (showing its tread). The player's task is to turn over one of the two tracks (revealing its tread), trying to match the tread (color) of the third card. The turn is scored a "save" if the treads match or a "strike" if the treads clash. The two treads are removed from the field and the remaining track is turned to reveal its tread. This becomes the tread to match on the next turn. Two new tracks are dealt from the deck, a track is turned, the treads are scored, etc. Play continues until all cards (except the last two, which do not count) have been paired and scored. The object of the game is to minimize strikes on a trip through the deck.

Experiment

The goal of our experiment was to measure how people track the changing odds in TRACS. The probe in our experiment was a confidence meter that players set before each turn to indicate their subjective belief in the color (tread) of each shape (track) on the field. We used two different confidence meters (Figure 1), both based on a spectrum that runs from 100% Red to 100% Blue. One confidence meter displayed a discrete set of qualitative values on an octal scale. The other confidence meter displayed a continuous set of quantitative values on a decimal scale.

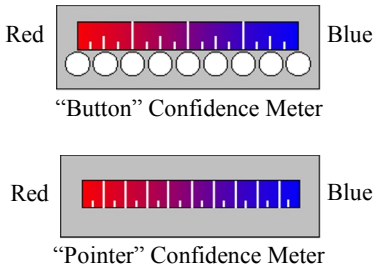


Figure 1: Two different confidence meters.

We tested 43 adults playing 10 games each. Subjects were tested on a personal computer using a mouse to set the confidence meter. There were no time limits, but each game was typically completed in less than 5 minutes. Each subject played in two blocks of 5 games, one block with each confidence meter in balanced design to control for fatigue and learning effects. The two blocks were separated by a short break. Before data collection, subjects read a playbook that described the cards and rules, watched a demo and played a practice game. All games were played with random shuffles of the deck and all 43*10 = 430 shuffles were unique. The experimental results were similar for Button and Pointer confidence meters, so all data is combined here, rounding Pointer data to the nearest Button for consistency.

Analysis

Baseline Bias

The player's problem is illustrated in Figure 2, which shows the actual odds for a typical game. By convention, we measure odds in % Red, where % Blue = 100 - % Red. Figure 2 shows that the odds for each track type start at their baseline values (75% Red for triangles, 50% Red for circles and 25% Red for squares). However, the actual odds change (moving up/down on Figure 2) as tracks are turned to reveal their treads (moving right on Figure 2).

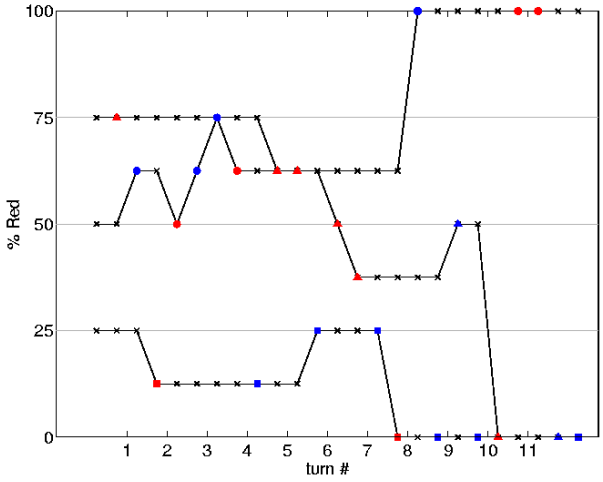


Figure 2: Change in actual odds (typical game).

Figure 3 illustrates a typical player's solution to the problem illustrated in Figure 2, as reported by the player's setting of the confidence meter for each track (before each turn). Relative to the actual odds (Figure 2), we see that the reported odds exhibit a bias towards the baseline odds. For example, after a minor adjustment near the start of the game, the player (Figure 3) reported constant odds for circles even after the actual odds (Figure 2) had moved far from the baseline. This Baseline Bias is explored further below.

Odds Inversions

Recall that the object of Straight TRACS is to turn the track that is most likely to match a given tread. At the start of the game, this is a simple task since the baseline odds specify which track to turn, e.g., triangle rather than circle to match Red. However, as the deck is depleted, the actual odds for two track types may become "inverted" with respect to their baseline configuration. This occurs whenever there is a crossover of two track types on the dynamic odds plot (Figure 2).

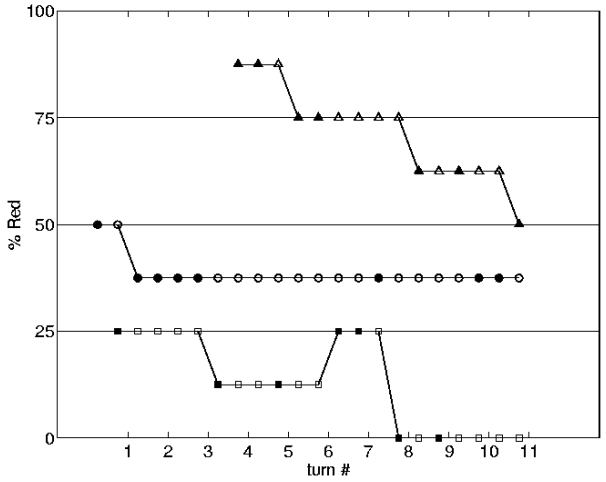


Figure 3: Change in reported odds (typical game).

For example, in Figure 2, a crossover near the middle of the game causes the odds to be less % Red for triangles than for circles for the remainder of the game. This is an inversion of the baseline odds relation between triangles and circles. Figure 3 shows that the player failed to detect this odds inversion.

As a gross measure of cognitive competence, we treat each odds inversion as a signal that a player must detect in order to minimize strikes in Straight TRACS. Figure 4 shows the total number of hits, misses and false alarms for this signal (for all players and all games). The relatively small number of hits compared to misses demonstrates that subjects exhibit a Baseline Bias. The occurrence of some hits and false alarms suggests that, although biased towards the baseline odds, subjects are at least trying to update odds, i.e., they are not just playing the baseline odds.

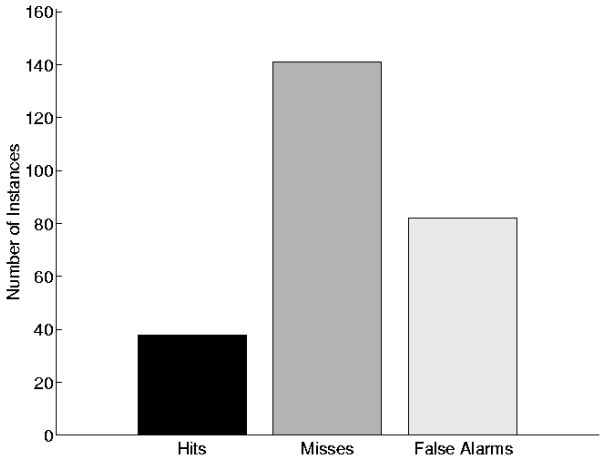


Figure 4: Detection of odds inversions.

Average Error

Each odds inversion (see above) involves a pair of tracks. As another measure of cognitive competence, we also examine confidence errors for single tracks. Figure 5 shows the average error versus turn in game, for human subjects and for a simulated agent that always plays the baseline odds.

Figure 5 shows that error increases with turn, i.e., as the actual odds deviate more from the baseline odds, for both the human subjects and the baseline agent. This shows that people have a Baseline Bias relative to the actual odds (zero error). Figure 5 also shows that the average error is higher for human subjects than for the baseline agent at the start of the game. This is a surprising result because: (1) The baseline odds are explicitly illustrated on the cards (treads) for the player to see. (2) The actual odds are obviously equal to the baseline odds at the start of the game. (3) Playing the baseline odds is a strategy that requires virtually no mental effort.

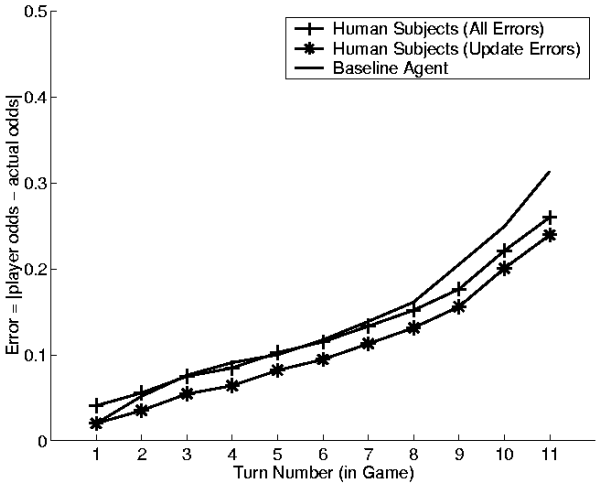
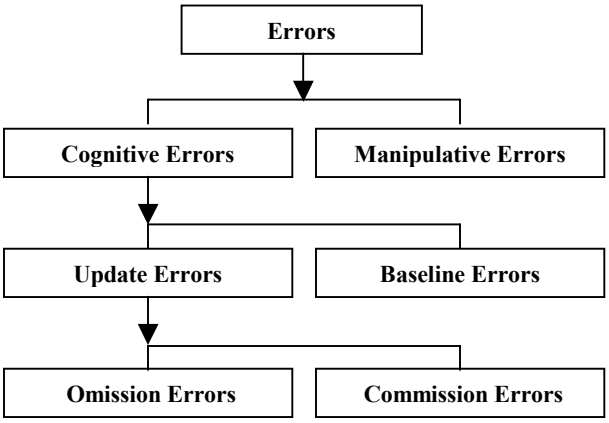


Figure 5: Average error in reported odds.

Kinds of Errors

To help explain Figure 5, we define a taxonomy of errors (Table 2). We first distinguish between Cognitive Errors, which are mental errors in judging odds, and Manipulative Errors, which are physical errors in moving the mouse to match the mind. We then distinguish between two kinds of Cognitive Errors: Update Errors are mental errors in updating odds relative to baseline odds, and Baseline Errors are mental errors in estimating the baseline odds themselves. Finally, we further distinguish between two kinds of Update Errors: Omission Errors are where no mental update is performed when it should be, and Commission Errors are where a mental update is performed but the result is incorrect.

Table 2: A taxonomy of errors.



The baseline agent makes no Manipulative Errors, no Baseline Errors and no Commission Errors, i.e., it makes only Update Errors of Omission. In fact, since the baseline agent never updates odds, its performance provides an upper bound on the magnitude of Omission Errors. Figure 5 shows that the baseline agent's Update Error is non-zero on turn 1. This is because the tread on the first field can cause a change in odds before the first turn. For example, assume that the cards on the first field are circle (track), Red circle (tread) and square (track). The baseline odds for circles are 4/8 Red, but since one Red circle is revealed as a tread on the field, the actual (updated) odds for circles are 3/7 Red. The same effect is magnified on later turns as more treads are revealed, hence the Omission Error increases with turn (Figure 5, curve for baseline agent).

For human subjects, the total error comprises Manipulative Error, Baseline Error and Update Error (Omission and Commission). The difference between total human error and baseline agent error on turn 1 is attributed to Manipulative Error and Baseline Error, which we assume are relatively independent of turn in game. Thus, the curve for total human error can be shifted downwards (curve * in Figure 5) to get an estimate of human Update Error.

This shifted curve for human error is directly comparable to the error curve for the baseline agent, which also includes only Update Error. The comparison (Figure 5) shows that human subjects are biased towards the baseline agent, relative to the actual odds (zero error). However, the shifted curve also shows that human subjects outperform the baseline agent (who does not update odds), and the difference grows with turn as the difference between actual odds and baseline odds increases. Thus, we conclude that human subjects make fewer Omission Errors than the baseline agent, and that the Commission Errors made by human subjects are not significantly larger in number and magnitude than the baseline agent's Omission Errors.

Anchoring and Adjustment

These results are consistent with the well-known heuristic strategy of "anchoring and adjustment" (Tversky & Kahneman, 1974). In our case, the baseline odds (with some Baseline Error, see above) are the anchor to which people make adjustments. The adjustments are, on average, better than pure anchoring but significantly worse than optimal adjusting.

To gain further insight into the adjustments, we examine how the confidence meter settings change from turn to turn (for the same track type). We define various magnitudes of adjustment (i.e., no-button jump, one-button jump, two-button jump, etc.) and compute the number of times each magnitude of adjustment was made. Figure 6 compares the results for human subjects to an optimal agent (playing the same games) who always sets the confidence meter at the actual odds. As expected from anchoring, human subjects most often make a no-button adjustment. This is in contrast to the optimal agent, who most often makes a one-button adjustment.

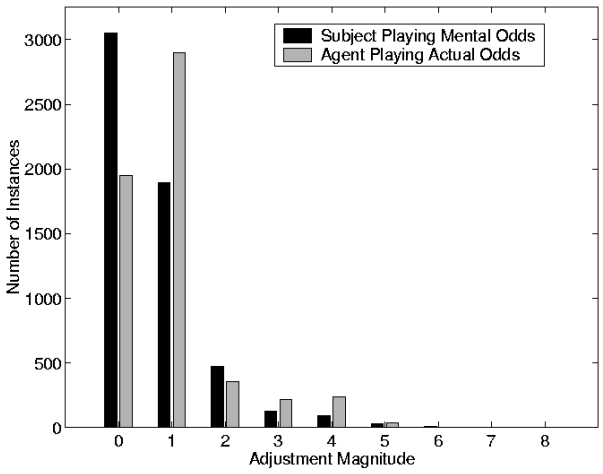


Figure 6: Number of adjustments (by magnitude).

Besides the magnitude of adjustment, we also examine various types of adjustments. We define five types of adjustments (Table 3) and compute the number of times each type of adjustment was made. Figure 7 compares the results for human subjects to an optimal agent (playing the same games) who always sets the confidence meter at the actual odds.

Table 3: Types of adjustments (anchor = baseline odds).

Type 0	No adjustment
Type 1	From on-anchor to off-anchor
Type 2	From off-anchor to more off-anchor
Type 3	From more off-anchor to less off-anchor
Type 4	From off-anchor to on-anchor

As expected from anchoring, Figure 7 shows that human subjects make many more Type 0 adjustments (actually non-adjustments) and less Type 1 and Type 2 adjustments than the optimal agent. Figure 7 also shows that the difference between Type 1 and Type 4 adjustments is smaller for human subjects than for the optimal agent. This suggests that, when people do move off the baseline anchor (Type 1) in an attempt to adjust odds, they often "lose it" and return to the baseline anchor (Type 4). The optimal agent moves off the baseline anchor (Type 1) more often and returns to the anchor (Type 4) only when the actual odds are equal to the baseline odds (i.e., the agent never "loses it").

These results (Figures 6 and 7) further support our conclusion that Baseline Bias in TRACS (Figures 4 and 5) is caused by a heuristic strategy of anchoring and adjustment.

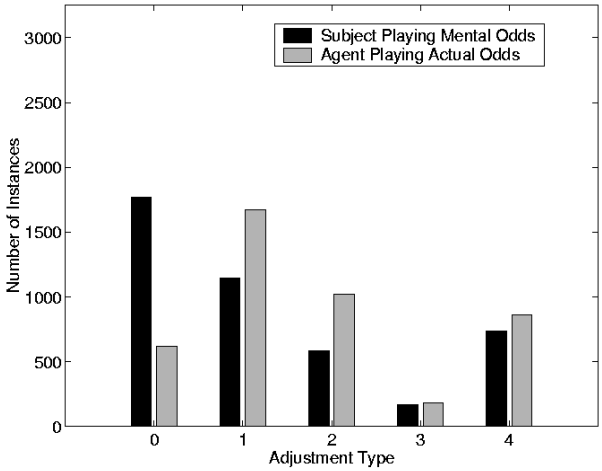


Figure 7: Number of adjustments (by type).

The question, of course, is how (exactly) do people decide when and how much to move off anchor? That is, what (exactly) are the mental limits that prevent more accurate adjustments? The answer is crucial if we are to explain and predict human performance in TRACS or any other domain where people must think probabilistically about things that are changing dynamically. Below we propose a theory of mental models that takes a first step in this direction.

Theory

Mind Sets

We claim (Burns, in press; 2002; 2001b; 2001c; 2001d) that people make sense of the world by forming probabilistic models in their heads (see Knill & Richards, 1996; Johnson-Laird, 1994; Gigerenzer et al., 1991).

We further claim that: (1) Mental models are bounded by natural regularities of the world. (2) Mental models have a normative basis within their natural bounds. (3) Cognitive competence can be explained and predicted by specifying the natural bounds and normative basis of mental models (i.e., in bounded rationality, see Simon, 1990).

More specifically, we claim that probabilistic mental models can be characterized as computational "mind sets" (Burns, 2002). Like the term "model", which is both noun and verb, we use the term "set" in dual sense to refer to both the mental representation of classes (e.g., declarative knowledge) and the mental operation of routines (e.g., procedural knowledge). The central tenet of our theory (Burns 2002) is that these mind sets are normative within their cognitive bounds.

As an initial test, below we sketch how our theory can explain Baseline Bias in TRACS. We also sketch how our theory can reconcile Baseline Bias with previous findings of Gambler's Fallacy and Base Rate Neglect in other probabilistic reasoning tasks (Tversky & Kahneman, 1974). This is a non-trivial test of the theory, since Gambler's Fallacy and Base Rate Neglect appear at first glance to be contrary to Baseline Bias.

Gambler's Fallacy and Base Rate Neglect

In Baseline Bias (in TRACS), people do not update the baseline odds when they should. Conversely, in Gambler's Fallacy, people update the baseline odds when they should not. Furthermore, in Base Rate Neglect, people discount or ignore the baseline odds altogether. How can we explain these differences? According to our theory, all three biases occur because people reason about probabilities with mind sets.

For Base Rate Neglect (Tversky & Kahneman, 1974; Koehler, 1996; Cosmides & Tooby, 1996), we suggest that people ignore the baseline odds in light of other evidence because they believe that the baseline odds reflect a less relevant (not applicable) set of occurrences. It is difficult in theory, let alone in practice, to aggregate probabilities that are derived from diverse sources with different pedigrees. As such, it is a bounded-Bayesian strategy to rely on the one source that is judged to be most relevant and reliable. Base rates that are judged irrelevant or unreliable are therefore neglected.

For Gambler's Fallacy (Tversky & Kahneman, 1974), we suggest that people update the baseline odds because they are judging odds for a finite (bounded) set rather than for an infinite set. For example, after seeing 10 heads and 2 tails, a gambler who believes the coin is fair will think that the future holds more tails than heads, simply because he thinks that the eventual (large but finite) set of many tosses for this coin will be balanced. As such, it is a bounded-Bayesian inference to conclude that the future odds are slightly higher for tails than for heads.

For Baseline Bias (in TRACS), we suggest that people want to update odds (as they tell us) but that it is simply beyond their cognitive capacity. To do so, players must count cards in each of six sets (see Table 1) and normalize to convert the counts to odds. These two tasks, i.e., concurrent counting and normalizing numbers, are naturally hard for the unaided mind (Dehaene, 1997; Dehaene, 1992; Gallistel & Gelman, 1992; Nickerson, 1996). Thus, with self-knowledge of mental limits, it is a bounded-Bayesian strategy to remain anchored to the baseline odds unless and until the evidence for an adjustment is compelling. For example, in the extreme case, pure anchoring to baseline odds (i.e., never adjusting) is the bounded-Bayesian strategy for a decision maker who knows that he cannot remember which cards have been revealed in the course of a game.

Conclusion

Our initial experiment and simulations show that TRACS provides a useful micro world for investigating how people make diagnoses and decisions under uncertainty. Our finding in Straight TRACS is that players exhibit a Baseline Bias, which we attribute to a heuristic strategy of anchoring and adjustment. We sketched a theory of set-based mental models that reconciles our finding with previous research on heuristics and biases. Our future plans are to use TRACS to investigate the mental limits of concurrent counting, normalizing numbers and other facets of cognitive competence in probabilistic and dynamic reasoning.

Acknowledgements

This research was supported by the MITRE Technology Program. Thanks to Craig Bonaceto, Eric Forbell and Fritz Behr for their work on the experiment and simulations.

References

Burns, K. (2001a). TRACS: A Tool for Research on Adaptive Cognitive Strategies: The Game of Confidence and Consequence. *Published on the World Wide Web,* www.tracsgame.com.

Burns, K. (2001b). Introducing TRACS: A unique game and a useful tool. *Annual Meeting of the Society for Judgment and Decision Making*, Orlando, Florida.

Burns, K. (2001c). A Bayesian basis for mental models. Draft manuscript.

Burns, K. (2001d). Mental models of line drawings. *Perception*, 30, 1249-1261.

Burns, K. (2002). Mind sets at TRACS: How people deal with changing odds. *40th Annual Bayesian Research Conference*, Studio City, California.

Burns, K. (in press). Mental models and normal errors. In Brehmer, B., Lipshitz, R., Montgomery, H. (Eds.), *How Do Professionals Make Decisions?* (Lawrence Erlbaum). Also presented at *5th Conference on Naturalistic Decision Making*, Tammsvik, Sweden, May 2000.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44,* 1-42.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics* (New York: Oxford).

Gallistel, C., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44,* 43-74.

Gigerenzer, G., Hoffrage, U., Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Gray, W. (in press). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments and micro worlds in basic and applied cognitive research. *Special Joint Issue of Cognitive Science Quarterly and Kognitionswissenschaft*.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition, 50,* 189-209.

Klein, G. (1998). *Sources of Power: How People Make Decisions* (Cambridge, MA: MIT Press).

Knill, D., & Richards, W. (1996). *Perception as Bayesian Inference* (New York: Cambridge University Press).

Koehler, J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1-53.

Nickerson, R. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin, 120,* 410-433.

Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology, 41,* 1-19.

Tversky A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Contradictions and Counterfactuals: Generating Belief Revisions in Conditional Inference

Ruth M.J. Byrne (rmbyrne@tcd.ie)
Psychology Department, University of Dublin,
Trinity College, Dublin, Ireland

Clare R. Walsh (cwalsh@tcd.ie)
Psychology Department, University of Dublin,
Trinity College, Dublin, Ireland

Abstract

Reasoners revise their beliefs in the premises when an inference they have made is contradicted. We describe the results of an experiment that shows that the belief they revise depends on the inference they have made. They revise their belief in a conditional (if A then B) when they make a modus tollens inference (from not-B to not-A) that is subsequently contradicted (A). But when they make a modus ponens inference (from A to B) that is contradicted (not-B) they revise their belief in the categorical assertion (A). The experiment shows that this *inference contradiction effect* occurs not only for factual conditionals but also for counterfactual conditionals. However, reasoners revise their beliefs in factual conditionals more than counterfactuals.

Belief Revision

Suppose you know the following well-established set of knowledge to be true:

If the car was out of petrol then it stalled.
The car was out of petrol.
What, if anything, follows?

You may conclude:

The car stalled.

But suppose additional knowledge comes to light at a later time and you discover the following is true:

The car did not stall.

What do you think you should believe to be true at this point?

New information can contradict previously held beliefs and inferences about the world. The ability to recognise inconsistency is a necessary step in revising beliefs (e.g., Legrenzi, Girotto, & Johnson-Laird, 2002). Once inconsistencies and contradictions are detected, they must be resolved (e.g., Elio & Pelletier, 1997). For example, you may decide to revise your belief in the conditional, and believe instead that the car being out of petrol does not necessarily mean that it stalled (it may be a diesel engine). Or you may revise your belief in the categorical, and believe instead that the car was not entirely out of petrol.

Dealing with contradictions is common not only in scientific discovery but also in everyday 'non-monotonic' inference. Which beliefs do people revise most readily? Conditionals can convey explanations, regularities or hypotheses about the world; categoricals can convey facts, data or observations (Elio, 1997). Revising the conditional or categorical is equally acceptable logically (Revlin, Cate, & Rouss, 2002). Yet most studies show that reasoners revise their belief in the conditional (Dieussaert, Schaeken, De Neys, & d'Ydewalle, 2000; Elio, 1997; Elio & Pelletier, 1997 Politzer & Carles, 2001; Revlin, et al, 2002).

Reasoners may prefer to revise some sorts of beliefs more than others because they accommodate new information by changing little of their existing beliefs (Harman, 1986). Minimal changes can be accomplished by altering beliefs that have the least explanatory power or informational content (Gardenfors, 1988). Categoricals convey more semantic information (they rule out more states of affairs as false) than conditionals (Johnson-Laird & Byrne, 1991); but conditionals and categoricals may differ in how entrenched they are in a belief system.

Intriguingly, some studies suggest that the belief reasoners revise depends on the inference contradicted. Consider a second problem:

If the car was out of petrol then it stalled.
The car did not stall.

You may conclude:

The car was not out of petrol..

But suppose the additional knowledge comes to light:

The car was out of petrol.

What do you think you should believe to be true? Once again you may decide to revise your belief in the conditional; or you may revise your belief in the categorical, and believe instead that the car did stall. The first example illustrates a modus ponens inference and the second illustrates a modus tollens inference, and Table 1 summarizes the structure of the two sorts of problem.

Table 1: Two types of belief revision problem

	Modus ponens	Modus tollens
1. Conditional	If A then B	If A then B
2. Categorical	A	Not-B
3. Conclusion	B	Not-A
4. Contradiction	Not-B	A

The Inference Contradiction Effect

Some studies show that reasoners revise their belief in the conditional more when a modus tollens inference has been contradicted, whereas they revise their belief in the categorical more when a modus ponens has been contradicted (Elio, 1997, experiment 1; Politzer & Carles, 2001). The possibility that the belief reasoners revise depends on the inference that has been contradicted, which we will call the *inference contradiction effect*, is puzzling. The contradiction establishes the same counterexample for both inferences, e.g., the car was out of petrol and it did not stall (A and not-B), yet the counterexample is accommodated differently in each case.

However, it is by no means clear whether an inference contradiction effect exists: some studies show the opposite pattern (Dieussaert, et al, 2000; Revlin, et al, 2002), and others show more revision of the conditional following a modus ponens inference, but equal revision of the categorical following modus ponens and tollens (Elio & Pelletier, 1997). The vagaries may arise because previous studies have asked participants to select from different sorts of - sometimes quite complex and constrained - options, e.g., to indicate denial or doubt about each of the statements, to choose one statement to reject, to rate degrees of belief, or to choose among various compound options such as 'disbelieve conditional and uncertain about categorical'. Our aim in the experiment we report is to establish whether an inference contradiction effect exists, and so we allowed participants to generate their own revisions, unfettered by pre-set selection options.

Previous studies have also presented a conclusion to participants prior to contradicting it, without requiring participants to indicate their evaluation of the inference. A participant who has not made the inference, or who does not agree that the presented inference is valid, may not need to engage in belief revision following the subsequent 'contradiction'. To guard against such a possibility, we allowed participants to generate the inferences they considered to follow from the premises, prior to presenting them with a contradiction, and in this way we ensured that their beliefs were genuinely contradicted.

Our conjecture is that an inference contradiction effect occurs because different cognitive processes are required to alter conditional and categorical beliefs following modus ponens and tollens inferences, and we return to this idea after we consider some new data.

Generation of Belief Revisions

We constructed a set of 8 problems, consisting of three modus ponens inferences, three modus tollens inferences, and two fillers based on quantifiers. The problems were based on a science fiction content about different aliens, their properties, living habits and so on (in other experiments we have examined causal and definitional contents, see Byrne and Walsh, 2002). The content and instructions were adapted from Elio & Pelletier (1997) and Politzer & Carles (2001). Participants were told they would be given 'an initial set of knowledge that was true and well established at the time you began exploring. There were no mistakes at that time'. They were given a set of premises on a page of a booklet (e.g., if A then B, A) and asked to write what, if anything, followed. On the next page, they were given the contradiction (e.g., not B). They were told this information was 'additional knowledge about the planet that has come to light at a later time. This knowledge is also true and well established. The world is still the same but what has happened is that knowledge about it has increased'. Their task was ' to try to reconcile the initial knowledge and the additional knowledge. You are to write down what you now believe to be true of all the knowledge you have at this point'.

The conditionals given to one group of participants were phrased in the indicative mood, e.g., 'If the ancient ruin was inhabited by Pings, then it had a force field surrounding it', and those given to a second group were in the subjunctive mood e.g., 'If the ancient ruin had been inhabited by Pings, then it would have had a force field surrounding it'. The participants were 28 undergraduates of the psychology department at the University of Dublin, Trinity College who participated for course credit

Belief Revision Responses

The sorts of revisions that reasoners spontaneously generated fall into three main categories:

1. Revisions or negations of the conditional. Reasoners indicated that the original interpretation of the conditional needed to be revised, saying, e.g., 'A does not mean must B', 'If A don't have to B', 'not all A's do B'. Or they denied its truth, e.g., 'that B if A is false', 'the original statement that A's B is incorrect'. Revisions of the interpretation were far more common than negations.

2. Revisions or negations of the categorical. Negating the categorical for modus ponens leads to the conclusion

'not-A', for modus tollens it leads to the conclusion 'B' via the double negation 'not not-B'. In other cases, reasoners deduced a new conclusion from the contradiction and the conditional. The contradiction for modus ponens is 'not B', and with the conditional leads (via modus tollens) to 'not-B and so not-A' (which is also the denial of the categorical). The contradiction for modus tollens is 'A', and with the conditional leads (via modus ponens) to ''B' (see also Elio & Pelletier, 1997). This tactic leads to the same conclusion as the previous one, but by a different process.

3. Reasoners affirmed the contradiction and the categorical, either in combination or separately. This tactic led to the conclusion 'A and not-B' (or equivalently, 'not-B and A'). Reasoners find it difficult to make the inference from 'not (if A then B)' to the conclusion 'A and not-B' (Handley, 1996), which supports the suggestion that the conclusion 'A and not-B' is reached by a different process from the conclusions in 1.

Revisions to Factual Conditionals

We report first the results for the participants who received indicative conditionals. They made the modus ponens and tollens inferences frequently (100% and 90% respectively) perhaps unsurprisingly given the content. Most participants generated one revision (81%) and we scored those who generated more than one by their first one (see Byrne & Walsh, 2002 for details).

Table 2: The percentages of revision types for modus ponens and tollens for indicative conditionals

	Modus Ponens	Modus Tollens	Mean
Revise conditional	33	54	44
Revise categorical	41	18	30
Affirm contradiction and/or categorical	8	18	13

Participants revised their belief in the conditional somewhat more than the categorical (44% versus 30%, binomial z = 1.32, 1-tailed p = .093). However, they revised their belief in the conditional more often when modus tollens was contradicted than when modus ponens was contradicted (54% versus 33%, Wilcoxon z = 1.94, p = .05), whereas they revised their belief in the categorical more often when modus ponens was contradicted than when modus tollens was contradicted (41% versus 18%, Wilcoxon z = 2.20, p = .03), as Table 2 shows. Some of their responses consisted of affirmations of the contradiction and the categorical (A and not-B) either together or separately (13%), as Table 2 shows. The remainder of responses consisted largely of explanations of the premises or contradiction, or assertions that none of the premises were true.

The results confirm earlier findings that reasoners revise their belief in the conditional more than the categorical; perhaps more importantly the results also confirm earlier findings of an inference contradiction effect, that is, the belief revised depends on the inference contradicted. In this experiment, the direction of the inference contradiction effect is that reasoners revise the conditional more following modus tollens and the categorical more following modus ponens (for similar results see Elio, 1997; Politzer & Carles, 2001).

The generated revisions show that reasoners do not revise their beliefs solely by rejecting or disbelieving one or both of the premises, nor by doubting or expressing uncertainty in one or both of them. Instead their revisions actively attempt to re-interpret the premises in a way that genuinely reconciles the conflicting information and resolves the contradiction, for example, by calling into question the necessity of the antecedent for the consequent. This sort of revision has not been identified in previous studies which relied on presented selections only. The generated revisions also show that reasoners do not confine themselves solely to revising their categorical or conditional beliefs; a third category of responses emerged which consisted of affirmations of (one or both of) the contradiction and the categorical. It is noteworthy that no participant generated a response which simply affirmed the conditional.

Revisions to Counterfactual Conditionals

The second group of participants received counterfactual conditionals in the subjunctive mood e.g., 'If the Spracks had had high-frequency sound sensor ears then they would have had tentacles'. A counterfactual seems to mean something different from its corresponding factual conditional (Costello & McCarthy, 1999; Ginsberg, 1986; Lewis, 1973; Stalnaker, 1968). It conveys the presupposition that the facts are 'Spracks do not have high-frequency sound sensor ears' and 'Spracks do not have tentacles'. When reasoners are given a surprise memory test for counterfactuals, they mistakenly identify that they were given these facts (Fillenbaum, 1974). They judge that someone uttering a counterfactual means to imply these facts (Thompson & Byrne, in press). They make the modus tollens inference more readily from a counterfactual than from a corresponding factual conditional (Byrne & Tasso, 1999). They make the modus ponens inference just as readily from both sorts of conditional.

Since counterfactual conditionals convey both the facts 'Spracks do not have high-frequency sound sensor

ears', 'Spracks do not have tentacles', as well as the suppositions, 'Spracks have high-frequency sound sensor ears', 'Spracks have tentacles', we considered that reasoners would not revise their beliefs in counterfactual conditionals as often as factual conditionals.

Table 3: The percentages of revision types for modus ponens and tollens for counterfactual conditionals

	Modus Ponens	Modus Tollens	Mean
Revise conditional	16	38	27
Revise categorical	53	13	33
Affirm contradiction / categorical	18	29	24

The results support our conjecture, as Table 3 shows. Once again, participants made the modus ponens and tollens inferences frequently (96% and 87% respectively). A comparison of the means in both tables shows that reasoners revise a factual conditional more than a counterfactual conditional (44% vs 27%, chi^2= 5.29, $p < .05$). For a counterfactual conditional, they often affirmed the contradiction and the categorical (together or separately). The results also show the presence of an inference contradiction effect, and its direction is the same for counterfactual as for factual conditionals.

Cognitive Processes in Belief Revision

Different cognitive processes may be required to alter conditional and categorical beliefs following modus ponens and tollens inferences from a factual conditional. The different effects of the counterexample, A and not-B, on the way reasoners revise their beliefs may arise because of the mental representations they have constructed in the course of making an inference.

Reasoners may understand conditionals by keeping in mind different possibilities (Johnson-Laird & Byrne, 1991). The explicit set of models for the conditional 'if A then B are as follows:

A B
Not-A not-B
Not-A B

where in the diagram 'not' is a propositional-like tag to indicate negation. Reasoners who interpret the conditional as a biconditional will construct the first two models in the set only. Regardless of their interpretation, reasoners may construct an initial set of models that makes some information explicit, but leaves other information implicit, because of the constraints of working memory:

A B
. . .

where the three dots represent an implicit model (Johnson-Laird & Byrne, in press).

Reasoners can make the modus ponens inference from this initial set of models. The categorical, A, is consistent with the explicit model:

A B

which supports the parsimonious conclusion, B. To make the modus tollens inference, they must flesh out the initial set of models to be more explicit. The information, not-B, cannot be incorporated readily into the initial set of models. However, it can be incorporated into the fleshed-out models, it eliminates two of them and it leaves a single model:

not-A not-B

which supports the parsimonious conclusion, not-A.

The process by which the two inferences are made differs. The modus ponens inference is made from the initial set of models, but the modus tollens inference is made only after fleshing out the models to be explicit, eliminating models that are inconsistent. This difference in the process of making the inferences may affect the revision of beliefs.

Consider the contradiction to the modus ponens inference. Reasoners must incorporate the contradiction 'not-B'. They made the inference, 'B', based on the initial set of models, and so they have the option of returning to the initial set to flesh them out to be more explicit. Faced with the contradiction, they may decide they need to think more fully about the possibilities compatible with the conditional. People often return to earlier possibilities to think about what might have been (e.g., Byrne & McEleney, 2000). When they do so they can incorporate the contradiction 'not-B' into one of the fleshed out models:

Not-A not-B

The new model indicates that the belief to revise is the categorical, A.

For the modus tollens inference, reasoners must incorporate the contradiction, A. They made the modus tollens inference by fleshing out the models to be more explicit and eliminating all but the model:

Not-A not-B

They do not have the option of returning to flesh out the initial models again, since they have executed that option in the process of making the inference. In the course of making the inference they have considered several alternatives and eliminated them, 'cashing out' the possibilities to one single remaining possibility. The contradiction 'A' cannot be incorporated into the existing model and so the belief to revise is the conditional, if A then B.

The essential revision principle may be that a contradiction can be incorporated into one of the possibilities compatible with the conditional, if these

possibilities have not been thought about and eliminated already. In the case of modus ponens, the only possibility compatible with the contradiction and the conditional (not-A not-B) is incompatible with the categorical (A) and so the categorical must be revised. In the case of modus tollens, the possibilities have been exhausted already in the course of making the inference, and so the conditional itself must be revised. The inference contradiction effect, that the belief revised depends on the inference made before the contradiction, may arise because different inferences require people to keep in mind different possibilities, which subsequently limits their room for manuever in incorporating contradictory information.

Reasoners can rely on background knowledge to add or eliminate possibilities (Johnson-Laird & Byrne, in press). As a result, when they have relied on knowledge to add or eliminate possibilities, their revisions may *not* be influenced by the inferences they have made (Byrne and Walsh, 2002). For example, given a causal conditional, 'if water was thrown on the campfire then it went out', and 'the fire did not go out', reasoners make the modus tollens inference, 'water was not thrown on the campfire'. But when the inference is contradicted 'water was thrown on the campfire' they can incorporate it by saying, for example, 'not enough water was thrown on the campfire'. Reasoners may even short-cut the process by 'matching' various models (Legrenzi, Girotto, & Johnson-Laird, 2002). The inference contradiction effect may be a feature of certain kinds of content.

Modifying or Abandoning Beliefs?

Previous studies have focused on what beliefs people disbelieve, deny or reject, doubt or are uncertain about. However, a contradiction can call for a revision to the original *interpretation* of the premises. A putative counterexample, A and not-B, does not necessarily mean that a conditional, if A then B, is false. Our participants generated revisions to the interpretation (e.g., 'A's do not necessarily have B's', 'Some other variable affected B, e.g., C') more often than they indicated disbelief, denial, rejection, doubt or uncertainty about the conditional's truth.

Reasoners may reach many different interpretations of a conditional (Johnson-Laird & Byrne, in press). One interpretation of 'if A then B' is that A is sufficient but not necessary for B, and a second is that A is both necessary and sufficient for B. These 'conditional' and 'biconditional' interpretations are inconsistent with the counterexample, A and not-B. But other interpretations are consistent with it, for example, that A is necessary but not sufficient for B. This 'reverse conditional' interpretation may be common in everyday reasoning (Byrne, Espino, & Santamaria, 1999).

The reverse conditional interpretation can occur when an additional requirement is made explicit, e.g., 'if the ruin was inhabited by Pings then it had a force field, if they had to protect their habitations then it had a force field'. The modus ponens inference from, e.g., 'the ruin was inhabited by Pings' is suppressed (Byrne, 1989). Reasoners say there is not enough information, or they incorporate the second requirement e.g., 'The ruin had a force field if the Pings had to protect their habitations' (Byrne, et al, 1999). They select options that refer to the second requirement (e.g. Diuessaert, Schaeken, Schroyens, & d'Ydewalle, 2000) and they judge the requirements to be conjoint (Byrne & Johnson-Laird, 1992). When both requirements are affirmed they readily make the inferences (Byrne, 1989; 1991), and they can be enhanced when reasoners know the additional requirements have been satisfied (Manktelow & Fairley, 2000). The suppression is increased when the additional requirement is emphasized, by phrasing it as a biconditional, 'if and only if the Pings had to protect their habitations…' (Byrne, et al, 1999), by relying on familiar content (Chan & Chua, 1994; see also Bonnefon & Hilton, in press), by qualifying its satisfaction (Stevenson & Over, 1995), or by specifying that the requirement was uttered by an expert rather than a novice (Stevenson & Over, 2001). Conditionals with many additional requirements lead to more suppression (Cummins, Lubart, Alksnis, & Rist. 1991; see also Elio, 1997). The conditions in which a reverse conditional are true can be specified with as much certainty as the truth conditions of a conditional or biconditional (but see Politzer & Braine, 1991; Stevenson & Over, 1995; 2001).

Our results show that revising belief in a conditional can mean modifying the original interpretation, for example changing from a conditional interpretation to a reverse conditional one, rather than abandoning belief in the truth of the conditional. In everyday life, just as in scientific thought, it may be rare to abandon entirely either a theory or a fact, upon discovery of another contradictory fact; instead reasoners may progress by attempting to modify their interpretation to restore consistency.

Conclusions

In everyday reasoning, the conclusions to inferences can be readily withdrawn in the light of subsequent information, that is, they are non-monotonic. An important task in everyday inference is the revision of beliefs in the light of contradictions. The results also show an inference contradiction effect, that is, reasoners revise a categorical belief when a modus ponens inference is contradicted and they revise a conditional belief when a modus tollens inference is contradicted. Our results show that reasoners revise their belief in a factual conditional more than a counterfactual

conditional. Our novel revision generation task allowed us to capture some of the rich re-interpretations that people produce to resolve contradictions through modifying rather than abandoning beliefs.

Acknowledgements

We thank Jean-Francois Bonnefon, Renee Elio, Uri Hasson, Phil Johnson-Laird, Mark Keane, and Guy Politzer for helpful comments on an earlier draft and Michelle Cowley and Michelle Flood for help with the experiment. The research was supported by the Dublin University Arts and Social Sciences Benefactions Fund.

References

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31, 61-83.

Byrne, R.M.J. (1991). Can valid inferences be suppressed? *Cognition*, 39, 71-78.

Byrne, R.M.J. & Johnson-Laird, P.N. (1992). The spontaneous use of propositional connectives. *Quarterly Journal of Experimental Psychology*, 45A, 89-110.

Byrne, R.M.J., Espino, O. and Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language*, 40, 347-373.

Byrne, R. M. J. & Tasso, A. (1999). Deductive reasoning with factual, possible and counterfactual conditionals. *Memory and Cognition*, 27, 726-740.

Byrne, R.M.J. & McEleney, A. (2000) Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1318-1331.

Byrne, R.M.J. & Walsh, C.R. (2002). Belief revision, the inference contradiction effect and counterfactual conditionals. *Manuscript in preparation.*

Bonnefon, J-P, and Hilton, D. (in press). The suppression of modus ponens as a case of pragmatic preconditional reasoning, *Thinking and Reasoning.*

Costello, T., & McCarthy, J. (1999). Useful Counterfactuals. *Electronic Transactions on the Web, 3,* 51-76.

Chan, D. & Chua, D. (1994). Suppression of valid inferences: syntactic views, mental models and relative salience. *Cognition,* 53, 217-238.

Cummins, D.D., Lubart, T., Alksnis, O. and Rist. (1991). Conditional reasoning and causation. *Memory and Cognition,* 19, 274-282.

Diuessaert, K, Schaeken, W., De Neys, W. & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Current Psychology of Cognition*, 19, 277-288.

Diuessaert, K, Schaeken, W., Schroyens, W. & d'Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking and Reasoning*, 6, 125-160.

Elio R. (1997). What to believe when inferences are contradicted. In M. Shafto & P.Langley (Eds). *Proceedings of the 19th Conference of the Cognitive Science Society.* Hillsdale: Erlbaum. pp. 211-216.

Elio, R. & Pelletier, F.J. (1997). Belief change as propositional update. *Cognitive Science*, 21, 419-460.

Fillenbaum, S. (1974). Information amplified: memory for counterfactual conditionals. *Journal of Experimental Psychology*, 102, 44-49.

Gardenfors, P. (1988). *Knowledge in flux.* Cambridge, MA: MIT Press.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence*, 30, 35-79.

Harman, G. (1986). *Change in view.* Cambridge, MA: MIT Press.

Handley, S. (1996). Explicit negation. *Phd thesis, University of Wales.*

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Erlbaum.

Johnson-Laird, P. N. & Byrne, R. M. J. (in press). Conditionals: a theory of meaning, inference, and pragmatics. *Psychological Review*.

Legrenzi, P. , Girotto V., & Johnson-Laird, P.N. (2002). Models of consistency. *Manuscript.*

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Manktelow, K.I. and Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning*, 6, 41-65.

Politzer, G. and Braine, M.D.S. (1991). Responses to inconsistent premises cannot count as suppression of valid inferences. *Cognition*, 38, 103-108.

Politzer, G. and Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking and Reasoning*, 7, 217-234.

Revlin, R., Cate, C.L., & Rouss, T.S. (2002). Reasoning counterfactually: combining and rending. *Memory and Cognition.*

Stalnaker, R.C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*. Oxford: Basil Blackwell.

Stevenson, R.J. and Over, D.E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, 48A, 613-643.

Stevenson, R.J. and Over, D.E. (2001). Reasoning from uncertain premises: effects of expertise and conversational context. *Thinking and Reasoning*, 7, 367-390.

Thompson, V. & Byrne, R.M.J. (in press). Making inferences about things that didn't happen. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* .

Anthropomorphic Agents as a User Interface Paradigm: Experimental Findings and a Framework for Research

Richard Catrambone (rc7@prism.gatech.edu)
School of Psychology, Georgia Institute of Technology, 274 5th Street
Atlanta, GA 30332-0170 USA

John Stasko (stasko@cc.gatech.edu)
Jun Xiao (junxiao@cc.gatech.edu)
College of Computing, Georgia Institute of Technology
Atlanta, GA 30332-0280 USA

Abstract

Research on anthropomorphic agent interfaces has produced widely divergent results. We suggest that this is due to insufficient consideration of key factors that influence the perception and effectiveness of agent-based interfaces. We propose a framework for studying anthropomorphic agents that can systematize the research. The framework emphasizes features of the agent, the user, and the task the user is performing. Our initial experiment within this framework manipulated the agent's appearance (lifelike versus iconic) and the nature of the user's task (carrying out procedures versus providing opinions). We found that the perception of the agent was strongly influenced by the task while features of the agent that we manipulated had little effect.

Introduction

If you could ask for assistance from a smart, spoken natural language help system, would that be an improvement over an on-line reference manual? Presumably the answer, in most cases, is yes, for two reasons. First, the spoken natural language aspect would allow you to speak your questions rather than having to type them. Generally this is a faster approach for most people. Second, the smart aspect would improve the chance of the help system finding the information you want even if you do not state the query using the correct or most appropriate terms.

The state of the art in this style of interface is a human user consultant. Does it matter that the user consultant has a face and that the face can have expressions and convey a personality? Would a face affect you in terms of your comfort and satisfaction with the interaction? Would the presence of a face make the help or advice you receive more persuasive? The answers to such questions have implications for the design of training systems, customer service, information kiosks, and many other applications.

Many people believe that anthropomorphic computer interfaces have great potential to be beneficial for a number of reasons. Agents could act as smart assistants, much like travel agents or investment advisors, helping people manage the ever-growing amount of information encountered today (Lyman & Varian, 2002). Further, a conversational interface appears to be a natural dialog style in which the user does not have to learn complex command structure and functionality (Laurel, 1990).

An anthropomorphic interface could use intonation, gaze patterns, and facial expressions, in addition to words, for conveying information and affect. The human face seems to occupy a privileged position for conveying a great deal of information, including relatively subtle information, efficiently (Fridlund & Gilbert, 1985). Anthropomorphic interfaces could make a computer more human-like, engaging, entertaining, approachable, and understandable to the user, thus harboring potential to build trust and establish relationships with users, and make them feel more comfortable with computers.

These potential advantages are balanced by strong negatives. Anthropomorphic agent interfaces are viewed by some researchers as being impractical and inappropriate. Current speech recognition, natural language understanding, and learning capabilities of computers still fall far short of any human assistant.

More specifically, it has been proposed that agent systems disempower users by clouding issues such as who is responsible for a system's actions (Lanier, 1995). Others feel that user interfaces are more beneficial when they clearly reflect the commands available to a user and present the objects that a user can act upon (Shneiderman, 1997). Furthermore, critics argue that agent interfaces may mislead both users and designers, increase user anxiety, reduce user control, undermine user responsibility, and destroy a user's sense of accomplishment (Shneiderman & Maes, 1997). Many current anthropomorphic or personified interfaces are viewed as being annoying, silly characters who hinder rather than enhance productivity (e.g., the Microsoft Paper Clip).

Although strong opinions have been voiced on both sides of this issue, relatively little careful empirical research on anthropomorphic interfaces has been conducted, and the results from this research have been contradictory or equivocal (Cassell, 2000).

Our goal is to develop a framework to systematically evaluate and understand the autonomous agent as a user interface paradigm. The present paper outlines the framework and an initial study that examines two issues within this framework. The first is whether the degree to which an interface agent is anthropomorphic has a measurable effect on users. Note that anthropomorphism is not a dichotomy but rather a continuum. One can think of interfaces with full fidelity video or 3D images of people to

more caricature-style characters to 2D cartoons of people or personified characters such as dogs or toasters.

The second issue is to what extent the nature of the task will influence a user's perception of an agent. Some tasks might be more likely to induce a user to imbue the agent with human-like qualities (such as if the user had to engage the agent in a debate) while other tasks might lead the user to view the agent simply as a reference tool (e.g., for providing reminders of keystrokes for a software application) with no "individuality."

Related Work

A few studies have revealed that anthropomorphic agents are attention-grabbing and people make natural assumptions about the intelligence and abilities of those agents. King and Ohya (1996) found that a dynamic 3D human form whose eyes blinked was rated more intelligent than any other form, including non-blinking 3D forms, caricatures, and geometric shapes.

One common trend discovered in studies is that anthropomorphic interfaces appear to command people's attention, both in positive and negative senses. Takeuchi and Nagao (1995) created conversational style interaction systems that allowed corresponding facial displays to be included or omitted. According to their metrics, the conversations with a face present were more "successful." Across two experiments they found that the presence of a face provided important extra conversational cues, but that this also required more effort from the human interacting with the system and sometimes served as a distraction.

Other studies have shown that the attention garnered by an anthropomorphic interface had a more positive effect. Walker, Sproull, and Subramani (1994) found that people who interacted with a talking face spent more time on an on-line questionnaire, made fewer mistakes, and wrote more comments than those who answered a text questionnaire. Koda (1996) created a Web-based poker game in which a human user could compete with other personified computer characters including a realistic image, cartoon male and female characters, a smiley face, no face, and a dog. She gathered data on people's subjective impressions of the characters and found that people's impressions of a character were different in a task context than in isolation and were strongly influenced by perceived agent competence.

An influential body of related work is that of Nass and his colleagues. Their efforts focus on the study of "Computers as Social Actors." They have conducted a number of experiments that examined how people react to computer systems and applications that have certain personified characteristics (Nass, Isbister, & Lee, 2000; Nass, Steuer, & Tauber, 1994; Rickenberg & Reeves, 2000). Their chief finding is that people interact with and characterize computer systems in a social manner, much as they do with other people. Furthermore, they suggest that findings in the social psychology literature (e.g., individuals with similar personalities tend to get along better than do those with different personalities) apply even when one of the two participants is a machine.

The studies cited above, and others, suggest that people are inclined to attribute human-like characteristics to agents and that a variety of factors might influence how positively the agents are viewed. Dehn and van Mulken (2000) provide a more extensive review of this literature.

A Framework for Research on Anthropomorphic Interface Agents

To effectively and systematically investigate the use of anthropomorphic interface agents, one needs to consider the key factors that will affect the usefulness of such interfaces. We propose an investigative framework composed of three key components: characteristics of the user, attributes of the agent, and the task being performed.

We believe that serious empirical study in this area must systematically address each of these factors and understand how it affects human users. Below, we provide examples of individual variables within each factor that could potentially influence user performance and impressions.

Factor 1: Features of the User

Potential users vary, of course, in many ways. However, there are certain features that may be quite likely to affect how useful a user finds an agent. These features include:

Personality: Researchers have identified what are referred to as the "Big Five" traits that seem to be quite useful in describing human personalities: extraversion, openness, agreeableness, neuroticism, and conscientiousness (e.g., McCrae & Costa, 1987). While any such breakdown is debatable, it seems reasonable to examine whether users' positions on these, or other, trait dimensions predicts how they will respond to agents.

Background Knowledge: A user who has a good deal of background knowledge in a domain might prefer an agent that is reactive and that the user can call upon when he or she needs some low-level bit of information or has a low-level task that needs to be done. Conversely, a user who is learning how to carry out tasks in a particular domain might welcome strategy advice from an agent, particularly if the agent can analyze the strategy and provide reasons for why the strategy might be altered.

Other Variables: Other user-related variables include gender, age, and computer experience.

Factor 2: Features of the Agent

Fidelity: Earlier studies suggest that more realistic-appearing, 3D human representations are perceived as being more intelligent, which could be viewed positively or negatively. However, realistic-appearing agents are more difficult to implement, so if user performance is improved by the presence of an agent, but does not vary according to appearance, simpler caricature style characters would be advantageous.

Presence: Is an agent's face always present on the screen or does the agent only appear when it is engaged in a dialog

with the user? One might hypothesize that an ever-present agent would make users uneasy by producing an effect of being watched or evaluated all the time.

Role: Should an agent act as a partner in the task or should it contribute only in clearly specified ways? For instance, an agent might be able to offer strategy guidance for design tasks. Alternatively, it might provide only lower-level procedural "how to" information.

Initiative: Should an agent proactively make suggestions and offer guidance or should it respond only when directly addressed? A proactive agent might be viewed as being "pushy" and might bother users, or it could be viewed as being extremely helpful and intelligent if it acts in situations in which the user is unsure of how to proceed or is so confused that he or she is unable to form a coherent help request.

Other Variables: Other agent-related variables to consider are expressiveness, speech quality, "gender," "personality," and competence.

Factor 3: Features of the Task

Tasks can vary in a variety of ways. Some tasks can be opinion-like (e.g., choosing what to bring on a trip) while others are more objective (e.g., solving a puzzle) in terms of assessing the quality of a solution. Some involve a good deal of high-level planning (e.g., writing a talk) while others are more rote (e.g., changing boldface words into italics). Tasks might be classified along some or all of the dimensions listed below:

Objectiveness: The situation might be an opinion-based one in which the user is seeking advice and recommendations on some topic (e.g., which items to pack for a trip to Europe). Alternatively, the user might be carrying out an objective task such as acquiring facts (e.g., finding the keystroke combination for a particular command in a software application).

Intent: The user could have a learning goal or alternatively may be carrying out a set of steps in a familiar domain. In the latter, the user might want help with low-level details whereas in the former the user is looking for guidance as to the structure of the domain.

Other Variables: Other task-related variables to consider are domain, degree of time pressure, duration, and consequences of the quality of task performance.

The number of variables within each factor is certainly larger than the number we have identified here. No doubt these factors will also interact. For instance, a novice attempting to carry out a task in a particular domain might welcome proactive comments/advice from an agent while someone with more experience could get annoyed.

With respect to measuring the usefulness of an agent, we have to consider which dependent measures are most appropriate. Towards the more objective end, a user's performance on a task in terms of accuracy and time--when such measures are meaningful--can give one indication of usefulness. Thus, time and errors would be appropriate measures for a text-editing task. Towards the more subjective end, a user is likely to have a number of affective reactions to an agent. These reactions might manifest themselves in terms of how much users liked the agent, how intrusive they found the agent, how they perceived the agent's personality, and how willing they are to use the agent in the future. We can certainly assess a user's liking and satisfaction towards an agent, but if the user can carry out the tasks more effectively with the agent regardless of liking and satisfaction, then how important are those variables? On the other hand, long-term use of an agent might be predicted by liking and satisfaction.

The likelihood of a user following an agent's advice might be another interesting measure of the usefulness of an agent. While advice-following would certainly be at least partly a function of the quality of the advice, it will also be impacted by how the user feels about the agent (how many children ignore the advice of their parents merely because it is the parents giving the advice?).

Experiment

Overview

One fundamental issue in the quality of agent interfaces is competence (Maes, 1994). It appears obvious that perceptions of anthropomorphic agent interfaces will be strongly influenced by the competence of the supporting software system and the quality of the replies and suggestions made by the agent. We chose to factor out competence as an influence. If our experiments uncover that people's performance is not enhanced and they dislike anthropomorphic user interfaces even though the system is competent, then that is an important and strong result that other researchers and developers need to understand. To remove competence as a factor, we employed a "Wizard of Oz" (Dahlback, Jonsson, & Ahrenberg, 1993) methodology (described below).

The experiment manipulated the agent fidelity and the task objectiveness variables because prior work and our framework suggest they seemed likely candidates to have an affect on the perception of agents. Usefulness was evaluated via both the performance and satisfaction dimensions. We hypothesized that user reactions to the agent would vary as a function of the objectiveness of task. A task that required the user to debate the merits of his or her opinion (about items to pack on a trip) might lead the user to feel the agent had more of a personality (for good or for bad) compared to a task in which the user made use of the agent more as a reference tool (i.e., reminding the user of keystroke commands for a text editor). We also hypothesized that users might find the agent to be more useful in its role as a reference source rather than as an entity that provides opinions. Finally, we expected that the more life-like the agent appeared, the more likely the user might be to ascribe qualities such as personality and intelligence to the agent, but objective performance would likely not be affected by appearance.

Participants

Thirty-nine undergraduates participated for course credit and were randomly assigned to conditions. Participants had a variety of majors and computer backgrounds.

Procedure and Design

Participants were run individually using a computer equipped with a microphone and speaker. Participants performed two tasks: a travel task and an editing task. The travel task was chosen to be a type of creative, opinion-based task in which interacting with an agent might be viewed as an opportunity to think more deeply about the task by discussing points of view about the importance of travel items. The editing task was chosen to represent an opportunity to use an agent primarily as a reference source rather than as a guide or teacher.

The travel task involved a hypothetical situation in which the participant had a friend who was flying overseas on his first international trip. The task was to recommend six items for the person to take with him from a pool of 12 items and to rank the six items in order of importance.

After the participant did the initial ranking using a simple software interface, a computer agent who supposedly had knowledge about international trips appeared. The agent made a predefined set of suggestions in which it recommended changing the rankings of four of the six choices and it agreed with the ranking of two other items. For example, the agent first suggested promoting the person's fourth item to the first position, demoting the first item but keeping it in the top six. The agent explained the reasoning for its suggestion at every stage and asked the participant what he or she thought about the suggestion. After the participant responded to the agent's comment on a particular item, the agent would say one of several conversational conventions (e.g., "OK, let's continue") so that it could move on to the next suggestion. After the agent finished providing feedback on the rankings, the original rankings were displayed on the screen and the participant was given the opportunity to change the rankings. After doing the re-ranking, participants filled out a questionnaire about the agent and were asked a few questions about the agent and related issues by the experimenter.

The editing task required participants to use an unfamiliar text editor to modify an existing document by making a set of prescribed changes to the document. Participants first viewed a short video that described the various functions (e.g., copy, paste) and the specific key combinations needed to issue the commands. Participants were then shown a marked-up document that required a set of changes such as deletions, insertions, and moves, and they were instructed that if at any time they could not remember the keystrokes for a particular function, they could ask the agent for help. Pilot testing was conducted to ensure that the number of commands was sufficiently large so that participants would be likely to need to ask the agent for help. After completing the editing tasks, participants again filled out a questionnaire about the agent and answered questions from the experimenter.

The agent was controlled through a Wizard of Oz technique. One experimenter was in the room with the participant to introduce the experimental materials, and a second experimenter was in an adjacent room, monitoring the questions and responses made by the participant. The second experimenter insured that the agent responded in a consistent manner using a prepared set of replies.

Two between-subjects variables were manipulated: type of agent (animated, stiff, iconic) and task order (travel task then editing task or vice versa). The left side of Figure 1 shows the face of the agent in the animated and stiff conditions. The animated agent (donated by Haptek Corp.) was 3D, with a female appearance--though somewhat androgynous--that blinked, moved its head, and produced certain facial expressions in addition to moving its mouth in synchronization with the synthesized voice. The stiff agent had the same face as the animated agent but moved only its mouth. The iconic agent (see the right side of Figure 1) was a light-bulb icon that had arrows appear whenever it spoke.

One design issue about this experiment should be flagged. Although our key task manipulation was the "objectiveness" of the task (i.e., the travel task being less objective and the editing task being more objective), the nature of the agent also was varied as a function of the task. The agent was completely reactive in the editing task; it provided information only when requested. However, in the travel task the agent provided feedback regardless of the participants' desire. A cleaner version of the experiment would have been to hold the "nature" of the agent constant across the tasks. We allowed this confounding to occur here because were interested in getting participants' reactions to certain human-like attributes of the agent but did not have the resources to run the additional conditions that would have been required to completely cross this factor with the task and appearance manipulations. In future work we plan to systematically investigate this reactive/proactive dimension.

Figure 1: Appearance of Agent in Animated and Stiff Conditions (left) and Iconic Condition (right).

Measures

Both objective and subjective measures were used. One objective measure was, for the travel task, whether

participants changed their rankings as a function of the agent's feedback. For the editing task we measured how long it took participants to complete the edits. The primary subjective variables were the responses to the individual items in the questionnaires and the answers to the questions posed by the experimenter. The questionnaire items used a five-point Likert scale (1 = strongly agree, 5 = strongly disagree) that addressed a number of qualities of the agent (see Table 2). The questions posed by the experimenter were open-ended and provided participants an opportunity to give their impressions about the agent's personality, helpfulness, and intelligence.

Results

In the data analyses we found that the task order manipulation did not have an effect, so in the interest of simplicity we will collapse across that factor in the presentation and discussion of the results.

Performance Measures. With respect to more objective measures, Table 1 shows that participants were more likely to change the rankings of items that the agent disagreed with compared to items that the agent agreed with, $F(1, 36) = 38.48$, $MSE = .07$, $p < .0001$). There was no effect of type of agent, $F(2, 36) = 0.9$, $MSE = .11$, $p = .42$. There was no interaction, $F(2, 36) = 1.25$, $p = .30$.

Table 1: Proportion of Travel Items with Changed Rankings as a Function of Type of Agent & Agent Advice.

	Animated (n =14)	Stiff (n =12)	Iconic (n =13)
Suggested Change	.82	.90	.77
Keep Rank	.57	.42	.38

The time (in seconds) to do the editing task did not differ as a function of agent (animated: 714.8, stiff: 568.7, iconic: 671.1); $F(2, 31) = 1.78$, $MSE = 37637.22$, $p = .19$ (5 participants did not do the editing task).

Questionnaire Responses. Table 2 shows the mean responses to the questionnaire items for the different agent conditions after the travel and editing tasks (there were 5 participants who did not do both tasks and they are excluded from Table 2). There was no effect of agent type for any of the questions. For two of the items, worthwhile and intrusive, there was an effect of task (worthwhile: $F(1, 31) = 15.68$, $MSE = .45$, $p = .0004$; intrusive: $F(1, 31) = 20.28$, $MSE = .23$, $p = .0001$). The agent was rated more worthwhile and less intrusive after the editing task compared to the travel task. These results make sense. First, the editing task required most participants to rely heavily on the agent to remind them of commands, thus making the agent seem worthwhile. Second, the uninvited critique of participants' rankings of travel items could certainly have seemed intrusive.

While group differences did not exist on most of the questionnaire items, it is interesting that for most items, the average response tended to be in the positive direction. Participants felt positively, on average, about the agent.

Interview Responses. While participants made a number of interesting and insightful comments about the agent in response to questions from the experimenter, a simple tally of responses shows reactions to the agent that again varied as a function of task. Virtually all participants found the agent helpful for both tasks. Participants were much less likely to consider the agent to have a personality after doing the editing task compared to the travel task. This makes sense because the agent was merely providing subjects with information on commands in the editing task. In the travel task the agent expressed its "opinions."

Finally, it is worth noting that in general the agent was perceived as more intelligent after the travel task than after the editing task. At one level this seems odd because the agent had all the answers for the editing task. However, as demonstrated by some participants' comments, the agent was perceived as very limited in the editing task; it knew about editing commands and probably little else (despite the fact that it also appeared to understand spoken language!). In the travel task though it presumably gave the impression of having sufficiently deep knowledge about travel such that it could give feedback on the importance of various items one might take on a trip. While some of the participants' responses to the agent indicated that they disagreed with its suggestions, they appeared to believe that the suggestions were at least thoughtful.

Discussion

In addition to the results reported above, we learned a great deal by observing participants' behaviors and responses in the sessions. One key question we had was how would the participants interact with the agent in the two different tasks. In the editing task, participants seemed very comfortable asking the agent for assistance. Participants requested help an average of 6.5 times. However, in the travel task participants seemed reluctant to engage the agent in a dialog. Only a few replied with more than a few words when the agent engaged them. There was clearly awkwardness to the interaction.

The agent's social abilities and personality (or lack thereof) were noted by a number of the participants. In the travel task, we intentionally had the agent begin the session saying, "Hello, [person's name]." Three participants explicitly mentioned this feature, one stating, when asked if the agent had a personality, "Yes, respectful. It said, '[my name]', and 'I agree with this.'...I thought that was very funny. That was really cool."

Other comments implying a personality included, "Seemed a lot like a travel agent that was in a hurry," and "helpful, but kind of annoying," and "he seemed almost irritated when I didn't agree with him." One participant who did the editing task first, stated after the task that the agent did not have a personality, "It was just directed at answering questions. It had no inflections." But when asked again after the travel task, the participant responded, "It was still mechanical, but you could feel the attempt at being more personable. It acknowledged my responses, asking me to elaborate. The responses were at a more

personal level." Participants' willingness to ascribe a personality to the agent based on a few comments by the agent in one task suggests that people might be predisposed to "finding" a personality in an agent. If the effects of seeing a personality in an agent can be better understood, such a predisposition might be exploited for good purpose by designers.

Conclusion

Anthropomorphic interface agents might be one of the best interface approaches ever devised. Or they might not. Equivocal results from prior research make it virtually impossible to decide this matter. The difficulty with prior work has been its lack of systematicity in examining key factors and the use of dependent measures that often did not appropriately assess subjective experience and objective performance.

In this paper we introduced a framework for systematically examining the effects of anthropomorphic agents on user performance and subjective responses. We performed an initial experiment within this framework that suggested that type of task may play an outsized role in the perception of agents. We plan to use our framework to guide additional studies and hope other researchers find it useful and that it will allow future experiments to build on each other more effectively than in the past.

Table 2: Responses to Questionnaire Items as a Function of Type of Agent and Task.

	Animated (*n*=12)		Stiff (*n* =12)		Iconic (*n* =10)		AVG
Agent was	Travel	Edit	Travel	Edit	Travel	Edit	Travel/Edit
Worthwhile	2.50	1.58	2.25	1.42	2.30	2.10	2.35/1.57
Intrusive	2.83	3.50	3.50	4.00	3.40	3.80	3.24/3.76
Friendly	2.67	2.67	2.42	2.50	2.40	2.80	2.50/2.65
Annoying	3.25	3.33	2.83	3.25	3.20	3.80	3.09/3.44
Intelligent	2.58	2.92	2.58	2.50	2.40	2.70	2.53/2.71
Cold	3.25	3.08	3.00	2.67	3.70	3.30	3.29/3.00
Agent has clear voice	2.33	2.58	2.58	2.33	2.50	2.40	2.47/2.44
Enjoyed interacting with Agent	3.08	3.17	2.75	2.83	2.70	2.90	2.85/2.97
Agent helped with task	2.25	1.50	1.67	1.50	2.00	2.30	1.97/1.74
Like to have agent	2.83	2.67	2.58	2.33	2.20	2.40	2.56/2.47

Note: Responses were on a scale from 1 (strongly agree) to 5 (strongly disagree).

References

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM, 43*, 70-78.

Dahlback, N., Jonsson, A. & Ahrenberg, L. (1993). Wizard of Oz studies — why and how. In *Proceedings of the 1993 International Workshop on Intelligent User Interfaces*, (Orlando, FL), 193-200.

Dehn, D.M. & van Mulken, S. (2000). The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies 52*, 1-22.

Fridlund, A.J. & Gilbert, A.N. (1985). Emotions and facial expression. *Science, 230*, 607—608.

King, W.J. & Ohya, J. (1996). The representation of agents: Anthropomorphism, agency and intelligence. In *Proceedings CHI ' 96 Conference Companion*, 289-290.

Koda, T. (1996). *Agents with faces: A study on the effect of personification of software agents*. Masters thesis, MIT Media Lab, Cambridge, MA.

Lanier, J. (1995). Agents of alienation. *Interactions 2*, 66—72.

Laurel, B. (1990). Interface agents: Metaphors with character. In B. Laurel (Ed.), *The art of human-computer interface design*. Reading, MA: Addison-Wesley, 355-365.

Lyman, P. & Varian, H. (2002). *How Much Information?* Find at http://www.sims.berkeley.edu/how-much-info/.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37*, 31-40.

McCrae, R. & Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

Nass, C., Isbister, K. & Lee, E. (2000). Truth is beauty: Researching embodied conversational agents. In J. Cassell, S. Prevost, J. Sullivan, and E. Churchill (Eds.), *Embodied conversational agents*. Cambridge, MA: MIT Press, 374-402.

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are social actors. In *Proceedings of CHI '94*, 72-78.

Rickenberg, R. & Reeves, B. (2000). The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *Proceedings of CHI 2000*, 329-336.

Shneiderman, B. & Maes, P. (1997). Direct manipulation vs. interface agents. *Interactions, 4*, 42-61.

Shneiderman, B. (1997). Direct manipulation versus agents: Paths to predictable, controllable, and comprehensible interfaces. In J.M. Bradshaw (Ed.), *Software agents*. Cambridge, MA: MIT Press, 97-106.

Takeuchi, A. & Nagao, K. (1995). Situated facial displays: Towards social interaction. In *Proceedings of CHI ' 95*, 450-455.

Walker, J.H., Sproull, L., & Subramani, R. (1994). Using a human face in an interface. In *Proceedings of CHI ' 94*, 85-91.

The Effect of Goal Constraints on Strategy Generation

Suzanne C. Charman (CharmanSC1@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff CF10 3YG, Wales, United Kingdom

Andrew Howes (HowesA@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff CF10 3YG, Wales, United Kingdom

Abstract

Given practice, people generate new more efficient strategies for achieving desired goals. However, some researchers have observed that even experienced users of computer systems persist with relatively inefficient strategies. One reason for these findings may be a reduced opportunity to use efficient strategies in tasks where higher goal constraints are present. In this study half of the participants completed a drawing task in Microsoft PowerPoint in which they had to design the layout of a computer room and study area; the other half completed an equivalent drawing task that involved no higher goal constraints. Those with higher goal constraints were slower to generate more efficient strategies. This can be accounted for by a reduced opportunity to use 'efficient' strategies. Experience in other computer packages and strategic knowledge also influenced strategy generation.

Introduction

When people learn a new skill they often move through a series of progressively more efficient strategies. Practicing a task does not only result in faster performance, but also leads to the generation of new strategies (Delaney, Reder, Staszewski and Ritter, 1998; Charman and Howes, 2001). For example, Charman and Howes (2001) found that people can successfully generate more efficient drawing strategies when using Microsoft PowerPoint as a result of practice on component procedures.

However, Carroll and Rosson (1987) observed that the skills of computer users "tend to asymptote at relative mediocrity" (p.1). Similarly, Bhavnani and John (1997) reported that even after a number of years of experience and formal training in a computer aided design package, many users had not adopted more efficient strategies. The reason, they suggest, was not related to the standard of interface design or experience with the package, but to an absence of strategic knowledge. Bhavnani, John and Flemming (1999) found that people stay with inefficient methods unless they are taught efficient strategies explicitly.

One explanation for the conflicting observations of Charman and Howes (2001) and Bhavnani and John (1997) is that participants had different primary goals. Whereas the task used by Charman and Howes (2001) involved participants reproducing simple pictures in Microsoft PowerPoint, Bhavnani and John (1997) observed CAD users completing real work tasks, which would have imposed higher goal constraints. It is possible that the presence of higher goal constraints inhibits the generation of more efficient strategies and/or reduces the opportunity to use them. Higher goal constraints when preparing a report or presentation, for example, might concern syntax and semantics, and when designing a building they might concern functionality and aesthetic quality.

Higher goal constraints may hinder strategy generation by changing the user's focus. This is consistent with the observations of Carroll and Rosson (1987). They found that people were unwilling to take time out to read a manual because they were 'end-product' focused, i.e. their paramount concern was with completing the tasks at hand. It is possible that Bhavnani and John's (1997) participants failed to generate more efficient strategies because they were focused on meeting higher goal task constraints derived from the work domain. The focus in Charman and Howes' (2001) study however was on the method for which more efficient strategies were available.

In addition, the presence of higher goal constraints may reduce the opportunity to use efficient strategies. When taking into consideration higher goal constraints, sub-goals tend to be smaller, and so strategies that exploit the iterative power of the computer package are not as beneficial. When working with higher goal constraints computer users may generate strategies that are efficient given the sub-goal structure of the task, but which appear inefficient when viewed from the perspective of the end product. E.g. It is possible to imagine an efficient way of drawing a *given* floor plan, but when a person is *designing* a plan they do so interactively, using the device as repository for partial solutions.

While, substantial efforts have been made to model strategy change (e.g. Shrager and Siegler, 1998), these models do not address the issue of *when* people deploy strategy generation mechanisms. These models instead have addressed details of the mechanisms by which new strategies are generated from existing strategies. For example, Crowley, Shrager and Siegler (1997)

proposed that people use both a metacognitive and an associative mechanism. The metacognitive mechanism is of particular interest here because it requires deliberate and resource intensive problem solving. Our interest in this paper is in the extent to which higher goal constraints moderate ability or opportunity to beneficially deploy metacognitive problem solving.

We predict that there will be a negative impact of higher goal constraints on the generation of efficient strategies. In the following experiment, whether or not participants had a higher level goal to meet was manipulated. The higher-goal task was to design the layout of a computer room and study space. In the no-higher-goal task participants copied and pasted an equivalent number of computers and desks into a large blank area. To complete the tasks a range of strategies varying in efficiency, with the same component procedures, could be used. Participants could copy and paste just one item (a computer or a desk) at a time, or could copy and paste multiple items at once. Previous experience and strategic knowledge were also examined as factors affecting strategy generation.

Method

Participants

Twenty-four undergraduates who were regular computer users, ranging in age from 18 to 26, took part in the experiment for 1½ hours of course credit or for payment of £6. All participants were given the same amount of credit or payment to take part in the study, no matter how long they took, in order to encourage efficient completion of the tasks.

Design

The study involved three between-subjects factors. The first was task type. In one condition participants were given a higher goal, where they were asked to design the layout of a computer classroom and a study area (see Appendix I). This higher goal gave rise to several design constraints that determined the manner in which the desks and computers could be arranged. The goal for these participants was to take into consideration the constraints outlined and also to consider the best use of space. In another condition participants did not have a higher goal in mind, they were asked to copy and paste an equivalent number of computers and study desks into a large blank space. A median split (over both task type conditions) on two pre-test measures, experience and strategic knowledge, created two more between subjects factors.

Procedure and Materials

The participants completed an informed consent form and then a short online spatial IQ test (Crampton and Jerabek, 2000), which consisted of ten questions, giving a score out of 100. Participants were then asked to complete a short questionnaire that asked about prior experience with Microsoft PowerPoint, as well as other software packages with drawing functions and Microsoft Word. The tuition phase was then completed, which ensured that the participants mastered basic drawing skills (drawing, moving, altering, fencing to select, copying and pasting a single shape). The participants were informed that they should only use functions identified in the tutorial stage. These included fencing, copying and pasting, but, for example, excluded duplication and grouping.

After the tuition phase the participants completed an open-ended questionnaire designed to assess knowledge about the device. Ten questions relevant to the key concepts particular to working with more than one item at a time were included. Five questions related to fencing multiple shapes with space between them and five related to the manipulation of multiple shapes.

The participants then completed a pre-test stage where they were asked to draw eight 2-shape items in as few moves as possible (Figure 1). The strategy used by the participant was coded and scored (1-7) according to the relative efficiency of the strategy based upon the coding framework outlined in Charman and Howes (2001). For example, participants were given a score of 1 if they drew each shape one by one, and a score of 7 if an exponential copying strategy was used (exponentially increase the number of items made each time copy and paste are used). This score was taken as a measure of strategic knowledge.

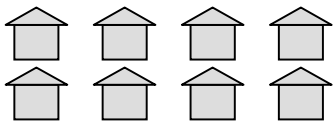


Figure 1: Pre-test task.

The main task of the experiment was then conducted. Participants were informed that there was an online help facility that they could choose to consult if they wished (this was available freely to the participants and could be accessed by selecting an open Internet Explorer window). For the main task, the no-higher-goal condition participants were given a key with sample desks and computers in it, and asked to reproduce 54 study desks and 148 computers in the space provided. The participants were all instructed that they could fence (to select), copy and paste the computers and study desks provided in the key.

In the higher-goal condition, participants were asked to plan the layout of a new extension to the Psychology building (Appendix I). In the proposed extension there was a study area where study booths were to be placed, and a computer classroom where computers were to be

placed. Participants were given design constraints for which visual measures were provided, such as making sure that there were gangways, access to desks and space between computers. Although participants were told that a design could include 148 computers and 54 desks, the task was over when the participant felt they had finished their design, with the constraints met.

Participants were instructed to complete the task in as few moves as possible. Finally participants filled in another device representation questionnaire. Microsoft PowerPoint 97 was used to carry out the drawing tasks.

Strategies

To complete the task, several strategies could be employed. It was possible to work with an individual shape, as the composite parts of each item (a computer or desk) were not grouped together. A better way to complete the task was to draw a fence around one item and then copy, paste and move the item. An even better strategy was to work with more than one item at a time. In order to do this a participant needed to know that multiple items with space between them could be selected at once (by drawing a fence around them, see Figure 2) and then manipulated (copied, pasted and moved) simultaneously. Finally, the exponential copying strategy allowed very fast completion of the task. Here the number of copies produced at once increases exponentially.

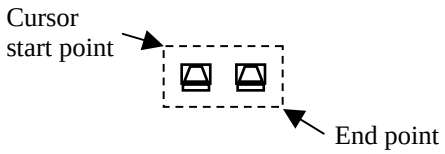


Figure 2: Fencing, by using the mouse to click and drag from the start point and releasing at the end point.

Results

Each move made by a participant was recorded, allowing a fine-grained analysis of performance. An individual move was taken to be either a key-stroke (e.g. delete) or a purposeful mouse-click (e.g. copy or select shape). Creating a fence was also counted as one move (although to do this one must click and drag).

Using efficient copying strategies did save time. The total time taken was negatively correlated with the proportion of moves where multiple items were worked with at once (r_s=-0.672, p<0.001). Only three participants visited the on-line help facility, for less than forty seconds each, so these results are not included in the analyses.

For analysis, participants were split into high and low experience groups based upon the experience questionnaire data, and also high and low strategic-knowledge groups based upon the pre-test strategy score achieved. The main analysis used was a between subjects 2x2x2 ANOVA with task type, experience and strategic knowledge as factors.

Task Type

Total Moves Those in the no-higher-goal condition (M=116.8) made fewer moves in total than the higher-goal condition (M=323.2) [F(1,16)=51.168, p<0.001, MSE=7916.1], and took fewer moves to make each item (M=0.6) than those in the higher-goal condition (M=1.9) [F(1,16)=37.444, p<0.001, MSE=0.4].

Excess Moves as a Proportion of Total Moves The fact that there were fewer moves made in the no-higher-goal condition may have been due either to reduced opportunity or to task focus. In order to further investigate strategy change as a function of opportunity we analyzed the excess moves as a proportion of total moves. For each task type the mean optimal number of moves was calculated (145 for higher-goal; 46 for no-higher-goal) and subtracted from each participants total number of moves to give the excess moves. The excess moves as a proportion of the total moves made was then calculated for each participant. There was no significant difference between the higher-goal group (M=0.45) and the no-higher-goal group (M=0.49) [F(1, 16)=0.363, p=0.555, MSE=0.05].

Strategy Generation Higher goal constraints impacted upon how soon strategies were generated. Participants occasionally started to complete the task by working with individual shapes. Most however started working with one item (computer or a desk) at a time. A better strategy was to work with more than one item at a time. The move on which this strategy was first used was recorded. The higher-goal condition (M=197.9) worked with more than one item significantly later on than those in the no-higher-goal condition (M=41.0), F(1,16)=18.729, p<0.001, MSE=17227.6.

The final progression in strategy use was to use an exponential copying strategy. A main effect of task type on the move when this strategy was first used was found [F(1,16)=13.820, p<0.01, MSE=21072.3]; those with a higher goal (M=229.8) generated the strategy later on than those with no higher goal (M=85.2).

However, while in both tasks the earliest opportunity to use each of the strategies was the same (move 7), the overall opportunity to use the strategies differed between tasks. These results may therefore reflect either reduced opportunity or a different task focus.

Experience

At the start of the experiment participants had either no experience, or very little experience, with the drawing functions in Microsoft PowerPoint.

Table 1: Interaction between task type and experience.

	High-Experience		Low-Experience	
Measure	*Higher-Goal*	*No-Higher-Goal*	*Higher-Goal*	*No-Higher-Goal*
Total time taken to complete the task	1060.5	384.3	1630.7	458.2
Total moves taken to complete the task	279.0	116.0	750.0	118.0

Performance A median split placed participants in either a high-experience or low-experience group, based upon their experience questionnaire score.

A main effect of experience on time taken to perform the task was found [$F(1,16)=18.102$, $p<0.001$, $MSE=57457.4$]. Unsurprisingly those with high experience ($M=696.4$s) performed the task faster than those with low experience ($M=1097.7$s). High-experience participants ($M=191.2$) also performed the task using fewer moves than those in the low-experience group ($M=254.0$), $F(1,16)=9.395$, $p<0.01$, $MSE=7916.1$.

Strategy Generation More interestingly, experience had an effect on how soon efficient strategies were generated. There was a main effect of experience on the move participants first worked with multiple items [$F(1,16)=7.024$, $p<0.05$, $MSE=17227.6$]. High-experience participants generated this strategy ($M=87.1$) earlier than those who had low experience ($M=157.7$). Those with high experience ($M=122.2$) also generated the exponential copying strategy earlier than those with low experience ($M=199.3$), $F(1,16)=5.856$, $p<0.05$, $MSE=21072.3$.

Strategic Knowledge

Performance A median split placed participants in either a high-strategic-knowledge or low-strategic-knowledge group, based upon their pre-test strategy score. A main effect of strategic knowledge [$F(1,16)=19.321$, $p<0.001$, $MSE=7916.1$] found that high-knowledge participants ($M=199.3$) performed the task using fewer moves than those in the low-knowledge group ($M=240.8$).

Interactions

There was a significant interaction between experience and task type for the total time taken to perform the task [$F(1,16)=12.906$, $p<0.01$, $MSE=57457.4$] and also for the total number of moves taken to perform the task [$F(1,16)=7.956$, $p<0.05$, $MSE=7916.1$] (see Table 1). Simple effects tests revealed that where experience was low, the presence of higher goal constraints had an effect on the time taken [$F_B@a2(1,16)=6.575$, $p<0.05$] and moves made [$F_B@a2(1,16)=11.457$, $p<0.05$]. Simple effects tests also found that experience had a greater effect on time taken [$F_A@b1(1,16)=8.350$, $p<0.05$] and moves made [$F_A@b1(1,16)=12.382$, $p<0.05$] where participants were given a higher goal. However, this interaction may have been due to a ceiling effect in the performance of the no-higher-goal condition.

Similarly there was an interaction between task type and strategic knowledge (see Table 2) for the number of moves taken to complete the task [$F(1,16)=7.544$, $p<0.05$, $MSE=7916.1$]. Simple effects tests revealed that strategic knowledge had a greater effect where participants were given a higher goal to consider [$F_A@b1(1,16)=4.509$, $p<0.05$]. However, again this interaction may be due to a ceiling effect.

Table 2: Total moves taken to complete the task.

	High-Strategic-Knowledge	Low-Strategic-Knowledge
Higher-Goal	262.6	444.3
No-Higher-Goal	72.5	139.0

Spatial IQ

A regression found that spatial IQ had a significant influence on the total number of deletes and undos used by a participant, ($\beta=-0.477$, $p<0.05$). This suggests that those with a high spatial IQ perform the task more accurately than those with a low spatial IQ, and therefore do not need to undo or delete as often. Those with a high spatial IQ may be better able to plan their actions, and so make fewer mistakes.

Mental Representation of the Device

The amount of experience a participant had on other drawing packages had a significant influence on their first device representation questionnaire score ($\beta=0.554$, $p<0.01$). A regression found that a participant's score on the first device representation questionnaire (DRQ) had a significant influence on the time taken to perform the pre-test ($\beta=-0.484$, $p<0.05$). The score that participants gained on the first DRQ also exerted influence on early improvement in the number of moves made to make each item in the main task ($\beta=0.611$, $p<0.01$).

From these data we can suggest that previous experience allows more accurate hypotheses about the operation of the device to be developed while the participant answers the questionnaire. This representation then supports the generation of faster and more efficient methods.

Case Study

One case study demonstrated a particularly strong effect of having a higher goal. Initially the participant had low device knowledge, but had an average spatial IQ and previous experience with computer packages. The participant completed the pre-test task very quickly and used a very good strategy. The strategy used in the pre-test involved the participant fencing, copying and pasting four items at once. However during the main task where the participant had to design the layout of the computer room and study areas, he did not use this strategy or the exponential copying strategy. Instead he fenced, copied and pasted each item one by one, this taking him 417 moves (M=220) and 1465 seconds (M=880). In this case it seems that the presence of a higher goal actually inhibited the use of a known and previously used strategy.

Discussion

When higher goal constraints were present participants made more moves and generated new strategies more slowly. Those with higher goal constraints made at least four times as many moves before generating more efficient strategies. In addition, those with low strategic knowledge or experience suffered diminished performance and took nearly twice as long to generate efficient strategies.

Our analysis indicated that the effect of higher goal constraints was entirely due to the way in which the design task reduced the opportunity for the use of the more efficient strategies. Once opportunity was accounted for, higher goal constraints had no significant effect on the number of moves made. This suggests that higher goal constraints might not change the ability of a user to generate an efficient strategy, rather they may change the problem such that the opportunity to use efficient strategies is reduced. Users with a higher goal may have demonstrated adaptivity to opportunity.

As opportunity could account for the difference in performance between the higher-goal and no-higher-goal conditions, we found no evidence that a higher work goal might inhibit strategy generation. We found no support for the hypothesis that users become so focused on meeting higher goal constraints that they do not concern themselves with the efficiency of the methods by which they complete the task. However, further study is required to assess the extent to which strategy generation might be inhibited by focus on higher goal constraints when opportunity is held constant.

We also found no evidence that higher goal constraints inhibited users from taking time out to learn about the device (following from Carroll and Rosson, 1987), as participants very rarely used the on-line help and all groups concluded the experiment with similar levels of device knowledge. More importantly, all the strategies were composed of the same known component procedures.

While the rate at which participants generated new strategies was slowed by a reduction in opportunity in the higher-goal condition, most participants showed a marked improvement in the efficiency of the strategies that they were using as the experiment progressed. Further, as participants made little use of the on-line help and did not stop performing the task to explore the package, the acquisition of device knowledge must have occurred while the task was being completed.

Together, our results suggest that it may be necessary to qualify the claim that people are unwilling to take time out learn (Carroll and Rosson, 1987). Our findings, while laboratory bound, indicate that people are willing to invest in the generation of more efficient strategies within the bounds of what they discover while using the device. They may not go to a manual, but they do think about the way that they achieve tasks, they do attempt to explain what they observe, and they do adapt their methods accordingly.

Finally, our findings suggest that Bhavnani, John and Flemming's (1999) conclusion that people do not generate efficient strategies without instruction may be premature. Our participants generated efficient strategies within the bounds of what the higher goal constraints allowed. These findings suggest that it may be beneficial, instead of teaching strategies explicitly, to encourage strategy generation during task performance. While Bhavnani, John and Flemming (1999) argue that strategies need to be taught, it may be better, in the long term, to ensure that users actually generate the strategy themselves. Evidence in the psychological literature suggests that there are substantial advantages to self-generation and self-explanation (Chi, Bassok, Lewis, Reimann and Glaser, 1989; Bielaczyc, Pirolli and Brown, 1995).

References

Bhavnani, S. K., & John, B. E. (1997). From sufficient to efficient usage: An analysis of strategic knowledge. *Proceedings of CHI '97*, 91-98.

Bhavnani, S. K., John, B. E., & Flemming, U. (1999). The strategic use of CAD: An empirically inspired, theory-based course. *Proceedings of CHI '99*, 183-190.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*(2), 221-252.

Carroll, J. M., & Rosson, M. B. (1987). The paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction*. Cambridge, M.A.: The MIT Press.

Charman, S.C., & Howes, A. (2001). The effect of practice on strategy change. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, (pp. 188-193). Mahwah, NJ: Erlbaum.

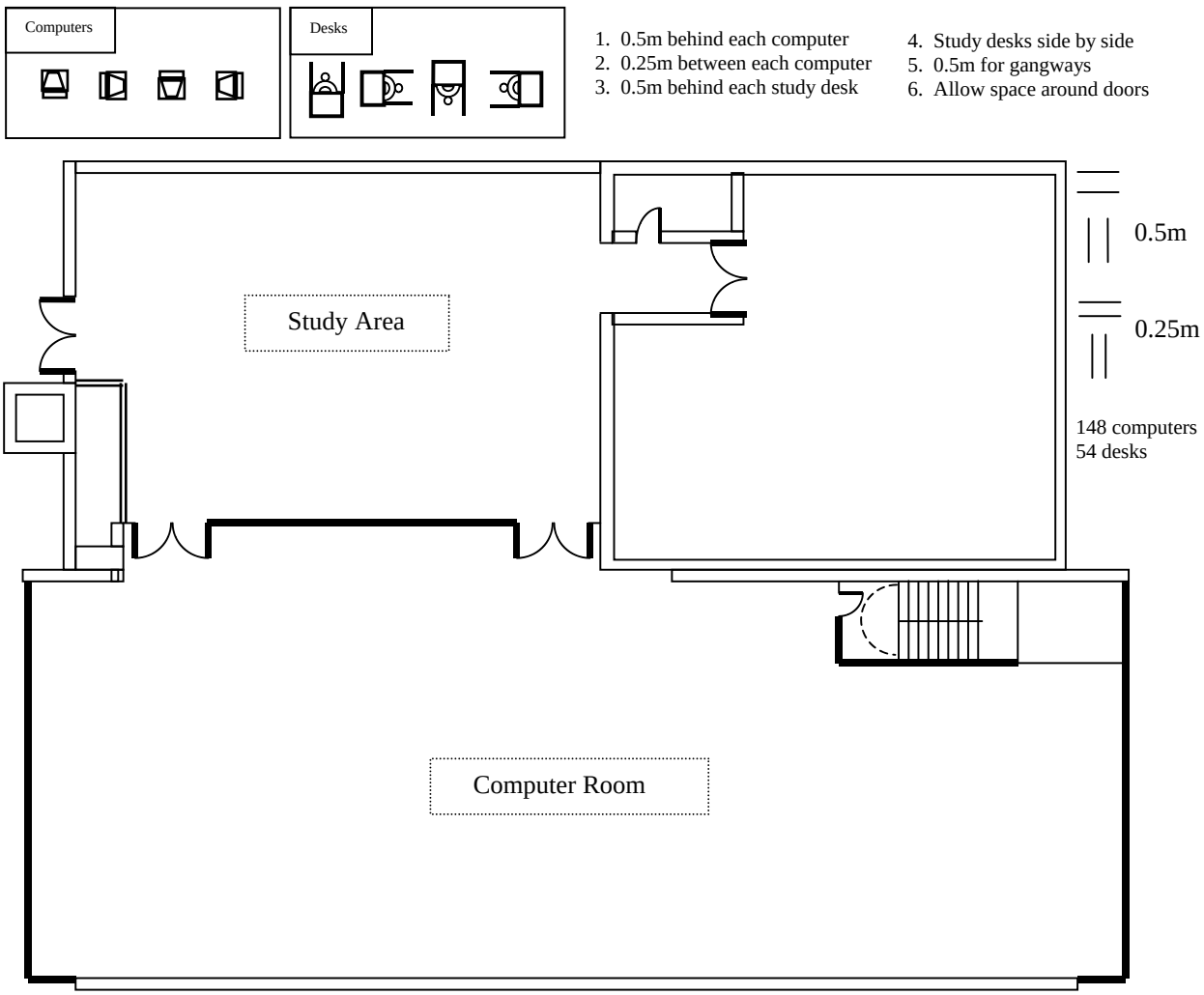
Chi, M.T.H., Bassok, M., Lewis, M., Reimann, P. & Glaser, R. (1989). Self –explanations: how students study and use examples in learning to solve problems. *Cognitive Science, 15*, 145-182.

Crampton, A., Jerabek, I. (2000). Spatial IQ Test. Available: http://www.queendom.com/tests/iq/spatial_iq.html (31.01.02).

Crowley, K., Shrager, J., & Siegler, R. S. (1997). Strategy discovery as a competitive negotiation between metacognitive and associative mechanisms. *Developmental Review, 17*, 462-489.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*, 1-7.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9*(5), 405-410.



Appendix I: The higher goal task. Participants were required to plan the layout of the computer room and study area using the items provided in the key.

Diagnosticity in Category Learning by Classification and Inference

Seth Chin-Parker and **Brian H. Ross**
(chinpark@s.psych.uiuc.edu) (bross@s.psych.uiuc.edu)
Beckman Institute, 405 North Matthews Avenue
University of Illinois, Urbana-Champaign
Urbana, IL 61801 USA

Abstract

Categories are learned in many ways, but the focus of much category learning research has been on classification learning. In classification learning, the diagnosticity of features is a primary influence on learning and the category representation. In this paper, we assess this influence of diagnosticity on a different means of category learning, inference learning. In two experiments, each with a different dependent measure, we found the expected result that classification learning led to strong sensitivity to the diagnosticity of the features, even to the exclusion of prototypicality (when controlled for diagnosticity). However, inference learners were significantly less sensitive to the diagnostic value of the features, and were sensitive to the prototypicality. This result provides further evidence for the idea that different types of category learning differentially influence the category representation and provides a better understanding of inference learning.

Introduction

Categories are critical for a wide variety of cognitive tasks, such as classification, inference, explanation, communication, and problem solving. Category learning reflects how it is that people acquire knowledge of the categories that will successfully support these uses. Any intelligent system that extends information from specific examples to other related occurrences needs to account for the processes related to category learning. Thus, developing an understanding of category learning is an important research endeavor in cognitive science.

Although categories may be learned in a number of ways, the focus of category learning research has been on classification, how items are assigned to categories (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). In classification learning, a subject is shown an item, asked to indicate its category membership (usually from a set of two possible categories), given feedback on their choice, and then allowed to study the item before the next item is presented. Through the learning trials of this classification task, the subject learns what items go into what category, thus developing a category representation that can be used later to answer questions about the category members or to classify novel items.

Category learning, however, does not just consist of classification learning. We learn categories for a variety of purposes, and how we learn categories is often tied to these uses. Category learning is based on not just classification, but on inference or explanation or problem solving (among other possibilities). A complete understanding of category learning requires considering additional category learning tasks and how they influence the category representation. The idea underlying this research is that different ways of learning about categories lead to different category representations, and that our real-world representations often derive from a variety of ways in which categories are learned and used. Although this notion has a certain intuitive appeal, only a small number of category learning studies have examined it.

One category learning task that has received attention over the last few years is inference learning (Anderson, Ross & Chin-Parker, 2002; Yamauchi, Love & Markman, 2002; Yamauchi & Markman, 1998, 2000). In this task, a classified item is presented with a category feature missing and the task of the learner is to choose the appropriate feature for that item. For example, if one were making inferences about different types of birds, one might be given a classified bird (e.g., yellow-rumped warbler) with a number of its features and asked to choose its food preference. Inference is a critical component of category use. Since people learn about categories as they make inferences and receive feedback on their predictions, inference learning is a natural direction to follow in category learning research. This task has also been the focus of recent research because it has many similarities to classification learning and is formally equivalent to classification if the category label is treated simply as another feature (Yamauchi & Markman, 1998).

Diagnosticity

A critical aspect of current theories of classification learning is the focus on the *diagnosticity* of the features, how predictive they are of category membership (Tversky, 1977). As people learn to classify category members, they learn to attend more to those features that help to distinguish the categories (e.g., Kruschke, 1992). In the simple case in which the categories may be distinguished on the basis of a single feature, all the attention may be focused on that feature. For more complex cases, the attention is distributed across the diagnostic features to maximize classification performance.

Inference learning, however, may not lead to such an exclusive attention to diagnostic features. In inference learning, the item is already classified, so the learner can focus on *that* category and what features occur with members of that category. This focus on a single category at a time makes information about the prototypical feature values more available (Anderson et al., 2002; Yamauchi & Markman, 1998, 2000). This information about the prototypical feature values for each category means that the category representation emphasizes the internal structure of the category, what it is that coheres the members of the

category. This focus during inference learning suggests that different information about the category would be acquired during learning when compared to classification learning.

Current Experiments

The goal of the current experiments is to investigate the role of diagnosticity in category learning with two different category learning tasks, classification and inference. Based on a large body of previous research, we expect that diagnosticity will be the primary influence in classification learning. Thus, the question of most interest is how inference learning is affected by feature diagnosticity (versus the internal structure of the category). The hypothesis is that inference learning will not lead to as strong an influence of diagnosticity as does classification learning. This hypothesis is of importance for two reasons. First, it questions a major assumption of current models of category learning, that the diagnosticity of features is the most important determinant of category learning. Second, it helps to provide a further understanding of another type of category learning, inference learning.

We used a common category structure, family resemblance, as shown in Table 1. In this structure, the prototype is chosen and the learning items from that category consist of items that are similar to the prototype, though they may be different from one another. In the experiments reported here, all the learning exemplars match the prototype on all but one of the features.

The diagnosticity of the features is manipulated by varying the overlap of the prototypes. With this manipulation, to be described in detail for each experiment, we could separately vary the prototypicality of the item (i.e., how similar it was to the prototype, reflecting the internal structure) and the diagnosticity of the features (in terms of how predictive they were of category membership, reflecting the relation of the two categories).

Based on results from previous studies, we can anticipate some of the results of these experiments. Classification learners should show a strong effect of diagnosticity, but not prototypicality. We also know the inference learners should show a sensitivity to prototypicality. The question remains to be answered as to whether the inference learners will show any effect of diagnosticity. If they do show a sensitivity to diagnosticity, it should be significantly less than that of the classification learners.

Experiment 1

Experiment 1 investigated the influence of diagnosticity on classification and inference learning with a forced-choice test at transfer. The critical test trials varied the diagnosticity and typicality to examine the influence of the learning conditions. The categories learned were fictional "bugs".

The manipulation of diagnosticity as a function of prototype overlap can be seen in Table 1. The target category is the one on the left, the Deegers (prototype 11111, indicating a particular set of values for each of the five binary dimensions). Along with this category, subjects either learned the Lokads (prototype 00011) or the Koozles (prototype 11000). Those features common to both prototypes are not diagnostic because they do not help one to determine category membership. As can be seen, both contrast categories overlapped Deegers on two features, but two different features. Thus, for subjects learning Deegers and Lokads, the last two features were not diagnostic, whereas the first two were not diagnostic for the subjects learning Deegers and Koozles. By varying the test features for Deegers, we can determine the extent to which diagnosticity is being used, as described shortly.

Table 1: Category Structure for Experiment 1.

	Deeger	Lokad	Koozle
Learning Exemplars	11110	00010	11001
	11101	00001	11010
	11011	00111	11100
	10111	01011	10000
	01111	10011	01000
Prototype	11111	00011	11000

Method

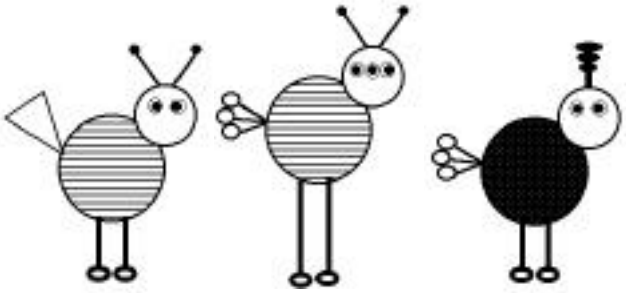
Design There were two learning conditions, classification learning and inference learning. Within each learning condition half of the subjects learned about Deegers and Lokads, and the other half learned about Deegers and Koozles. The position of the bug attributes within the category structure was balanced across subjects. This resulted in a total of ten experimental groups. Within the study and transfer blocks, items were randomly presented.

Subjects Fifty undergraduates from the University of Illinois participated for either course credit or pay. Ten subjects who did not meet the learning criterion (two classification, eight inference) were replaced.

Materials The materials were drawings of "bugs" labeled Deegers, Lokads or Koozles (see Figure 1). These bugs varied along the binary attributes of their antenna, legs, tail, body coloring and eyes.

Each subject learned two categories, the target category (Deegers) and one of the two non-target categories (either Koozles or Lokads). The abstract structures for the three categories can be found in Table 1. Each of the non-target category prototypes overlapped with the target category prototype on two features. The features that overlapped were

Figure 1: Example Prototypes for Bug Categories in Experiment One



Deeger Lokad Koozle

not predictive of category membership as they occurred equally often in each category. The remaining three features were diagnostic in that they occurred 80% of the time with one of the categories. Whether a particular feature was an overlap or diagnostic feature depended on which of the two non-target categories was learned with the Deeger category. For example, the Deeger specified by 11110 was seen by all subjects during learning. For a subject that learned the Lokad category along with the Deeger category, the first feature was a diagnostic feature, in that it was consistent with the Deeger prototype but not the Lokad prototype. For a subject that learned the Koozle category along with the Deeger category, the first feature was an overlap feature because it occurred in both the Koozle and Deeger prototypes. This design of the category structures allowed the same item to vary in terms of its diagnosticity between subjects, depending only on the category set learned, while the item prototypicality remained constant.

Table 2 specifies the terminology used to describe the bugs. The bugs varied on two dimensions, their relation to the category prototype (prototypicality) and the number of diagnostic features maintained. Along the first dimension, there were three possible levels: prototype, typical and atypical, which indicated whether there were 5, 4, or 3 features consistent with the prototype. For the second dimension, the item could maintain 1, 2, or 3 of the diagnostic features. For a subject that learned the Deeger-Lokad combination, 11110 would be a typ3 Deeger: it is typical because it has four features consistent with the prototype 11111, and it has a diagnosticity of 3 because it has all three of the diagnostic features for Deeger (111--). However, for a subject that learned the Deeger-Koozle combination, 11110 would be a typ2 Deeger: again it is typical because it has four of the prototype features, but now it only has 2 of the diagnostic features (--111).

The study items were the five typical bugs for each category being learned. For the classification subjects, these bugs were always seen complete. In the inference condition, the bugs were missing one feature that was consistent with the prototype. The inference subjects also saw the two possible features that were choices for the missing feature.

The transfer items were pairs of bugs from a given category. There were two critical types that allowed an examination of the influence of diagnosticity: typ/typ and typ/atyp items. The six typ/typ transfer items were pairs of bugs that were matched in prototypicality (both were typical) but varied in the number of diagnostic features. One of the bugs in the pair would be a typ2, while the other was a typ3, the specification being dependent on which category set had been learned. The typ/atyp pairings varied both in terms of how many features were consistent with the category prototype and the number of diagnostic features. Of the six typ/atyp transfer items, three of the pairs consisted of a typ2 and an atyp3. These transfer items pitted typicality to the category prototype against the diagnosticity of the features. The other three pairs consisted of a typ3 and an atyp1. For these pairs, both typicality and diagnosticity were in agreement as to which bug was the better category member. The content of the typ/atyp pairs was also dependent on the category set learned; the typ2/atyp3 pair for someone who learned Deeger-Koozle was the typ3/atyp1 pair for someone who learned the Deeger-Lokad set, and vice-versa. The critical typ/atyp pairs are the typ2/atyp3, but the typ3/atyp1 pairs allow full counterbalancing of items.

There were nine additional filler transfer pairs in the Deeger task, as well as 12 pairs for the other category that did not address the issues of interest.

Table 2: Abbreviations for Items

	# of Prototype Consistent Features	# of Diagnostic Features Maintained
Proto	5	3
Typ3	4	3
Typ2	4	2
Atyp3	3	3
Atyp2	3	2
Atyp1	3	1

Procedure Subjects were given verbal instructions prior to the study phase. All subsequent instructions and reminders appeared on the computer screen. Learning and testing were done on Macintosh computers using Psyscope (Cohen, MacWhinney, Flatt, & Provost, 1993). All subjects were debriefed both verbally and with a written statement as to the general intent of the experiment.

In the classification condition, subjects saw a bug presented in the center of the screen. Subjects indicated which category they thought the bug belonged to by pressing the "D" key for Deeger, the "K" key for Koozle or the "L" key for Lokad. Feedback was given as to whether the choice was correct or incorrect, and the subject was shown the bug again along with the correct name to study. This study time was self-paced. The learning phase continued for a minimum of four blocks, ten exemplars per block, until the subject was able to correctly identify nine of the ten bugs within a block.

In the inference condition, the subjects were presented with an incomplete bug (one of the five attributes was absent from the bug picture) centered on the screen. The category label was presented to the left of the bug, and the two possible features for the missing attribute were presented on the right side of the screen, one above the other. The subject indicated a choice by clicking the mouse on one of the two features. The position of the correct feature was randomized across learning trials. Feedback was given, and self-paced study time was allowed. The learning criteria were the same as in the classification condition.

Following learning, all subjects completed the same forced choice test. No feedback was given to the subjects during the transfer phase. Each subject completed the target (Deeger) test first. Subjects saw two possible Deegers on the screen, one centered on the right-hand side of the screen and the other centered on the left-hand side of the screen. Below the pictures appeared the question, "Which of these bugs is most typical of a Deeger?" Once the subject clicked on one of the pictures, a box appeared around the choice and the prior question was replaced by "How confident are you of your choice?" along with a number scale going from one (guessing) to seven (very sure). The subject clicked on a

number to indicate his or her confidence. Once the target category transfer was completed, the subject was informed that the same type of questions would be asked about the other bug category. The procedure for the non-target category transfer was identical.

Results and Discussion

For the typ/typ transfer items, the mean proportion of choice for the typ3 Deeger was calculated. Also, the mean confidence scores were determined; confidence scores when selecting the typ2 were multiplied by –1 (no preference between the bugs would result in a mean confidence of zero). These results examine how important the diagnosticity was to each condition when the prototypicality is held constant. The classification learners chose the bug with more diagnostic features well above chance (m = .79), $t(19) = 4.74$, $p < .01$, while the inference learners did not (m = .56), $t(19) < 1$. These results are also reflected in the group confidence ratings. The classification learners' confidence rating (m = 3.44) was significantly greater than 0, $t(19) = 5.38$, $p < .01$, while the inference mean confidence rating (m = 0.70), was not, $t(19) = 1.22$, $p > .10$. The proportion of the classification learners who chose the typ3 bug over the typ2 bug was significantly greater than the proportion of inference learners, $t(38) = 2.70$, $p < .01$. The mean confidence score of the classification learners was also significantly different from the score of the inference learners, $t(38) = 3.19$, $p < .01$.

For the typ/atyp items, the mean proportion of times the typical Deeger was chosen over the atypical Deeger was determined. The mean confidence rating was calculated by multiplying the confidence scores when choosing the atypical bug by –1. Of interest are the bug pairs that pitted typicality against diagnosticity, the typ2/atyp3 pairs; choosing the typ2 bug meant that overall typicality to the prototype was of primary importance while choosing the atyp3 bug indicated that diagnosticity was driving the decision. The classification learners chose the typ2 bug (m = .35) marginally less than the inference learners (m = .57), $t(38) = 1.81$, $p < .10$. However, the mean confidence score for the classification learners (-2.22) was significantly lower than that of the inference learners (0.76), $t(38) = 2.49$, $p < .05$.

For both transfer measures, the typ/typ and typ/atyp items, the classification learners showed significantly more dependence on diagnostic information than the inference learners when making decisions about the category members. In the typ/typ measure, the inference learners did not show an influence of diagnosticity. The typ/atyp results are more difficult to assess in this regard since there is no clear baseline to compare performance to. These results show a clear difference in the role of diagnosticity for the two different types of category learning.

Experiment 2

Experiment 2 examined the same issues, but had three differences from Experiment 1 to increase generality and the number of critical observations. First, during transfer, a typicality rating task was used rather than a forced-choice task. Second, to allow us to use all the transfer data to test the hypothesis, we changed the design so that each category had a critical contrast (see Table 3). As before, one subject might learn Deegers and Lokads, whereas another learned Deegers and Koozles. However, by adding two more counterbalancing groups (Himlit and Lokads, Himlit and Koozles), all the categories had critical test items. Third, although the items were again fictitious bugs, new features were constructed.

Table 3: Category Structure for Experiment 2

	Deeger	Lokad	Koozle	Himlit
Learning Exemplars	11110	00010	11001	10100
	11101	00001	11010	10111
	11011	00111	11100	10001
	10111	01011	10000	11101
	01111	10011	01000	00101
Prototype	11111	00011	11000	10101

Method

Design As in Experiment 1, there were a classification and an inference condition. There were four possible category combinations a subject could learn [Deeger-Koozle, Deeger-Lokad, Himlit-Koozle, Himlit-Lokad]. Within each possible combination, there was a balancing as to the order of the categories presented during the typicality rating task. The presentation of items within both study and transfer blocks was random.

Subjects Sixty-one undergraduates from the University of Illinois participated for either course credit or pay. One subject's data were lost due to a computer error and 12 subjects (five inference, seven classification) were replaced who did not meet the learning criterion.

Materials The materials for this experiment were again drawings of "bugs" (Deegers, Himlits, Koozles and Lokads) that consisted of five binary attributes: legs, wing, eyes, antenna, and tail. Across subjects, each attribute was balanced as to whether it served as an overlap or diagnostic attribute. The bugs seen during learning were the five typical bugs (one feature of each was not consistent with the prototype) from each of the two categories being learned, and they were presented as in Experiment 1.

The typicality rating task consisted of 16 bugs for each category learned. Five of the bugs were the typical bugs (one inconsistent feature each) which had been seen during learning. The other 11 bugs were novel to the subject at the time of the typicality rating task. These bugs consisted of the category prototype along with 10 atypical bugs, each of which had two features that were inconsistent with the prototype. The levels of prototypicality and diagnosticity were specified as in Experiment 1 (refer to Table 2).

Procedure The procedures during the learning phase for this experiment were very similar to those in Experiment 1. The primary differences were that the classification subjects used the mouse to click on the category label and the study time

was restricted to two seconds per bug. The learning criteria were the same as Experiment 1.

Following the learning phase, the subjects were given instructions on the computer screen explaining the typicality rating task. During this task, a single bug appeared centered on the screen along with the question, "How typical is this bug of a [Deeger, Himlit, Koozle, Lokad]?" Underneath the picture were the numbers from one ("Not at all typical") to seven ("Very typical"). The subject indicated their rating of each bug by clicking the mouse on one of the numbers.

Results and Discussion

What influences the typicality ratings for the two learning groups: diagnosticity, prototypicality or both? The short answer is that the ratings of classification learners were influenced only by diagnosticity, whereas both diagnosticity and prototypicality affected the ratings of the inference learners.

The group means for each of the bug types are provided in Table 4. The effect of diagnosticity becomes clear when the column means are examined. Collapsing across the levels of prototypicality, it is evident that classification learners showed a large effect of diagnosticity, whereas inference learners showed some effect, but a smaller one. The ANOVA supports this interpretation with no main effect of learning, $F(1, 46) < 1$, a significant main effect of diagnosticity, $F(2, 92) = 147.87$, $p < .001$, and an interaction between the factors, $F(2, 92) = 13.17$, $p < .001$. As can be seen at the column mean level (and also within each row), the loss of a diagnostic feature reduced typicality ratings almost 1.5 (on a 1-7 scale) for the classification learners, whereas it had much less of an effect on inference learners .

The effect of prototypicality requires a more careful examination. Collapsing over the diagnosticity levels, the two learning conditions show very similar row means, supported by the ANOVA indicating that there is no main effect of learning condition, $F(1, 46) < 1$, a significant effect of prototypicality, $F(2, 92) = 102.96$, $p < .001$, and no interaction between the two, $F(2, 92) < 1$. However, the difference between conditions is hidden by the fact that prototypicality in this analysis is confounded with diagnosticity as can be seen in Table 4. For classification learners, within each column there is no effect of prototypicality. For example, as long as the item has all three diagnostic features, the rating given by the classification learners does not depend at all on whether it has the two non-diagnostic features consistent with the prototype (prototype, 5.92), just one of them (typical, 5.98), or neither of them (atypical, 5.88). A similar result is seen with the items that maintained two diagnostic features; on average, the classification learners actually rated the atypical items 0.12 higher than the typical items. However, inference learners show large effects of prototypicality at each level of diagnosticity. For the test items with all three diagnostic features, each non-diagnostic feature makes a difference of about 0.7. The inference learners dropped their typicality rating about 0.9 when a non-diagnostic feature was removed from the items that maintained two diagnostic features.

Thus, in Experiment 2, the typicality ratings of classification learners were influenced only by diagnosticity, whereas both diagnosticity and prototypicality affected the ratings of inference learners.

Table 4: Mean Typicality Ratings from Experiment 2

Classification Learners

	Number of Diagnostic Features			
	3	2	1	mean
Prototype	5.92	-----	-----	**5.92**
Typical	5.98	4.39	-----	**5.02**
Atypical	5.88	4.51	2.93	**4.18**
	5.94	**4.47**	**2.93**	

Inference Learners

	Number of Diagnostic Features			
	3	2	1	mean
Prototype	6.06	-----	-----	**6.06**
Typical	5.24	4.98	-----	**5.08**
Atypical	4.63	4.07	3.66	**4.00**
	5.29	**4.37**	**3.66**	

General Discussion

The diagnosticity of features plays a critical role in current theories of category learning. These experiments investigated the role of diagnosticity for two different means of category learning, classification and inference, and found an important difference. For both experiments, classification learners relied on the diagnostic features when making decisions about category members. Inference learners were sensitive to both the diagnosticity of the features (although much less so than the classification learners) and the relationship of the item to the category prototype. These results indicate that these two different ways of category learning lead to different emphases in the category representation. Yamauchi, Love, and Markman (2002), using a non-linearly separable category structure, also found that inference learners did not show an effect of feature diagnosticity, while the classification learners did, when predicting missing features of items following learning.

The results of this study rule out the strong hypothesis that the inference learners would not be at all sensitive to the diagnostic value of the features. However, this conclusion needs to await further research for two reasons. First, there were only two categories, so the probability that cross-category comparisons might be made (e.g., about the internal structure learned by inference) was probably much greater than in many more realistic situations. When we learn about items there is normally not such an obvious and closely related contrasting category being learned. Second, it is possible that some or all of this influence occurred at test. For example, the effect of diagnosticity in inference learning was most evident when the items were less typical category members. Since these items would have had more features in common with the non-target category, their presence may

have prompted the inference learners to consider that other category (although it was not necessary to do given the design of the transfer tasks). Again, this seems to be much less likely in more realistic category situations. It is important to keep in mind that although the inference learners did show some sensitivity to the diagnosticity of the features, it was sometimes tiny and always significantly less than the classification learners. Despite these possibilities, it is important that future research more fully examine how the category representation is influenced by inference learning.

It is also interesting to consider the classification learners. Their lack of sensitivity to the non-diagnostic features that were part of the prototype suggests an extreme focus in the representation they develop. Other results also point out some difficulties that arise from category learning based solely on classification. Yamauchi and Markman (1998, Exp. 2) found that varying the order of classification and inference learning resulted in a situation where a block of classification learning prior to inference learning was not beneficial (although inference learning prior to classification learning was). Chin-Parker and Ross (in press) showed that classification learners were not sensitive to within-category correlations whereas inference learners were sensitive to this relational structure. Anderson et al. (2002) also found that classification learners were less accurate than inference learners when classifying individual features of category members. These lines of research suggest that classification learning may lead to a category representation that is good at determining category membership of items but is impoverished with regards to other category information.

The question remains as to why classification and inference learning lead to such different category representations. Even though the two learning tasks can be considered formally equivalent, they impose very different demands on the learner. In the classification learning task, a subject is shown an item and predicts the category label using the available information. If a piece of information is not diagnostic, such as "flying" when learning to distinguish birds and bats, it is not important and not incorporated into the category representation. The current experiments and formal modeling (Krushke, 1992) have shown that diagnosticity is the primary concern during classification learning. As noted earlier, the inference learning task focuses the learner on a single category, promoting the acquisition of information about the internal structure of that category. This would make the inference subjects sensitive to what are the most likely features given the category membership. If an item is labeled as a "bird" and then a prediction is made about how it will get from one tree to another, the correct prediction would most likely be "flying", and that piece of information is incorporated into the category representation. Recognition of the features that distinguish birds from bats is not important in this situation, so those diagnostic features would not be made salient.

A major challenge now will be to formalize the differences that exist between the various means of category learning and to incorporate that information into a category learning model. Such a model would be useful for any endeavor within the cognitive sciences concerned with learning from past experience.

In closing, it is important to remember that when we learn about categories in more realistic situations, it is by a combination of different tasks, such as classification, inference, and explanation. The limited representation that is developed as a result of classification learning does not appear to be much like the flexible, dynamic representations that underlie our knowledge of real world categories. However, the same may be true of a category representation that is developed as a result of any single category learning task. It may the combination of various learning tasks that creates a flexible and dynamic representation. To understand category learning as it exists outside of the laboratory, we are going to have to develop a more integrated approach to category learning (e.g., Solomon, Medin, & Lynch, 1999).

Acknowledgments

This work was supported by Grant NSF SBR 97-20304 from the National Science Foundation.

References

Anderson, A., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition, 30,* 119-128.

Chin-Parker, S., & Ross, B. H. (in press). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition.*

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments, & Computers, 25,* 257-271.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 1922-1944.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 1207-1238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences, 3*, 99-104.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Yamauchi, T., Love, B. C. & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 585-593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.

Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28*, 2064-2078

Comprehension Monitoring and Regulation in Distance Collaboration

Kwangsu Cho (kwangsu@pitt.edu)
Christian D. Schunn (schunn@pitt.edu)
Alan M. Lesgold (al@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260, USA

Abstract

Comprehension monitoring and regulation in a distance learning situation were examined in comparison to individual learning through an error-detection paradigm. The collaborative learning condition produced significantly better learning and monitoring. These results were interpreted as the effect of regulative interaction in the collaboration. Then, the specific interactions of 3 good and 3 poor pairs were contrasted to examine their interaction pattern in terms of monitoring and regulation. The results showed that the good pairs had a higher level of monitoring and regulative interaction. Also when the good and poor groups successfully monitored comprehension failure, the poor groups implemented less effective regulation strategies.

To understand text, learners need to integrate successively encountered information from that text into a coherent and well-integrated (mental) representation (Kintsch, 1998). According to Kintsch this comprehension process proceeds in a piecemeal way, sequentially developing a bigger and more coherent representation. This process tends to be prone to errors such as representation of incorrect information and/or misrepresentation of correct information due to omissions, inconsistencies, and/or anomalous and unclear text. When these comprehension failures occur, learners should be able to use metacognitive monitoring to detect the failures and regulation strategies to repair them and thus construct a more coherent understanding of the text in order not to end with a lack of understanding or misunderstanding.

However, despite the significance of monitoring and regulation strategies to text understanding, learners tend to fail to detect their own misunderstandings (Markman, 1979), ignore incorrect information (Otero & Kintsch, 1992), and overestimate their own understandings (Glenberg, Wilkinson, & Epstein, 1982) and capabilities (Presseley & Ghatala, 1990). Learners are too often satisfied with their *faulty* understanding to challenge given tasks and hence fail to trigger regulation processes. Accordingly, various efforts such as metacognitive strategy training, setting up explicit comprehension goals, or self-generating feedback have been made to improve learners' comprehension.

Considering that effective learning often takes place in social settings, and that individual learners' comprehension could be affected by their peers' comprehension, it seems worthwhile to examine comprehension monitoring and regulation in collaborative learning situations. More specifically, comprehension monitoring and regulation seem especially critical in distance collaboration situations where a lot of learning takes place from text. Therefore, the goals of this research are to examine whether distance collaboration improves individuals' comprehension monitoring and regulation abilities, as well as the conditions that make distance collaboration produce effective or ineffective text comprehension.

Monitoring and Regulation in Collaboration

Monitoring and regulation have been considered critical in effective face-to-face collaboration because they help learners construct a more coherent understanding. First, externalizing thinking and understanding through communication might help collaborators better monitor and regulate their performance (Miyake, 1997) because it causes thinking and understanding to become objects that can be sharable and manipulable between collaborators (Miyake, 1986). While learners working alone are often subject to self-confirmation bias, learners can benefit from working with peers thanks to a 'checking mechanism' in collaboration (Miyake, 1986) that advances comprehension monitoring and regulation. Second, the division of cognitive processes in collaboration (Dillenbourg, Baker, Blaye, & O'Malley, 1995) may play a part in improving monitoring and regulation in collaboration. For example, one peer might take the role of leader, while another peer might take the role of monitor (e.g. Miyake, 1986). In the process of collaboration, many errors made are detected and corrected by partners (Miyake, 1986; Resnick & Salmon, 1993). Also, Karabenick (1996) recently showed that learners may have better comprehension monitoring after receiving questions from colearners. Third, comprehension monitoring and regulation could be easily implemented when peers have conflicting perspectives. As Piaget's socio-cognitive conflict theory suggests, collaborating individuals with different understandings of the same task may advance their understanding in the process of resolving their differences. Fourth, regulating comprehension problems seems fundamental to collaboration processes because the regulation process in collaboration may be activated automatically (Schegloff, 1991), and incorrect elements of their representation might then be fixed through communicative interactions such as engaged discussion (Kruger, 1992), elaboration or arguments (van Boxtel, van

der Linden, & Kanselaar, 2000), or other repairs (Lumpe & Staver, 1995).

Because collaborating learners may have higher chances of monitoring (detecting that there might be something wrong) and of regulation (knowing what could be correct knowledge), it might be straightforwardly expected that collaborating learners will have better comprehension than isolated learners. However, when distance between learners is involved in learning, the above inference seems complicated because people working together at a distance report various kinds of difficulties that seem to deteriorate collaboration. First, the lack of nonverbal communication cues in distance communications may torture clear communications (Armstrong & Cole, in press) that help to manipulate thinking. Second, distance learners have difficulty in grounding communications and spend a long time doing so (Kiesler, Siegel, McGuire, 1984). Third, some studies report that cognitive conflicts are not only well detected (Armstrong & Cole, in press), but also rather emotionally charged with no easy method of cognitive resolution. Finally, anonymous individuals who are placed in group distance learning situations tend to be less supportive of each other because of low perceptions of group cohesion and conformity.

Therefore, one could propose the following model of effective distance collaboration. When comprehension failures occur, they should be detected. If not, the failures might end up with non- or mis-comprehension. Once the failures are detected, they should be repaired. If not, the failures also might lead to non- or mis-comprehension. To test the model, we hypothesized that if interactions between individuals working at a distance (e-Pairs) are sufficiently effective, they will be better than the individuals working alone (Singles) in learning scores because of better comprehension monitoring and regulation and that good e-Pairs will be better than poor e-Pairs in comprehension monitoring and regulation.

Method

Comprehension monitoring and repairing during distance collaboration was compared to monitoring and repairing during individual learning. Unlike typical face-to-face collaboration studies that emphasize ecological validity, we wanted better experimental control and a wider range of data.

Participants. Sixty-nine undergraduates (Male = 27, Female = 42) taking introductory psychology courses volunteered in this study. All the participants received course credits for participation. The first language of all the subjects was English. They all reported that they had experience using chatting interfaces on the internet and were familiar with these interfaces. Randomly, thirty-seven of them were assigned to an individual learning condition (Singles: male=15, female=22) and the other 32 to the collaborative learning condition (e-Pairs: male=12, female=20). All e-Pairs participants were randomly paired with a same sex partner. One pair was removed from the data analysis because of a problem with the interface.

Materials. Two expository texts about theories of knowledge representation were used. One text concerned symbolism and the other connectionism. The text content was selected because undergraduate students were not familiar with these theories of representation, and this allowed us to minimize the pre-knowledge effect and maximize the purity of comprehension monitoring and regulation strategies. Each text consisted of 15 sentences and had two versions: Consistent and inconsistent. Following Markman (1979)'s *error-detection paradigm,* inconsistent versions had contradictory or inconsistent information at the 5th, 10th and 15th sentences. For example, the first five sentences used in the symbolism text were *(1) One of the major theories about representation is called symbolic representation. (2) The symbolic representation view is that the human mind represents information as a language-like or symbolic form. (3) Because most of us think and all of us write linguistically, we tend to couch our ideas in symbols like a natural language form. (4) We can understand thought, belief, problem solving in a language-like symbolic form. (5c) Thus, in this view, symbols (roughly, words) are used to represent information in the human mind. (5i) Thus, in this view, symbols (roughly, words) are not used to represent information in the human mind.* Here the 5c was a consistent sentence, while 5i was an inconsistent sentence. Thus, when subjects came to read the fifth sentence, either (5c) or (5i) was displayed to them. Detecting the first inconsistent sentence located at the 5th position was manipulated to be the easiest, that of the second at 10th position the middle, and that of the last at the 15th position the most difficult in terms of the amount of correct representation needed to detect the inconsistency.

Interface. A computer interface (*see* Figure 1) was used to manage the experiment automatically, to collect data, and to provide an environment in which the participants could work. The interface for the main experiment session consisted of five units: (1) a new sentence display unit, (2) a history window, (3) a monitoring detection task unit, (4) a comprehension self-rating slider, and (5) an IRC (internet relay chatting) as a distance communication channel. The new sentence display unit was used for displaying each new sentence. When each new sentence appeared, the previous sentence moved up to be located at the bottom in the history window which accumulated all the previous sentences. Thus, the participants could focus on comprehension instead of memorizing sentences. The distance communication channel was an internet relay chatting interface where each pair communicated without any verbal and nonverbal interaction. The individual learning condition was identical except for not having the distance communication channel.

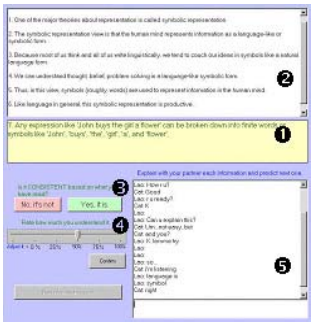


Figure 1: Computer interface

Comprehension monitoring task. There were two monitoring tasks: Detection and comprehension self-rating. The detection task was to decide whether or not each sentence was consistent with previous sentences. The self-rating of comprehension was measured with a rating scale labeled with 0%, 25%, 50%, 75%, and 100%, indicating the approximate percentage of the meaning that the subject believed he/she understood. However, self-rating measures appeared unreliable and were thus removed from the results.

Comprehension regulation interaction coding scheme. Each episode (the period between the end of one sentence and the start of the next new sentence), was evaluated in terms of the level of monitoring and regulation quality exhibited. The conversation levels were coded using the following hierarchical scheme: 0: off-task – coded when an episode consisted of task-unrelated things; 1: Checking answers – coded when an episode consisted only of asking for and providing each other's answers; 2: Rephrasing – coded when an episode consisted of providing answers and rephrasing the given sentence as their rationale; 3: Explanation – coded when an episode included integrating, relating, or generating information to explain answers, 4: Elaboration – coded when subjects proceeded to elaborate or clarify each other explanation, and 5: Negotiation – coded when an episode was resolved with an agreed cognitive solution. This scheme was hierarchical in that the higher, the better in comprehension as a continuum from low level monitoring (Checking answers) to high level regulative behavior (Negotiation). When multiple levels in an episode appeared together, the highest one was selected to represent the quality of interaction of the episode. Two coders independently coded two randomly selected groups for the analysis and achieved a 0.84 inter-coder reliability.

Procedure. Each participant was randomly assigned to either the Singles condition or the e-Pairs condition. The participants went through an instruction session, a pretest session where they answered 20 multiple choice questions about the main texts, a warm-up session that had two short texts to familiarize them with the interface, a 2nd instruction session that was exactly same as the 1st instruction, a main task session, and a posttest session. In the instructional sessions, they read that they would study, with or without a partner, two draft texts that might or might not have inconsistent information. They were also instructed that they should explain the meaning of each sentence – to themselves in the Singles or with their peer in the e-Pairs – and that they would get bonus credits based on their performance. Both the pre- and post-test comprehension questions, and their alternatives about the main tasks, were completely randomized. In the main task session, all the participants in the e-Pairs were randomly paired with a same-sex partner with whom they had no interaction before the main task session. The order of presentation of the two texts was randomized and the selection of either version was counterbalanced. Thus, the participants studied two texts but only one of them was an inconsistent version. In each episode identified as each sentence level interaction, whenever a new sentence appeared, the participants individually read the sentence, performed the comprehension monitoring tasks (the detection and self-rating task) with no means of communication. Then they studied the sentence by explaining the meaning of the sentence either alone in the Singles or together with a peer through the distance interaction channel in the Pairs. When they decided to finish studying, they hit a button to request another sentence, at which point the communication channel was automatically disabled. These sentence level activities repeated until the end of the two tasks. Note that the e-Pairs were not asked to reach an agreement, did not have an interaction before the main task session, and only their first names were shown on the communication channel interface.

Results

1-1. e-Pairs will be better than Singles in learning

A one-way ANOVA showed no significant difference in the pretest scores between the Singles and e-Pairs. Another one-way ANOVA was done with the learning defined as the difference between the post- and pre-test scores. The e-Pairs (M = 4.47, SD = 1.73) were significantly better than the Singles (M = 2.62, SD= 1.88) ($F(1,65)$ = 5.13, p = .03). Finally, the effect size as *Cohen's d* was 1.02. These results provided a rationale to conduct further analyses.

1-2. e-Pairs will be better than Singles in monitoring

The detection task performance as comprehension monitoring was examined (*see* Figure 2). In the consistent versions, the e-Pairs (M = .92, SD = .06) and Singles (M = .88, SD = .04) were not significantly different on any sentence. However, in the inconsistent versions, the e-Pairs were significantly better for the first inconsistent sentences, the easiest one (5th sentence: for e-Pairs M = .75, SD = .18; for Singles M = .18, SD=.01) at text 1: $F(1,32)$ = 18.03, p = .00 and at text 2: $F(1,33)$ = 5.05, p = .02), not for the second (10th) and the third (15th) inconsistent sentence.

2-1. Good e-Pairs will be better in monitoring

To examine what mechanisms drive effective distance collaboration, three good (M = 13.3, SD =2.1) and three poor e-Pairs (M = 4.0, SD = 7.0) were selected, t (4) = 5.74, p =.00). This selection was made after removing e-pairs where peers had large knowledge differences, to avoid looking at extreme cases. No significant differences between the good and poor Pairs were found in the pretest scores (Good: M=11.0, SD=2.0; Poor: M=13.0, SD=5.6), the pretest difference between members in each pair (Good: M=1.0, SD=.0; Poor: M=1.0, SD=1.0), the mean number of turns per episode (Good: M=6.7, SD=.3.3; Poor: M=6.4, SD=5.0), the total time spent per group (Good: M = 52.7 min, SD=15.0; Poor: M=47.2, SD=14.3), and time per episode (Good: M=1.76 min, SD=.5; Poor: M=1.6, SD=.5). These non-significant indices formed the baseline against which to examine differences in the level of monitoring and regulation quality in collaboration.

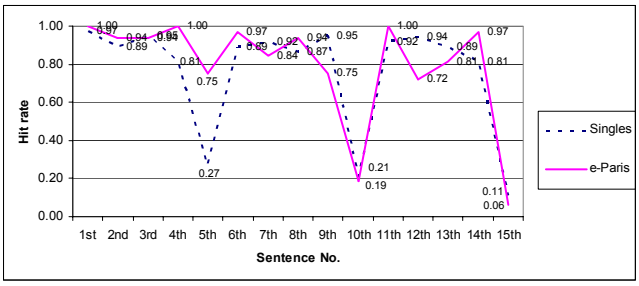


Figure 2. Comprehension monitoring task performance

Then, the hypothesis that high e-Pairs will be better in comprehension monitoring was tested. There were no significant differences at all three inconsistent sentences, although the pattern looks similar to that of the comparison between the e-Pairs and Singles.

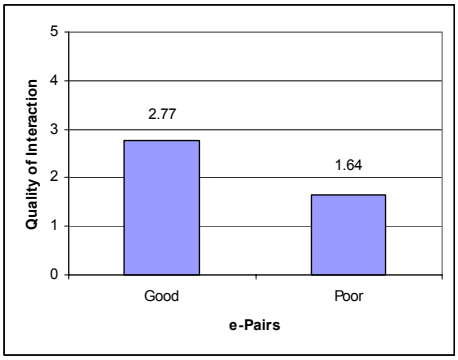


Figure 3. Quality of regulative interaction in general

2-2. Good e-Pairs will have higher quality of monitoring and regulative interaction than poor e-Pairs

Another difference was found in the level of monitoring and regulation quality in their interaction. The good e-Pairs (M = 2.77, SD = .97) had a significantly higher level of regulative interactions than the poor e-Pairs (M = 1.64, SD= 1.0), $F(1,178)$ = 58.40, p = .00 (*see* Figure 3). In general, the good e-Pairs interaction quality was around explanation level, whereas the poor e-Pairs interaction quality was between just checking answer and rephrasing (refer to coding scheme in the method section).

2-3. Given successful monitoring, good e-Pairs will be better than poor e-Pairs in regulative interaction.

Before answering the question, we examined when e-Pairs had longer conversations, which means they tried to do something more like repairing. Based on each individual decision on the detection task before starting each interaction period, each episode was categorized into one of three categories: both members' answers or perspectives were 'same and correct', 'same and incorrect', or different. Then comparisons were made between good e-Pairs and poor e-Pairs. According to a two-way ANOVA (F(2,174) = 8.2, p = .00) and its S*cheffe*, the only significant difference was on the category dimension, especially between the level *different* and the others (*see* Table 1).

Table 1: The mean number (SD) of turns in episode

Categories	Good	Poor	Mean
Same & Correct	6.3 (3.1)	5.7 (4.8)	6.0 (4.0)
Same & Incorrect	5.0 (2.5)	6.8 (4.2)	6.0 (3.3)
Different	9.8 (3.0)	8.8 (4.9)	9.2 (4.3)
Mean	6.7 (3.3)	6.4 (5.0)	6.5 (4.1)

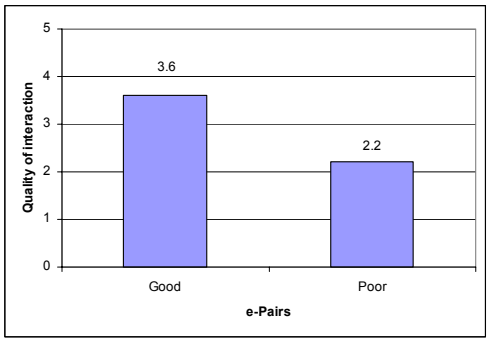


Figure 4: Quality of interaction given successful monitoring

Then we examined the hypothesis that good e-Pairs will have higher quality of interaction (*see* Figure 4). Regulative interaction qualities between the good and the poor e-Pairs were compared when they all had successful monitoring. Given the total 24 episodes (good: n = 10; poor: n = 14) where peers in an e-Pair had different perspectives that signaled there might be something *wrong*, their interaction qualities were examined. Interestingly, the poor e-Pairs (M = 2.21, SD =1.31) showed a significantly lower level of regulative behavior than the good e-Pairs (M= 3.60, SD = .52), $F(1,23)$ = 9.95, p = .00. The interaction quality of the good e-Pairs was between the explanation and elaboration

level, while that of the poor e-Pairs was around rephrasing. For example, the following episode from a good e-pair shows that when peers had different opinions they tried to resolve the difference.

1: Lao	i put incorrect b/c i had no clue what that was about
2: Lao	sorry:(
3: Cat	haha
4: Cat	that's alright
5: Lao	i just thought that the info sounded like conflicting symbols
6: Cat	it's just saying that by adding another symbol to a sentence you can make it a fact
7: Cat	the sentence is kinda weird
8: Lao	oh ok
9: Lao	yeah it is
10: Cat	maybe the next sentence will be about displacement
11: Lao	ok

However, another episode from a poor e-pair shows that after they checked their answers they did not try to resolve their comprehension failure.

1: Cu	hmm
2: Ja	i wasn't sure about this one
3: Cu	me either
4: Cu	I chose incorrect
5: Ja	oh, i chose correct. i don't know why though...
6: Cu	me either
7: Ja	oh well

Discussion

Comprehension processes are error-prone because they are constructive and approximate. Learners need to be error sensitive to attain error-proof comprehension. In this study, we examined the role of collaboration in improving comprehension monitoring and regulation in a distance communication situation, a matter that had not been investigated before. With a relatively well-controlled collaboration experiment, we first showed that distance collaboration is more beneficial to learning than working alone. In addition, performance in detecting contradictory information is also somewhat better in collaboration. Therefore, the better learning that occurred in the e-Pairs may be attributed to the process of collaboration.

Furthermore, to examine the role of collaboration in comprehension monitoring and regulation in detail, 3 good e-Pairs and 3 poor e-Pairs were examined. The good e-Pairs were not significantly better than the poor e-Pairs in comprehension monitoring (error detection). However, the regulative interaction quality of good e-Pairs' interactions was generally higher than that of poor e-Pairs. In general, the good e-Pairs interaction quality was around the explanation level, whereas the poor e-Pairs interaction quality was between just checking answer and rephrasing.

Another interesting finding was from the comparison when both the good e-Pairs and the poor e-Pairs had successful monitoring. The poor pairs' regulative interactions were not highly activated even though their comprehension problems were monitored explicitly, while the good groups tended to indulge in higher level of regulative interaction.

Therefore, the results can be interpreted as supporting the claim that participants in distance collaboration benefit from collaborative interaction by improving their detection of comprehension failures, and implementing repair processes through regulative interaction. Also, the results support the research model that states that when comprehension failures or cognitive conflicts happen they should be detected and repaired to achieve correct comprehension or learning.

Thus, the model explains why some research on cognitive conflict finds increased learning while other research does not. As the model states, cognitive conflicts do not necessarily result in learning unless the conflicts are detected and resolved. In this experiment, no case was found to reach a cognitive resolution coded as negotiation. Instead, a lot of cases ended up with social negotiation. Here social negotiation means that conversants agree to blur their conflicts without reaching a clear resolution, as seen in the example conversation from the good e-Pair. Interestingly, there was also no instance of flaming, which is frequently reported in distance collaboration studies.

The so-called 'checking mechanism' (Miyake, 1986) may be a key for suppressing self-confirmation bias that may be dominant in solo learning. Self-confirmation bias is a tendency to stick to an already held explanation rather than developing alternative explanations. This tendency, when learning alone, tends to block learners from changing their representation by suppressing (Otero & Kintsch, 1992) and/or ignoring (Chinn & Brewer, 1993) inconsistent information that does not match with their representations. However, the confirmation bias in a group may be smaller, because groups are better than individuals at rejecting presuppositions (Gorman, Gorman, Latta & Cunningham, 1984), so long as they entertain hypotheses and alternative ideas, and consider justifications (Okada & Simon, 1997).

The results of this research are consistent with other research in the collaboration community. For example, Brown and Campione (1986) argued that "understanding is more likely to occur when a student is required to explain, elaborate, or defend his or her position to others; the burden of explanation is often the push needed to make him or her evaluate, integrate, and elaborate knowledge in new ways" (*p*. 1060). Also, Forman and Cazden (1994) identified *parallel, associative,* and *cooperative* interaction patterns, of which *cooperative* is the highest level – characterized as constantly monitoring, guiding and correcting each other's

work. Additionally, Barron (2000) argued that after contrasting a high-achievement group with a low-achievement group, greater monitoring for coordination between members would result in higher results. Therefore, collaboration might be an ideal way to improve individuals' monitoring and regulation abilities.

Finally, some aspects of this study should be noted that may limit generalizations of the results. One is that this experiment was highly controlled compared to other face-to-face collaboration research. We tried to separate the collaboration period from individuals' comprehension monitoring decision periods, to examine the effect of collaboration on individual learners' comprehension. Also, we tried to remove socially confounding variables. For example, the participants in each pair did not interact before the main tasks. Although this may appear to limit the *ecological* validity of this study in terms of face-to-face collaboration, it seems acceptable in terms of *e-cological validity* since distance collaboration is often between anonymous individuals. Also it may provide a cleaner demonstration of the cognitive effects of collaboration on learning.

Acknowledgments

This study was supported by the NetLearn project funded by a Technology Innovation Challenge Grant from the U.S. Department of Education to New York City Community School District No. 2 and the University of Pittsburgh and by Microsoft Corporation. We thank Randi Engle, Heisawn Jeong, Lelyn Saner, Mark McGregor, Patrick Jermann, Amy Soller, and Brad Morris for their sincere help.

References

Armstrong, D. J. & Cole, P. (In press). Managing distances and differences in geographically distributed work groups. To appear in P. Hinds & S. Kiesler (Eds.), *Distributed Work* (working title).

Barron. B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Science*, 9 (4), 403-436.

Brown, A. L., & Campione, J. C. (1986). Psychological theory and the study of learning disabilities *American Psychologist*, 41, 1059-1068.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63, 1-49.

Dillenbourg, P., Baker, M., Blaye, A., O'Malley, C. (1995) The Evolution of Research on Collaborative Learning. In Reimann, P., Spada, H. (eds.) *Learning in human and machines. Towards an interdisciplinary learning science*, London: Pergamon.

Forman, E.A. & Cazden, C.B. (1985). Exploring Vygotskian perspectives in education: The cognitive value of peer interaction. In J.V. Wertsch (Ed.), *Culture, communication and cognition: Vygotskian perspectives*. New York: Cambridge University Press.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10(6), 597-602.

Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology*, 75,65-79.

Karabenick, S. A. (1996). Social influences on metacognition: Effects of colearner questioning on comprehension monitoring. *Journal of Educational Psychology*, 88, 689-703.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39, 1123-1134.

Kintsch, W. (1998). *Comprehension*. Cambridge Univ. Press.

Kruger, A. C. (1992). Peer collaboration: conflict, cooperation, or both? Social Development, 2&3, 165-182.

Lumpe, A. T. & Staver, J. R. (1995). Peer collaboration and concept development: Learning about photosynthesis. *Journal of Research in Science Teaching*, 32(1), 71-98.

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 50, 643-655.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, 10, 151-177.

Miyake, N. (1997). *Making internal process external for constructive collaboration*. In Marsh, J., Nehaniv, C., and Gorayska, B., eds., Cognitive Technology, 119--123. IEEE Computer Society.

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, 21(2), 109-146.

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science*, 3(4), 229-235.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist*, 25(1), 19-33.

Resnick, L., & Salmon, M., et al. (1993). Reasoning in conversation. *Cognition and Instruction*, 11, 347-364.

Schegloff, E. A. (1991). Conversation analysis and socially shared cognition. In L. B. Resnick, J. M. Levine, & S. D. Teasley, *Perspectives on socially shared cognition*. Washington, D.C.: American Psychological Association.

van Boxtel, C., van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction*, 10, 311-330.

Second Order Isomorphism: A Reinterpretation and Its Implications in Brain and Cognitive Sciences

Yoonsuck Choe (choe@tamu.edu)
Department of Computer Science
Texas A&M University
3112 TAMU
College Station, TX 77843-3112

Abstract

Shepard and Chipman's second order isomorphism describes how the brain may represent the relations in the world. However, a common interpretation of the theory can cause difficulties. The problem originates from the static nature of representations. In an alternative interpretation, I propose that we assign an *active* role to the internal representations and relations. It turns out that a collection of such active units can perform analogical tasks. The new interpretation is supported by the existence of neural circuits that may be implementing such a function. Within this framework, perception, cognition, and motor function can be understood under a unifying principle of analogy.

Introduction

One of the central tenets in neuroscience is that neurons receive incoming spikes, process that spatial or temporal information, and then pass on the transformed information for further analysis. Also, neurons that fire together develop strong connections (Bliss and Collingridge 1993). Thus, the neurons represent features in the input, and connections encode relational context among features. This viewpoint is analogous to the second order isomorphism by Shepard and Chipman (1970; below, just S&C). However, a problem can arise depending on how we interpret S&C's theory.

The difficulty comes from the *static* role assigned to representations. In this paper, the representations and the relations are given an active role. When working as a collection, these active units can perform an analogical function. In fact, a similar active approach has been employed in previous work, resulting in the emergence of analogical (Hofstadter 1985; Mitchell 2001) or metaphorical (Narayanan 1999) functionality.[1] An important observation advanced in this paper is that the function of active representations and relations are very similar to that of neurons, and specific circuits in the cortex and the thalamus can actually *implement* analogical functions. Analogy is commonly attributed to higher cognitive faculties only, but it does not always have to be the case (Chalmers et al. 1992); it may be part of a larger set of human brain function including perception and motor function. I will discuss in the end how such an analogical framework can allow us to better understand the nature of cognition and brain function.

[1] Analogy and metaphor are closely related in that they refer to similarities in relations and attributes although the relative degree in each may differ (Gentner 1989).

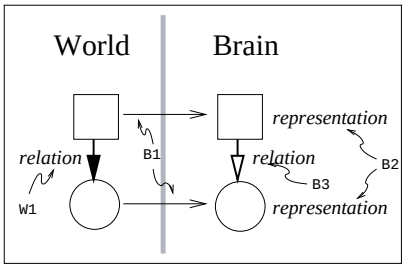
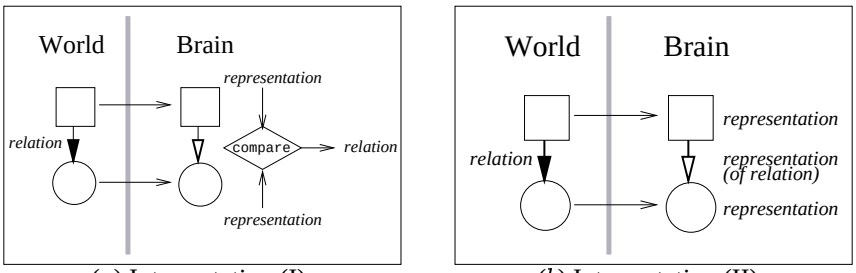


Figure 1: **S&C's Second Order Isomorphism.** There are two objects, one square and one round in the world (on the left). The internal representations in the brain of these two objects are shown on the right. The vertical arrows represent the relations between the objects. The two horizontal arrows represent mapping from the world to the brain which is initiated by sensory transduction. Note: The square and circle on the right (in the brain) are just there for the ease of reference. They can be removed without causing any change in content (this applies to the rest of the figures).

Common Interpretations

Under second order isomorphism the brain needs to find the relation between the (1) relations between external objects and (2) relations between internal representations (figure 1). S&C's theory seems to be more appropriate in modeling how our brain represents the world than Locke's Isomorphism (Edelman 1998, 1999). In physical terms, we can interpret the figure as follows: (1) relation in the world (W1; coincidences in sensory events) (2) arrows from world to brain (B1; sensory transduction), (3) representations in the brain (B2; afferent connections), and (4) relation in the brain (B3; lateral connections). Of these, let us focus on what is available in our brain (B1–B3). If we take for granted the information our sensory transducers tell us, we can drop B1 from our discussion and focus on just B2 and B3.

An implicit message in figure 1 is that two objects are represented, and some brain process then judges the relationship between the two (the open arrow). Making this point more explicit, we can illustrate S&C's theory as in figure 2*a* (the diamond box). We can see that a difficulty can arise in such an interpretation; something has to perform the comparison function, but this creates an ever increasing levels in a hierarchical way (i.e. higher areas



(a) Interpretation (I). (b) Interpretation (II).

Figure 2: **Common Interpretations of Second Order Isomorphism.** (a) An explicit comparison mechanism is necessary to judge the relations between the two representations. (b) The comparison box is replaced by a representation of relation. However, this figure still requires a third-party to *evaluate* the representation of relation.

judging the output relation in the lower areas). However, as Hilgetag et al. (1996) noted, it is hard to determine a strict hierarchy among cortical maps (in this case, between visual areas). Also, as Zeki (2001) suggests, integration of these information may be a nonhierarchical process. Thus, representing something and delaying the interpretation until later may not work very well.

One can argue that the lateral connections *represent* the similarity relation, not requiring a separate interpreter (figure 2b). However, we still need something to evaluate (or interpret) the resulting representation. Thus, this reformulation just replaces the need for one kind of interpreter with another.

Assigning an Active Role to Relation

What can a relation be if it should not be a representation? The problem seem to come from representations and relations playing a *static* role. What if we assign an active role to the representations as shown in figure 3a? In the figure, I assigned an active role to the relation arrow itself, allowing one representation to invoke another. Thus the activation of the internal representation of the square *invokes* (or turns on) that of the circle, and vice versa.

Now consider how can we use this new active relation (note that it is *directional*) to describe the relations in the world. First, we have to know what kinds of relations exist in the world. There are two basic relations: spatial and temporal relations. Spatial relations are between objects, and they are causally bidirectional.[2] On the other hand, temporal relations are between events, and they are causally directional. When one event precedes the other, the reverse cannot happen simultaneously.

In the brain, action potentials only propagate in one direction along the axon, and the adaptation of synapses tend to learn causality (Song et al. 2000). Such connections are ideal for implementing temporal relations, but what about spatial relations? If we pair a unidirectional arrow from A to B with a reciprocal one from B to A, then we can indeed represent spatial relations with only directional arrows. If representations for object A and B simultaneously activate through mutual excitation, then they can represent the spatial relation between the two. So, let us update our figure again to include backward relational arrows (figure 3b). We can now think in terms of *temporal* relations only, because spatial relations seems to be a special case of temporal relations (at least in the brain).

[2]Note that *causal* simply means that one event precedes the other in time.

The Role of Active Relations and its Neural Basis

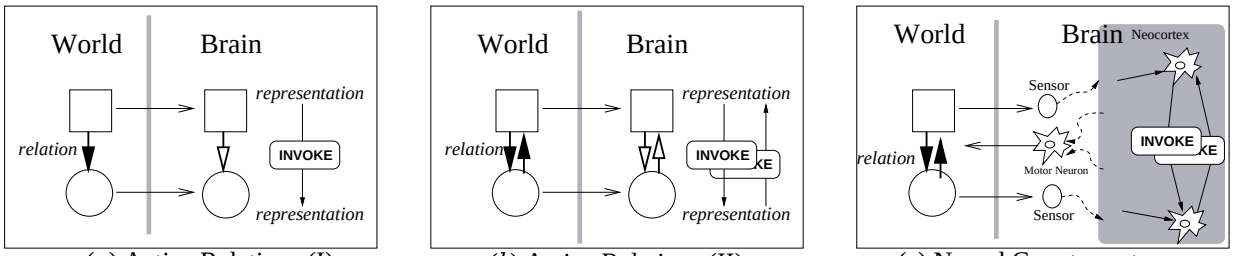
In the previous section, I replaced the representation for relation by an active relation. What about the representations for the objects (or events)? Representation is an inherently static term (like a symbol), thus, we should take a more active viewpoint and ask *what action occurs* when a neuron detects a pattern in its incoming input, rather than focusing only on what a neuron represents. Knowing what representations *do* may be as important as knowing what they *stand for*.

To discover the relationship between things in the world, we need the motor capabilities as much as we need sensors. Thus, between the world and the brain there must be a backward arrow from the brain to the world. The resulting diagram is shown in figure 3c. This addition is crucial in learning the relations in the world (O'Regan and Noe 2001). The final diagram looks very similar to the basic circuitry in our brain. How can this final figure help us understand the mechanisms of the brain? The key is to understand what is the *action* taken by a neuron, no less than to know what it represents.

Active Relations: A Primitive for Analogical Processing

Now we have a single active functional unit: a neuron that fires a spike along the *active relational arrow* as soon as it detects a certain input feature. This unit alone cannot achieve much, neither can a serial chain of such units. The true power of this simple unit is revealed when it is used in a massively parallel way. This may be an obvious line of thought because that is what our brains seem to do. However, it turns out that the collective effort of these simple units can embody a simple yet powerful function of analogy.

We have to simplify matters to see how such neurons can process analogy. Let us assume there are six neurons in an imaginary creature's brain inhabiting the world



(a) Active Relations (I) (b) Active Relations (II) (c) Neural Counterparts

Figure 3: **A New Interpretation of Second Order Isomorphism.** (a) In this new interpretation of S&C's theory, an active role is assigned to the relation arrow. Notice that the INVOKE arrow is a single arrow, not one arrow going into and another leaving out of the box. Thus, the rounded box signifies that the arrow actually performs an *action*. (b) Backward relational arrows are added in the brain to account for the mutual, but directional nature of relations. With two relational arrows, both spatial and temporal relations can be implemented. (c) The neural counterparts of (b) are shown. The limiting term *representation* is removed, and the motor reaction (backward arrow from the brain to the world) is added. Sensory transducers are also explicitly shown.

of fruits (figure 4). After the fruit brain experiences the world of fruits, it will learn the co-occurrences between features and establish relational arrows as shown in the figure (arcs with arrows). Also suppose that the brain is partitioned into several specialized map areas (or partitions), as in cortical maps. Now suppose <apple>, <orange>, and <word-red> were presented to the creature simultaneously. If we track the activation, we can see that these detectors will turn on: apple detector, orange detector, color-red detector, color-orange detector, and finally, word-red detector. These activations are *input-driven*. Because the neurons are active, as soon as they detect what they are familiar with, they send out signals through the relational arrows horizontally across the cortex. As a result of this second order activation, the word-orange detector turns on, even without input. Now, here is the crucial moment. We can ask this question: *which neuron's firing was purely cortically-driven?* (note that we can view this as a filtering process). The result of the filtering is then <word-orange>. The significance of this observation is that this process is very similar to solving analogical problems. The input presented to the creature is basically an analogical query: <apple>:<orange> = <word-red>:<?>. The filtered cortical response <word-orange> can then be the *answer* to this query.[3] Thus, active neurons can perform a rudimentary analogical function when the responses are filtered properly.

However, things can get complicated when combinations of objects are used as a query. Let us extend the creature's feature detectors to include concepts of small and big (not shown in the figure). Then we can allow the creature to learn the relations again. We can then present an analogical query like this: <big><apple>:<small><apple> =<big><orange>:<?>. In this case, if we follow the same steps as above we come across a problem. Because the answer we expect (i.e. <small> <orange>) already appeared in the query, if we look for purely

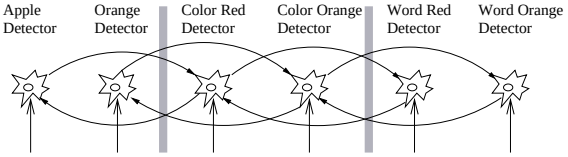


Figure 4: **World of Fruits.** A brain with object, color, and word detector neurons is shown. The six neurons each respond to these input features as labeled above. At the bottom is the fruit world, and the thick vertical arrows represent afferent input. The horizontal arcs are the relational arrows that point to their most frequently co-occurring counterparts that have been learned through experience. The gray vertical bars represent the partitioning of the brain into separate map areas (from the left to right, object map, color map, and word map). Note that for simplicity, the word-orange detector connects only to the color-orange detector, but not the orange detector, i.e. it is a word-color-orange detector, not a word-object-orange detector.

cortically-driven activations, the answer will be <word-red> <word-orange>. However, we can overcome this problem if we ask: *what are the most cortically-driven activities in each partition of the brain*? Because <big> and <apple> appeared in the input twice but <small> and <orange> appeared only once, the latter two can be selected, as well as the purely cortically driven activities listed above. Thus, even for derived activities that are input-driven, those that are less input-driven can survive and the correct analogical response can still be found among such activities that are more cortically-driven within each partition (or area). Note that <color-orange> also survives the filtering, but what is more important here is that a simple filtering process as described above can generate a *small subset of potential answers* to analogical queries. Although the simple analogical query presented above has a straight forward answer, in more complex analogical problems, there can be multiple answers depending on the interpretation (Hofstadter and Mitchell 1994; Mitchell 2001).

In this section, I have shown that active neurons that detect input features and establish relational contexts can collectively perform a rudimentary analogical function.[4]

[3] There is an issue of how the presence of <word-red> can affect the outcome at all. This problem will be discussed later in the discussion section.

[4] Analogical tasks can become much more complex than the ones shown here. The example in this paper is decidedly simple

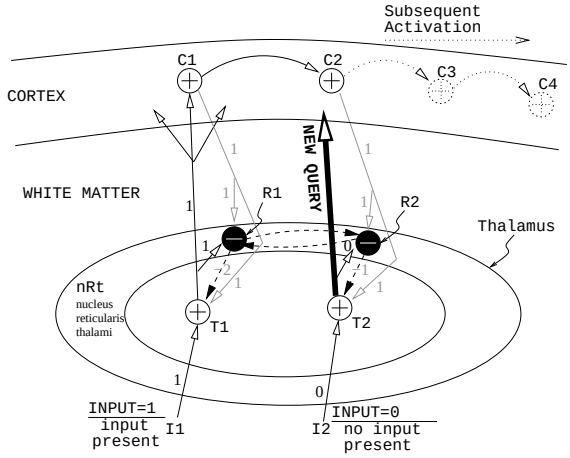


Figure 5: **Analogical Filtering in the Thalamus.** The diagram shows a simplified thalamo-cortical loop that can perform analogical completion and selection and then propagate the selection back to the cortex. All connections shown are based on known anatomy of the thalamus and the cortex (Mumford 1995). I1 and I2 are input fibers, T1 and T2 are thalamic relay cells, and R1 and R2 are inhibitory nRt cells. C1, C2, C3, and C4 are cortical neurons (each is a set of neurons ranging multiple layers in a single cortical column). The neurons are either excitatory (+) or inhibitory (–), and the arrows are axons (pointing in the direction of action potential propagation). The numbered labels on each arc show the activity being carried. Black solid arrows are ascending fibers to the cortex and the cortico-cortical connections (relational arrows), and gray solid arrows are cortico-thalamic feedback. Black dashed arrows are inhibitory connections. The diagram shows a scenario when an input was presented to C1, which excites C2, and in turn generates the feedback from C2 to T2, which is then retransmitted to the cortex as a new query (ascending thick black arrow). The selection decision for further propagation to the cortex depends on the relative excitation and inhibition T1(T2) receive from C1(C2) and R1(R2). On the right of C2 (dotted) in the cortex is the subsequent cascade of analogical completions. Note that to avoid clutter, reciprocal connections in the cortex are not shown.

But does the brain function in such a way? In fact, an exact circuit that may be implementing such a function exists in the brain.

Neural Basis of Analogical Completion and Filtering

Two basic mechanisms are needed to account for the proposed analogical function: completion and filtering. Below, I will discuss how the cortico-cortical connections and the thalamo-cortical loop can implement these two mechanisms.

Completions may be accomplished by the long-range cortico-cortical connections (Mumford 1992). As mentioned earlier, synapses are strengthened when the presynaptic activity precede postsynaptic activity (Song et al. 2000), thus the connections can implement causal relations. Also, specific patterns of connections observed in animals (e.g. visual cortex of monkeys; Blasdel 1992) show how such patterns can implement specific completion functions. Computational models also showed how such patterns can encode feature co-occurrence and how they can dictate the performance of the model (Choe 2001; Geisler et al. 2001).

For filtering, a separate mechanism is necessary. In the thalamo-cortical loop, there exists a massive feedback from the cortex to the thalamus and an inhibition mechanism within the nucleus reticularis thalami (nRt) on the surface of the thalamus (Mumford 1995). This particular architecture has been thought to be involved in the analysis and synthesis of new memories (MacKay 1956), active blackboard (Harth et al. 1987; Mumford 1995), global workspace (Newman et al. 1997), and finally, generating attention and consciousness (Crick and Koch 1990). It turns out that these feedforward and feedback connections from nRt to the cortex together with the nRt inhibitions can filter the feedback from the cortex to promote the most cortically-driven feedback, i.e. the analogical answers. Let us first see how the purely cortically-driven activities are selected (figure 5). In the thalamus, ascending fibers (T1 to C1) branch out and excite the inhibitory nRt neuron R1 (T1 to R1). When the feedback from C1 to T1 comes back, it branches and stimulates R1. As a result, if the descending feedback had a matching ascending signal, the inhibition T1 receives is twice as high as other neurons in the thalamus that are activated by purely cortically-driven feedback (i.e. that of T2). If the synaptic weights are appropriate (i.e. $w_{\mathrm{TC}} = 2$ and $w_{\mathrm{TR}} = 1$)[5], at T1 the feedback will cancel out, but at T2 the feedback will survive the inhibition and be retransmitted to the cortex (the *new query* arrow). Such a surviving cortical feedback, together with the input stimulus at the next moment form a new analogical query to the cortex, and the same process is repeated. That is, C2 elicits activities in C3, and in turn C4 through the thalamo-cortical loop (note that they can be quite far away). For the selection of the *most cortically driven* feedback, the mutual inhibitions in the nRt layer (e.g. between R1 and R2) may disinhibit (inhibiting an inhibitory neuron results in less net inhibition at the target; figure 5) each other and allow the more cortically driven feedback to go back to the cortex, even when all current cortical activities are input driven.

Discussion

The neural mechanisms described in this paper can only account for simple kinds of analogies, and in some case it can even seem as simple pattern completion. For example, <orange> = ? will result in the same answer <word-orange> as in the *Active Relations:...* section. How can the term <word-red> in the original query affect the outcome at all? For this, I believe that among many possible completions, the general map area (i.e. the partitions in figure 4) that are activated by input gets higher preference. In this example, the fruit-map, word-map and color-map will turn on, thus purely cortical activations in other general maps (say odor-map, etc.) will not be as salient as that of <word-orange>. Thus, in this

to clearly illustrate the basic mechanism.

[5]Here, w_{YX} is the synaptic connection strength from neuron X to neuron Y.

way, the presence of <word-red> can indeed affect the outcome of the analogical query. A more precise neural mechanism for this kind of selection among areas (or maps) needs to be investigated further.

Researchers regard the analogical capability as the crux of high-level cognition (see Gentner et al. 2001 for a collection of current work on analogy). However, analogy does not need to be limited to high-level cognition. Recent results suggest that analogy may be needed in perception as well (Morrison 1998), and such an ability emerging in perceptual systems may even be a crucial *requirement* for cognitive development (Chalmers et al. 1992). Then it is not unthinkable that motor function also employ analogy in a similar manner (cf. sensory motor contingency theory by O'Regan and Noe 2001), thus we can then start to understand perception, cognition, and motor function under the unifying framework of analogy.

How can such a diverse functionality be integrated under a single framework of analogical processing? Massive connections exist within and across different functional areas in the brain, and the sensory/motor maps are topologically organized, i.e. nearby neurons are responsive to nearby features in the sensory space (Kohonen 1982; von der Malsburg 1973). Within each map, the feature detectors and the cortico-cortical connections learn to encode the relations (Choe 2001). It is possible that cognitive maps also have a topological organization where nearby areas learn to encode similar concepts, such as semantic maps or episodic memory maps (Miikkulainen 1993). When the sensory, cognitive, and motor maps are connected in an orderly way preserving their local topology, analogies can be drawn within and (more importantly) *across* different functional domains.

Within this huge number of maps specializing in different tasks, a cascade of multiple analogical completions can be going on in parallel, synchronized at each moment by the 40Hz rhythm to hold an instantaneously coherent state (Gray 1999; Mumford 1995). Such a state can then pose as another analogical query, and the process can repeat. When the cascade reaches a motor area, behavior will be generated. Memory content may also enter the analogical cascade, and this quasi-static contribution can prevent the continuously changing input stream from causing random cascades, thereby maintaining a more goal-directed and stable behavior. Specific mechanisms of how the memory content enters the thalamo-cortical loop, and how analogies are archived in long-term memory should be studied further.

Neuroscience research has revealed a lot about perception and motor abilities in the brain, but understanding the cognitive faculty still remains elusive. Investigation into cognitive functions can proceed under the analogical framework, where we can infer the functionality of the higher areas by backtracking the connections to the perceptual and motor areas and study their topology and analogical links. Specific predictions regarding the layout of the higher centers may be made based on the topology of the lower centers and the connection structure between the two, and experiments can then focus on verifying these predictions. For example, there are orientation maps with smoothly changing orientation preference in V1 (primary visual cortex; Blasdel 1992), and there are object maps in TE (temporal area E; Tanaka 1996) that also change smoothly (for example, rotation of a head). The analogical framework predicts that there will be an orderly mapping from V1 to TE that preserve such local topology across different representation spaces.[6] Similar mappings may exist between sensory and cognitive areas, and if such a mapping is found, we can start to understand the abstract cognitive functions based on concrete perceptual architecture.

The strong connection made in this paper between analogical function and specific neural circuitry can help us better understand both. The functionality of the target area of a neuron can be studied to understand *what action occurs* when a neuron detects a certain feature in the input. Such a study can reveal the kinds of relations implemented in the brain, thus providing us with insights into what kinds of analogies are possible. The mechanisms of neural circuits can also be further revealed by carefully designed analogical tests in perception, cognition, and motor function, and also in a combination of these different domains. Different types of unimodal and cross-modal analogical tasks can reveal how the different cortical areas are related and how they invoke each other. In studying such mappings across tasks and modalities, understanding the co-occurrence statistics of natural signals becomes increasingly important as they may give us a hint on how the connections are organized in the brain (Choe 2001; Simoncelli and Olshausen 2001).

Conclusion

In this paper, I analysed the difficulties that the common interpretations of S&C's second order isomorphism can cause in understanding the brain. I proposed an *active* role for representations and relations, and it turned out that collectively they can perform an *analogical function*. An important connection between analogical function and a specific brain circuit was then established, providing support for the new interpretation. This new viewpoint allows us to understand perception, cognition, and motor function under a unifying framework of analogy, and it can help us take a more focused approach in brain and cognitive sciences.

Acknowledgments

I am greatly indebted to Sang Kyu Shin, James Bednar, Risto Miikkulainen, Un Yong Nahm, Marshall Mayberry, Bruce McCormick, and Jyh-Charn Liu for their feedback and encouragement. I would also like to thank the anonymous reviewers for their helpful comments and pointers. This research was supported in part by Texas A&M University, and the Texas Higher Education Coor-

[6] Although the pathway from V1 to TE is not direct, involving V2, V4, and TEO areas, but successive mappings within this path can reveal how V1 and TE are topologically mapped. Also, see Edelman (1995) for more on smooth representation spaces for complex visual objects.

dinating Board ARP/ATP program under grant #000512-0217-2001.

References

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12:3139–3161.

Bliss, T. V. P., and Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, 361:31–39.

Chalmers, D. J., French, R. M., and Hofstadter, D. R. (1992). High-level perception, representation, and analogy. *Journal of Experimental and Theoretical Artificial Intelligence*, 4:185–211.

Choe, Y. (2001). *Perceptual Grouping in a Self-Organizing Map of Spiking Neurons*. PhD thesis, Department of Computer Sciences, The University of Texas at Austin, Austin, TX. Technical Report AI01-292.

Crick, F., and Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in The Neurosciences*, 2:263–275.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, 5:45–68.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449–498.

Edelman, S. (1999). *Representation and Recognition in Vision*. Cambridge, MA: The MIT Press.

Geisler, W. S., Perry, J. S., Super, B. J., and Gallogly, D. P. (2001). Edge Co-occurrence in natural images predicts contour grouping performance. *Vision Research*. 711–724.

Gentner, D. (1989). The mechanisms of analogical learning. In Vosniadou, S., and Ortony, A., editors, *Similarity and Analogical Reasoning*, 199–241. New York, NY: Academic Press.

Gentner, D., Holyoak, K. J., and Kokinov, B. N., editors (2001). *The Analogial Mind: Perspectives from Cognitive Science*. Cambridge, MA: The MIT Press.

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, 24:31–47.

Harth, E., Unnikrishnan, K. P., and Pandaya, A. S. (1987). The inversion of sensory processing by feedback pathways: A model of visual cognitive functions. *Science*, 237:184–187.

Hilgetag, C.-C., O'Neill, M. A., and Young, M. P. (1996). Indeterminate organization of the visual system. *Science*, 271:776.

Hofstadter, D. (1985). Waking up from the boolean dream, or, subcognition as computation. In *Metamagical Themas*, chapter 26. New York, NY: Basic Books.

Hofstadter, D. R., and Mitchell, M. (1994). The copycat project: A model of mental fluidity and analogy-making. In Holyoak, K. J., and Barnden, J. A., editors, *Advances in Connectionist and Neural Computation Theory*. Norwood, NJ: Ablex Publishing Corporation.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

MacKay, D. (1956). The epistemological problem for automata. In Shannon, C. E., and McCarthy, J., editors, *Automata Studies*, 235–251. Princeton, NJ: Princeton University Press.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

Mitchell, M. (2001). Analogy-making as a complex adaptive system. In Segal, L., and Cohen, A., editors, *Design Principles for the Immune System and Other Distributed Autonomous Systems*, to appear.

Morrison, C. T. (1998). *Situated Representation: Solving the Handcoding Problem with Emergent Structured Representation*. PhD thesis, Bringhamton University; State University of New York.

Mumford, D. (1992). On the computational architecture of the neocortex, pt. II, the role of the cortico-cortical loop. *Biological Cybernetics*, 65:241–251.

Mumford, D. (1995). Thalamus. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, 153–157. Cambridge, MA: MIT Press.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '99, Orlando, FL)*, 121–128. AAAI Press.

Newman, J., Baars, B. J., and Cho, S.-B. (1997). A neural global workspace model for conscious attention. *Neural Networks*, 10:1195–1206.

O'Regan, J. K., and Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):883–917.

Shepard, R. N., and Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17.

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216.

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, 3:919–926.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.

von der Malsburg, C. (1973). Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetik*, 15:85–100.

Zeki, S. (2001). Localization and globalization in conscious vision. *Annual Review of Neuroscience*, 24:57–86.

Age Differences in Transitory Cognitive Performance

Sy Miin Chow (smc9x@virginia.edu)
Department of Psychology; University of Virginia; P.O. Box 400400
Charlottesville, VA 22904-4400 USA

John R. Nesselroade (jrn8z@virginia.edu)
Department of Psychology; University of Virginia; P.O. Box 400400
Charlottesville, VA 22904-4400 USA

Abstract

Short-term performance data from a complex computerized cognitive test called SYNWORK1 were examined for age differences in transitory performance fluctuations in samples of 55 older and 57 younger adults. Profile analysis indicated that the older adults' performance trajectories were essentially parallel to those of the younger adults', but with the older adults performing at a consistently lower level on all four subtasks of SYNWORK1. These apparent age differences in level of performance were reduced substantially when a simple graphical approach was used to examine the performance trajectories. These results extend our knowledge concerning the nature of intraindividual variability while illustrating again some of the methodological inadequacies inherent in research comparing age differences in levels of cognitive performance when common statistical assumptions are even mildly violated. The competence of older adults can be underestimated based on a single measure of a group mean, thus leading to further risk of missing important learning strengths of older adults.

Selection and selection effects have received a considerable amount of attention from behavioral scientists (Nesselroade, 1988; Nesselroade & Thompson, 1995) and still remain one of the obstacles researchers must somehow overcome. The primary concerns, however, revolve around selecting a representative sample of participants from the population of interest (e.g. Cronbach, Gleser, Nanda & Rajaratnam, 1972) and valid indicators to represent the underlying construct under study (e.g. Little, Linderberger & Nesselroade, 1999). These, of course, capture only two of the ten possible dimensions defining empirical data in Cattell's data box (Cattell, 1966; Little et al., 1999), namely, the persons and variables dimensions, among other possible design configurations. Another relatively familiar dimension of the data box, occasions of measurement, has also been discussed rather extensively, especially in comparing the relative merits of cross–sectional versus longitudinal research design (e.g. Kraemer, Yesavage, Taylor & Kupfer, 2000). Another kind of selection effect that is inherent in almost any research designs, but has rarely been addressed, is the effect of averaging data across participants or occasions of measurement. In a recently published article, Newell, Liu and Mayer–Kress (2001) question the common practice of averaging data across participants or occasions, presumably to remove the transient, noise–like changes from trial–to–trial, or during the "warm–up" phase at the beginning of a practice session, with the goal of singling out a global learning trend that is characteristic of all the participants across all the trials. As suggested by Lamiell (1981), both idiographic and nomothetic approaches have their own merits in answering certain research questions. However, when a group mean is used as the only index of a group's performance, the end of searching for a global trend in learning does not always justify the means of levelling out the individual differences in this aspect.

Idiographic and Nomethetic Approaches to Modeling Change

Over the past few decades, the importance of an idiographic approach (Allport, 1937; Murray, 1938) to studying human behavior has gained increased recognition. Considerable efforts have been devoted to integrate idiographic and nomothetic approaches in psychological research, thus allowing researchers to capture both the intraindividual variability, and the interindividual differences in various aspects of human behavior (Baltes & Nesselroade, 1979). Repeated assessments of the same individual often yield information on intraindividual variability in aspects thought to be relatively stable over short time-span, such as cognitive abilities and intelligence (see e.g., Horn, 1972; May, Hasher & Stoltzfus, 1993; Stigler, 1994), personality styles and other belief systems (e.g., Shoda, Mischel & Wright, 1994; Kim, Nesselroade & Featherman, 1996), as well as other more transient state-like fluctuations in affective states (e.g., Larsen, 1987; Shifren, Hooker, Wood & Nesselroade, 1997; Mumma 2001).

While many researchers are moving away from performing means comparisons at the aggregate level, the idea of taking a group mean as the unbiased estimator of the group, as well as the population that it represents, is so deeply entrenched in contemporary data analytic techniques that a majority of the between–group comparisons essentially

involve means. Although it is a well–known fact that other measures of central tendency, such as the median, might be a better estimator of a group's performance when certain statistical criteria are not met, or when there are outliers in the data that could potentially skew the results of one's analysis, most of the available statistical tests, such as ANOVA, are based on means comparisons. Even in cases where intraindividual variability is modeled explicitly, such as in growth curve analysis (McArdle & Epstein, 1987; Francis, Fletcher, Stuebing, Davidson & Thompson, 1991; Meridith & Tisak, 1990), conclusions on intraindividual changes in levels of performance still derive primarily from means. Adding to these methodological difficulties of capturing representative idiographic information, of course, are the conceptual difficulties of summarizing the interindividual differences in a way that is helpful for making empirical decisions. A graphical approach is a useful supplement to other more rigorous statistical approaches, as it helps to depict a summary picture of both the intraindividual and interindividual aspects of change.

Transitory Changes in Cognitive Performance as Meaningful Intraindividual Changes

Despite increased awareness of the limitations of using an aggregate measure to represent a group, the same limitations that exist when applied to individual data were not addressed as often. Just as a group mean does not necessarily represent the group as a whole, an individual's mean score is limited in its own way. The transient fluctuations observed during the initial phases of an individual's learning history should not unthinkingly be regarded as "outliers" that ought to be levelled out. Unfortunately, most experimental studies aimed at capturing deterministic dynamics in transitory learning fluctuations as observed during the "warm-up" phase are limited to studies in the area of motor development (e.g. Adams, 1961; Schmidt, 1982; Thelen, 1994). Individuals' cognitive learning curves as reported by some researchers (e.g. Salthouse, Hambrick, Lukas & Dell, 1996) do in general resemble the learning curve of motor skills. However, very few studies have focused on examining the patterns of transitory changes in human cognitive performance and even fewer studies emphasize learning as an ongoing refinement of errors in the face of external perturbations.

Objectives of This Study

In this paper, we examine transitory changes in individuals' short-term adaptive responses and compare these responses among adults of different ages. We also demonstrate some of the inadequacies inherent in comparing adults of different ages on the basis of their group means when common statistical assumptions are even mildly violated. Finally, we present a simple graphical approach that serves as a supplement to means comparisons at a group level, and an alternative to summarizing changes at an individual level.

Method

Participants

This sample consisted of 55 older adults (36 female and 19 male), and 57 younger adults (38 female and 19 male). The older adults' ages ranged from 60 to 93 years ($M = 73.73$, $SD = 7.26$), and the younger adults' from 17 to 28 years ($M = 19.18$, $SD = 1.75$). The younger adults were recruited from the undergraduate participant pool in a southeastern university and were rewarded with course credits for participation in this study. The older adults were recruited from the Charlottesville community through newspaper advertisements and direct solicitation at senior centers . Both age groups rated themselves as in reasonably good health, with self-ratings of health averaged between average to good (on a 4-point scale from 1—excellent to 4—poor, $M = 2.27$ and $SD = 1.00$ for older adults; $M = 1.91$ and $SD = 0.87$ for younger adults).

Materials

The cognitive performance of the participants was assessed using a computer program called SYNWORK1. The SYNWORK1 program is a computerized multi–tasking test environment designed by Elsmore (1994) to examine an individual's ability to perform multiple tasks concurrently. Figure 1 shows the four primary tasks that are measured on SYNWORK1, including a memory task, a self–paced arithmetic task that involves using mouse–click to manipulate the plus and minus panels to adjust the sum from four zeros (0000) into the correct sum of two given numbers, a visual monitoring task, and an auditory discrimination task. Points are given for correct responses and taken away when a task is neglected or performed incorrectly. An individual's total score is constantly updated and is shown in the middle of the four task quadrants.

Procedure

Due to the complexity of the SYNWORK1 program, the participants were first given a short training session during which they practiced the four SYNWORK1 tasks one at a time, followed by a one–minute session during which they practiced the four tasks simultaneously. After the participants were told to strive for their best possible scores, we began recording their performance data on SYNWORK1 over nine consecutive trials, with each trial lasting about one minute.

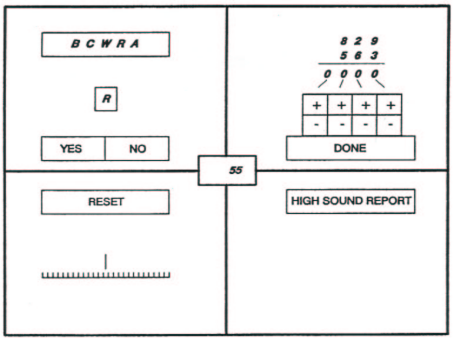


Figure 1: The SYNWORK1 program designed by Elsmore (1994) to test human synthetic work performance

To induce some short–term transitory fluctuations in the participants' performance, we manipulated the difficulty levels of the SYNWORK1 program throughout the nine trials by speeding up or slowing down the timing at which each task appeared. During the more difficult sessions, we also increased the number of addends for the arithmetic task, and the number of target letters the participants had to remember for the memory task. This sudden increase in the difficulty levels of the four tasks was expected to induce a small amount of perturbations in the individuals' responses and the individuals' ability to adapt to these sudden perturbations was taken as an indication of their adaptive behavioral patterns. During a particular trial, the SYNWORK1 program was governed by one of three possible sets of parameters we classified as easy, mid–difficulty, and difficult. The participants were exposed to one of two sequences of altering difficulty levels in the order of {M, E, M, D, M, E, M, D, M}, or {M, D, M, E, M, D, M, E, M}, where M represents mid–difficulty, E represents easy, and D represents difficult sessions. The total number of correct memory, arithmetic, visual, and auditory responses for each of the nine trials were used as the primary dependent variables in this study to determine if the two age groups had different performance profiles on each of the tasks.

Results

Prior to analysis, the participants' task scores over nine trials were screened for outliers and departure from other statistical assumptions. Several outliers were detected among younger adults who performed too poorly on the memory task, older adults who performed too well on the arithmetic task and the visual task, and among younger adults on the auditory task due to the restricted range of scores observed in that age group. These outliers were retained in subsequent data analysis as they were thought to reflect reasonable range of fluctuations in performance. No missing values were observed in the data. When subject to a MANOVA test, the task scores of individuals exposed to the two sequences of difficulty levels were not statistically different from each other. After confirming that, a profile analysis was used to evaluate the differences in level, parallelism, and flatness of the two groups' responses on the four tasks, with age group as the independent variable.

Profile Analysis

The descriptive statistics of the two age groups are presented in Table 1. Consistent with findings reported in the aging literature, the younger adults were found to perform at a significantly higher level on the memory task, $F(1,110) = 132.17$, $p < .001$; the arithmetic task, $F(1,110) = 223.69$, $p < .001$, and the auditory task, $F(1,110) = 106.56$, $p < .001$, as determined by using Wilk's criterion. The visual task was the only task that did not show significant difference in levels between the two age groups.

Table 1: Descriptive Statistics of the Older and Younger Adults.

	Mean		**Standard Deviation**	
	Old	Young	Old	Young
Memory	5.77	9.47	2.92	2.20
Arithmetic	0.21	1.84	0.60	1.45
Visual	70.80	72.22	22.51	21.49
Auditory	0.81	1.78	0.83	0.76

When averaged across groups, scores on all four tasks were found to deviate significantly from flatness by Hotelling's criterion on all four SYNWORK1 tasks ($p < .01$ for all). This simply confirmed that the experimentally imposed perturbations did lead to some fluctuations in the participants' performance. There might also be some learning taking place, as shown by the slight increase in means over the nine trials. Using Wilk's criterion, however, the older and younger adults' profiles were not found to deviate significantly from parallelism, except on the arithmetic task, $F(8, 103) = 2.98$, $p = .005$. The arithmetic task might be too cognitively demanding for the older adults, thus causing them to avoid the task altogether. Aside from this, the two age groups did not react to the experimental perturbations in a statistically different way over the nine trials.

Graphical Representation of Intraindividual Variability

The conceptual difficulties of summarizing information at an individual level can be illustrated using Figure 2, in which all the individuals' total scores on

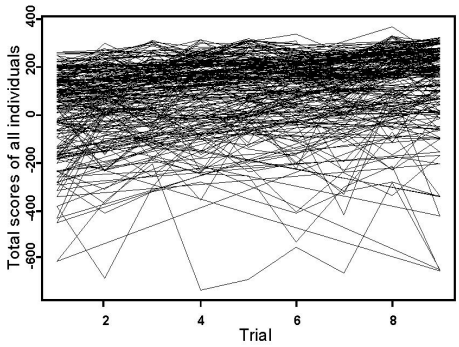


Figure 2: The total scores of all the individuals in the sample.

SYNWORK1 are shown over the nine trials. Different individuals have their own distinct trajectories, making it hard to summarize the performance of the group as a whole. The conventional means plot (see Figure 3) may be helpful in providing a summary picture, but could also mask important information in the data. To further examine the performance of the two age groups as captured by other descriptive measures, the 10th, 50th and 90th percentiles of the two groups' scores on the four tasks were plotted and compared using S–Plus' Hmisc library (Harrell & Alzola, 2001). Due to space limitations, we chose to include only the means and percentiles plots for the memory and auditory tasks here.

As shown in the plots in Figure 4, the age differences in performance levels reduced substantially when the percentiles plots were used to represent the performance of the two age groups [1]. When compared to adults from the same age group, older adults who performed at the 90th percentiles on the memory and arithmetic tasks (omitted here) were found to show performance that closely resembled younger adults whose scores were near the median levels of their peers[2]. Older adults in this category were observed to make more marked improvements on the memory and arithmetic tasks toward the end of the experiment. A similar pattern was observed on the auditory task, except that the relatively well performing older adults (i.e., the 90th percentile group), started out with performance trajectory that was identical to the trajectory of younger adults who performed at the median level, but after the 7th trial, caught up to the performance of younger adults who performed at the 90th percentile level and performed at exactly the same optimal level after the 7th trial. The younger adults who performed

[1]We also plotted the trajectories of younger and older adults at less extreme percentile levels (e.g. 25th and 75th percentiles). A similar, but smaller magnitude of reduction in age differences was observed.

[2]The percentile levels were calculated separately for each task and each age group.

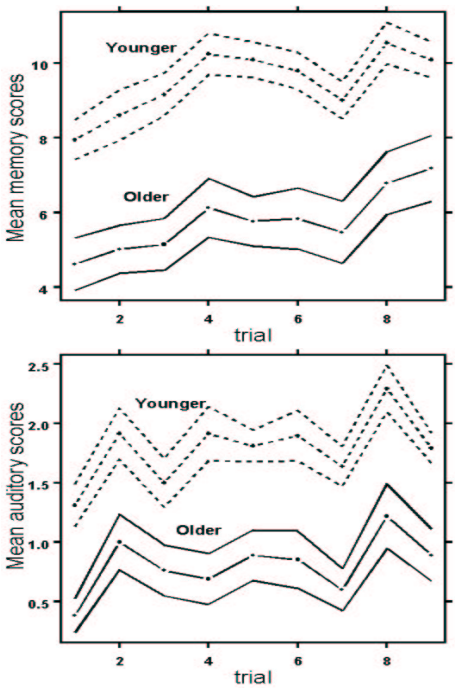


Figure 3: The means and the 95% confidence limits of the older and younger adults' scores on the memory and auditory tasks.

at the 90th percentile levels were performing at the optimal level on the memory, visual and auditory tasks very early on, achieving the maximum possible scores during almost all of the trials. Their only source of improvement in scores derived primarily from the self–paced arithmetic task, on which their scores continued to improve throughout the course of this experiment. On the other hand, older adults at the 90th percentile level were observed to have rather uniform improvements on all four tasks, and more trials were required before they could attain the same level of performance the well performing younger adults could achieve at an early phase.

Another important source of variability constituted by age stemmed from the lack of clear improvements observed among older adults who performed at the 10th percentile level. The relatively poor performing younger adults (those at the 10th percentile level) were able to capitalize on the memory task and auditory task as time progressed much better than older adults of comparable performance level. Nevertheless, older adults at the 10th percentile level did demonstrate their own learning strengths on the visual task—they attained a considerable amount of improvement on the visual task from trial one to trial two, and maintained a rather persistent level of performance before a sharp decrement in scores was observed during and after the 8th trial, presumably due to fatigue and decreased attention span. When the

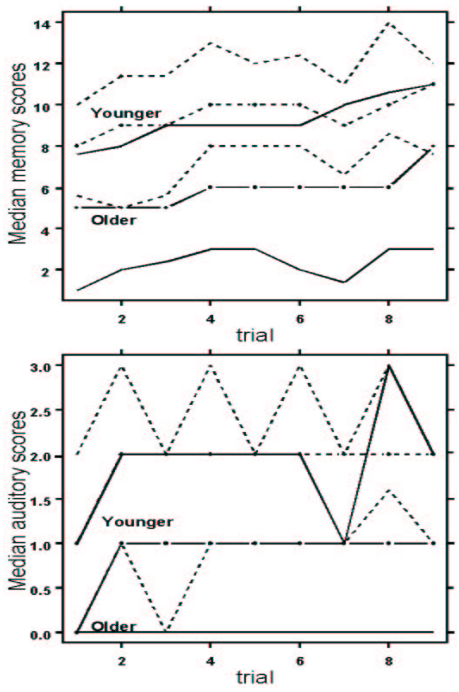


Figure 4: The 10th, 50th and 90th deciles of the older and younger adults' memory and auditory scores.

plots of individuals from less extreme percentile levels (e.g. 25th and 75th percentiles) were examined, similar fluctuation patterns that differed slightly in levels were observed, revealing interesting individual differences in performance fluctuations within, as well as between different age groups.

Discussion

In this study, SYNWORK1 was used as an active interface for capturing the age differences in performance fluctuations when individuals were faced with ongoing external perturbations. As Jones and Conrad's (1933) quotation of Thorndike's remark put it, "...individual differences amongst those of the same age...enormously outweigh differences between ages" (p. 258–59). The common approach of comparing age differences in levels of cognitive performance by using group means inevitably under–represents the complexity underlying the variability in performance both within and between different age groups, especially when the two age groups have unequal variances in many respects. Due to the lack of a clear definition of what constitutes a normative representation of the older adults population, the issue of identifying and eliminating outliers becomes tricky. Results from our profile analysis showed findings that were consistent with those reported in the literature (e.g. Erber, 1976). Essentially, adults of different age groups were found to exhibit similar, significant improvements on cognitive or intelligence tests as practice effects accumulated, but the younger adults' level of performance was almost always higher than the older adults' on all the measurement occasions.

Using group means and the changes in means as the sole indicators of the performance of these older and younger adults might be informative in its own way, as demonstrated by the profile analysis in this study, but fails to acknowledge the differences in performance profiles among individuals of the same age group. In fact, the younger adults in this study, who were all recruited from the same university and were often thought of as representing a rather homogeneous group, showed different dynamics in their performance fluctuations when the trajectories of individuals from different percentiles were compared. The trajectories plotted using the percentiles of the younger and older adults also revealed very consistent patterns in performance fluctuations that reflected our experimentally imposed alterations in task difficulty levels very accurately. More importantly, the age differences in levels of performance were reduced substantially when the percentile scores of these two age groups were examined, revealing some of the older adults' unique strengths in learning that started surfacing at a relatively more gradual pace than for the younger adults.

Of course, younger adults of the higher ability group might be encountering ceiling effects on those tasks from a very early phase. In addition, with the data from the present study, there is no way for us to determine whether the high–performance younger adults will resemble the high–performance older adults in any way when they get older. However, by using a simple graphical approach, we presented some of the inadequacies of using group means as the only representation of the dynamics of the group as a whole because a researcher may risk bypassing some of the interesting dynamics within the group by not looking at the results offered by other alternative methods.

Acknowledgments

This research was supported by the National Institute Aging grant 5 R01 AG18330 awarded to the second author.

References

Adams, J. A. (1961). The second facet of forgetting: A review of warm–up decrement. *Psychological Bulletin*, *58*, 257–273.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal

research in the study of behavior and development. New York: Academic Press.

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of Multivariate Experimental Psychology* (1st ed.). Chicago, IL: Rand McNally.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measures*. New York: Wiley.

Elsmore, T. F. (1994). SYNWORK1: A PC–based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 421–426.

Erber, J. T. (1976). Age differences in learning and memory on a digit–symbol substitution task. *Experimental Aging Research*, *2*(1), 45–53.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Davidson, K. C. & Thompson, N. M.(1991). Analysis of change: Modeling individual growth. *Journal of Consulting and Clinical Psychology*, *59*(1), 27–37.

Harrell, F. E. & Alzola, C. F. (2001). *An introduction to S–Plus and to the Hmisc design libraries*. Unpublished manual.

Horn, J. L. (1972). State, trait and change dimensions of intelligence. *British Journal of Educational Psychology*, *42*, 159–185.

Jones, H.E., & Conrad, H.S. (1933). The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty, *Genetic Psychology Monographs*, *13*.

Kim, J.E., Nesselroade, J. R. & Featherman, D. L. (1996). The state component in self-reported worldviews and religious beliefs of older adults: The MacArthur successful aging studies. *Psychology and Aging*, *11*(3), 396–407.

Kraemer, H. C., Yesavage, J. A., Taylor, J. L. & Kupfer, D. (2000). How can we learn about developmental processes from cross–sectional studies, or can we? *American Journal of Psychiatry*, *157*, 163–171.

Lamiell, J. T. (1981). Toward an idiothetic psychology of personality. *American Psychologist*, *36*(3), 276–289.

Larsen, R. L. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology*, *52*(6), 1195–1204.

Little, T. D., Lindenberger, U. & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*(2), 192–211.

May, C. P., Hasher, L. & Stoltzfus, E. R. (1993). Optimal time of day and the magnitude of age differences in memory. *Psychological Science*, *4*(5), 326–330.

McArdle, J. J. & Epstein, D. B. (1987). Latent growth curves within developmental structural equation. *Child Development*, *58*(1), 110–133.

Meridith & Tisak (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Mumma, G. H. (2001). Increasing accuracy in clinical decision making: Toward an integration of nomothetic-aggregate and intraindividual-idiographic approaches, *The Behavior Therapist*, *24*(4), 77–94.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Nesselroade, J. R. (1988). Some implications of the trait–state distinction for the study of development over the life span: The case of personality. In P. B. Baltes, D. L. Featherman & R. M. Lerner (Eds.), *Life–span development and behavior* (Vol 8). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nesselroade, J. R. & Thompson, W. W. (1995). Selection and related threats to group comparisons: An example comparing factorial structures of higher and lower ability groups of adult twins. *Psychological Bulletin*, *117*(2), 271–284.

Newell, K. M., Liu, Y,& Mayer–Kress, G. (2001). Time scales in motor learning and development. *Psychological Review*, *108*(1), 57–82.

Salthouse, T. A., Hambrick, D. Z., Lukas, K. E. & Dell, T. C. (1996). Determinants of adult age differences on synthetic work performance. *Journal of Experimental Psychology: Applied*, *2*(4), 305–329.

Schmidt, R. A. (1982). *Motor control and learning*. Champaign, IL: human Kinetics.

Shifren, K., Hooker, K., Wood, P. & Nesselroade, J. R. (1997). Structure and variation of mood in individuals with Parkinson's disease: A dynamic factor analysis. *Psychology and Aging*, *12*(2), 328–339.

Shoda, Y., Mischel, W. & Wright, J. C. (1994). Intraindividual in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Society Psychology*, *67*(4), 674–687.

Siegler, R.S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*, *3*, 1–5.

Thelen, E. & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and acion*. Cambridge, MA: MIT Press.

Reminiscence and Arousal: A Connectionist Model

Eric Chown (echown@bowdoin.edu)
Department of Computer Science, 8650 College Station
Brunswick, ME 04011 USA

Abstract

In recall tasks, increased levels of arousal soon after presentation time leads to short-term performance that is contradictory to standard memory models. Despite the fact that long-term recall is excellent in such situations, short-term recall is poor, worse than in the long-term case. This article presents a model, based upon Hebb's cell assembly construct, to account for this puzzling data. The system, called MultiTrace, has previously been used to model a lexical priming task and was adapted with only minor changes for this task.

Introduction

The relationship between arousal and short-term memory has presented a problem for standard memory models for nearly 40 years. In a pair of studies in the early 1960s Kleinsmith and Kaplan (1963; 1964) found that while there is a positive correlation between increases in arousal and long-term memory, in the short-term case the reverse is true. Further, in cases of where arousal markedly increases, short-term recall is substantially worse than long-term recall, an effect called "reminiscence." Because these results challenge standard memory models in which short-term memory is necessarily stronger than long-term memory, they have often been criticized, but just as often have been replicated (Eysenck, 1977; Weingartner & Parker, 1984; Revelle & Loftus, 1990). To date, no widely accepted explanation has ever been provided to account for the data (see, for example, Revelle & Loftus, 1990). However, more modern views of memory, updated with an increased understanding of the underlying neural mechanisms involved, can now provide a plausible explanation of the data. This article presents an existing model that can account for the data in a way that is surprisingly close to the original account of Kleinsmith and Kaplan.

The model used in this article, called MultiTrace (Sonntag, 1991; Chown, 1994; Forbell & Chown, 2000), is a variant on the cell assembly construct proposed by Hebb (1949) and later extended by Kaplan, et. al. in their TRACE system (1991). What makes the cell assembly model so attractive for this task is that, unlike many other neural network models, cell assemblies have complex temporal dynamics directly affected by the physiological properties of neurons. MultiTrace is an extension of the single cell assembly TRACE model to allowing sequence learning through the inclusion of multiple cell assemblies.

This article will begin with a discussion of the Kleinsmith and Kaplan data and some of the theoretical problems it raises. Next the MultiTrace model is presented, highlighting parts of the model directly relevant to this task. Finally, the original experiments are modeled in MultiTrace and the results are compared to the originals.

Reminiscence

The learning paradigm used by Kleinsmith and Kaplan was a simple paired-associate task. In the first experiment, subjects were shown a word for four seconds and then the same word and a number for another four seconds. The recall task was to remember the number when the word was given as a cue. Changes in galvanic skin response (GSR) deflection were used as the measure of arousal response. These changes were sorted into "high" and "low" categories for each subject. In the long-term case (one week), Kleinsmith and Kaplan got the results that they expected – recall was better for words in the "high" category. The surprising result was that in the short-term cases (2 minutes and 20 minutes) this was not the case, and further, recall in the "high" categories was worse than in the long-term case. When the arousal deflection was "low", the recall curves generated were what would normally be predicted – short-term recall was very good but decayed over time (Figure 1). In the "high" case, however, Kleinsmith and Kaplan found that the recall curve was essentially an inverted U (Figure 1). To address criticism of the original study, Kleinsmith and Kaplan replicated their work the following year, but used nonsense syllables instead of words, and got qualitatively similar results.

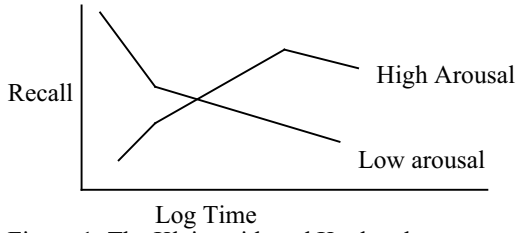


Figure 1: The Kleinsmith and Kaplan data.

The major question raised by the Kleinsmith and Kaplan data is why recall is so poor in the short-term case in the high deflection condition. Among other things, the results would seem to discount the notion that long-term memory is just a decayed version of short-term memory, or that memories transition from short-term into long-term storage.

Kleinsmith and Kaplan proposed that the neural circuits involved in the task rapidly reverberated under conditions of increased arousal and were relatively unavailable because of that reverberation. They went on to theorize that this same reverberation was responsible for the better long-term recall due to greater perseverative consolidation. This explanation was discounted due largely to the lack of an accepted mechanism that could account for the unavailability of the circuits in the short-term. Kleinsmith and Kaplan proposed neural fatigue as the mechanism, but this is a property of neurons that only started to become accepted more than 20 years later (Ito, 1992; Artola & Singer, 1993) and was criticized at the time for being implausible.

Most of the alternative explanations proposed in the interim involved some combination of consolidation and interference in conjunction with separate short and long-term memory mechanisms. None of them were satisfactory (Revelle & Loftus, 1990), and work in the area appears to have died out due to the lack of attractive theories. The model presented here is closely related to Kleinsmith and Kaplan's original proposal. The neural circuits in this case are cell assemblies. In this paradigm learning comes as the result of correlated neural activity (this is usually called Hebbian learning). One effect of an increase in arousal is to generate intense and focused activity in the brain (Oades, 1985). In a Hebbian paradigm one byproduct of intense activity is that the areas of focused activity experience more correlated activity and therefore more learning. As this activity stretches out over a few seconds it essentially serves as the memory consolidation period. On the other hand, this intense activity also tends to fatigue the neurons that have been repeatedly firing. The net result is that these fatigued cell assemblies temporarily require an unusual amount of stimulation in order to become reactivated.

The MultiTrace Model

Hebb developed the cell assembly construct to address questions concerning the temporal nature of neural processing. Hebb needed a way to explain how neurons could hold information over time (e.g. the psychological concept of "set") even though they essentially pass through information. He solved this problem by proposing that cell assemblies consist of a large collection of neurons that are highly interconnected. These connections form a kind of loop that enables the cell assembly to effectively hold information through the reverberation of the loop once it becomes active. In this way cell assemblies are neural analogs of "symbols." Hebb's theory was rejected after early models showed that the recurrent connections tended to lead to out of control activity (Rochester, et. al., 1956). This was because Hebb had cautiously omitted inhibition from his model because no direct physiological evidence existed for it at the time. More recently, however, cell assembly theory has undergone something of a revival as experimental evidence for their existence has been found (Amit, 1995) and several models have been proposed that extend Hebb's original conception (Kaplan, et al, 1991; Hetherington & Shapiro, 1993; Amit, 1995; Horn, et al., 2000).

TRACE

MultiTrace is based upon one such model, the TRACE (Tracing Recurrent Activity in Cognitive Elements) model of Kaplan et. al. (1991). TRACE is based on the idea that there must be counterbalancing forces to offset the tendency of cell assemblies to continually reverberate. The most important of these forces are *inhibition*, which provides a mechanism for selection and for competition with other cell assemblies, and *fatigue* which ensures that cell assemblies do not stay active indefinitely.

TRACE is modeled as a set of difference equations that capture the biological properties of the population of neurons that comprise the cell assembly. The equations are similar to population models used by biologists. Rather than trying to work out the interactions of neurons individually, the Kaplan group decided to work at a slightly higher level in order to understand their collective behavior. TRACE units, therefore, are more complex than typical neural network units, but are still at a higher level than neurons.

The crux of any cell assembly model is that activity in a cell assembly is the basis of cognitive processing. Perception, for example, would correspond to a cell assembly reaching some internal threshold of activity (defined as the percentage of active neurons at any given time step). TRACE added several theoretical constructs to Hebb's theory that are relevant to this article; they are inhibition, fatigue, and short-term connection strength.

Inhibition serves numerous roles in cognition, including selection and providing a mechanism for perceptual competition. These will be discussed in more detail in the discussion of the MultiTrace extension to TRACE.

Fatigue provides a kind of shut-off mechanism to cell assemblies, ensuring that the cognitive system does not literally get stuck on a single thought. It has also been

speculated that fatigue may play a central role in learning in much the same way that muscle fatigue is crucial to getting stronger. Neurons, like muscles are physical systems that require fuel (e.g. transmitter substances) to work efficiently. With extreme usage that fuel supply can run out and essentially damage the system. With muscles it is the repair of the damage that makes the muscle stronger. It may be that neural fatigue is a kind of signal to rewire synapses in order for the system to run more efficiently in the future.

The important function of short-term connection strength with regard to this article, is that it provides a mechanism for short-term memory. Short-term connection strength (STCS) is based on post-tettanic potentiation, the property of neurons that once they fire they are temporarily more likely to fire again. In other words, once a cell assembly has been active, for a short time it will be easily reactivated (or recalled). STCS's counterpart in TRACE is long-term connection strength (LTCS) which comprises the long-term structure of the brain and is what changes with learning.

The TRACE equations are shown in Table 1. As difference equations, they are updated on each time step (set at 10 milliseconds in TRACE).

Table 1: The TRACE equations

Update Equations	**Delta Equations**
$A(t+1) = A(t) + \Delta A$ $F(t+1) = F(t) + \Delta F$ $S(t+1) = S(t) + \Delta S$ $L(t+1) = L(t) + \Delta L$	$\Delta A = P - M$ $P = (A + \bar{A}\, I)\bar{A}\, V$ $M = (A^{\theta_L} + A\bar{A}^{\theta_C})\bar{V}$ $\Delta F = \varphi_g A\, \bar{F} - \varphi_d F$ $\Delta S = \sigma_g A\, \bar{S} - \sigma_d S$ $V = \frac{1}{v}(S + L)\bar{F}\, A_R$ $I = I^{exc} - I^{inh}$ (see text)
θ_l : unit loss θ_c : inhibitory competition v : normal factor φ_g : fatigue growth φ_d : fatigue decline σ_g : STCS growth σ_d : STCS decline	A: activity F: neural fatigue S: STCS L: LTCS I: network input A_R: Arousal Equation Note: $\bar{X}$ denotes quantity (1 – X)

MultiTrace

MultiTrace extends TRACE by putting TRACE units into a larger architecture consisting of multiple cell assemblies. In addition to the properties of the individual units that come from the TRACE model, units also have properties associated with the architecture of the brain. There are two properties that are important with regard to this article and they have previously been used with MultiTrace to model a lexical priming task (Forbell & Chown, 2000). The first property is what Kinsbourne (1982) called "the functional distance principle." This is the property that similar concepts tend to interfere with each other more than dissimilar ones. In neural terms, similar concepts are processed near each other in the brain and will tend to inhibit each other. This provides a means for perceptual competition among other things. The second property is that cortex consists of a large number of layers. It is possible to think of these layers vertically, with the perceptual layer on the bottom and more abstract layers higher in the hierarchy. In the original TRACE paper it was proposed that each layer consists of cell assemblies with different temporal properties (in the context of the model this is achieved by varying the parameter settings).

Figure 2 is drawn from (Forbell & Chown, 2000) and demonstrates both of these properties. The lexical layer is more abstract than the phoneme layer and is therefore at a higher level in the cognitive hierarchy. In this case the lexical layer is one level up since words consist of sequences of phonemes. Each layer has slightly different properties (for example, perceptual layers will need to recover from fatigue very quickly so the perceptual system can always be ready for the next input) and similar items within a layer tend to compete with each other through lateral inhibition. In the figure, for example, "BLAST" will inhibit "BLACK" more than it will inhibit "BLUES" since it is more similar to "BlACK."

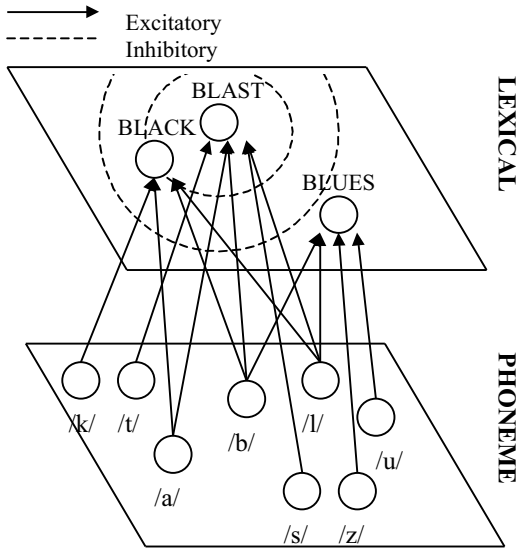


Figure 2: The two-tiered connectionist architecture. Similar words are processed near each other and will tend to inhibit each other based upon distance.

The main extensions to the original TRACE model added in MultiTrace is to use a more complex input function that combines several sources of inhibition as well as excitation. The input equations found here are the same as in (Forbell & Chown, 2000) and more details on the derivations can be found there.

For clarity a unit's input can be divided into inhibitory and excitatory components. Because of the functional distance principle, neighboring units will tend to inhibit each other. The amount of inhibition is a function of the distance between units and the activity and fatigue of the inhibiting unit:

$$I_{jk}^{inh} = \frac{A_j(1 - F_j)}{D_{jk}}$$

j: source unit
k: target unit
D: distance

There is an additional inhibitory factor in the model termed "regional inhibition" which is designed to limit the total activity in any given cognitive layer. Therefore the total inhibition for a unit, k, is the sum of the inhibition from other units, plus its regional inhibition:

$$I_k^{inh} = \frac{1}{L}(\sum_{j=1}^{n} I_{jk}^{inh}) + R(\sum_{i=1}^{n} A_i)$$

n: number of units in a layer
R: regional inhibition factor
L: lateral inhibition factor

Excitatory input is computed in a conventional connectionist manner. The only difference is that units have both long-term and short-term connections between them:

$$I_{jk}^{exc} = (LTCS_{jk} + STCS_{jk})A_j$$

$$I_k^{exc} = \sum_{i=1}^{n} I_{jk}^{exc}$$

j: source unit, k: target unit
n: number of incoming connections for unit *k*

Once the excitatory input is computed it is run through a standard sigmoid function to force it to values between 0 and 1.

Learning in MultiTrace is based upon the Hebbian learning rule that connections between units are increased when they fire nearly simultaneously. This is modulated by the fatigue (raised to the power *P*) of the unit as well as a learning rate, Λ.

$\Delta LTCS_{kj} = \Lambda * A_k * A_j * F_k^P$

Simulation and Results

The experiments simulated in this article are from (Kleinsmith & Kaplan, 1964). The only changes to MultiTrace from (Forbell & Chown, 2000) were in parameter settings related to operating at a different level of cognition than the perceptual experiments in that work. The parameters internal to units are largely based upon those in the original TRACE paper (Kaplan, et. al., 1991) since it was meant to model higher cognition. With regard to the experiments, the important parameters are the ones affecting fatigue, short-term connection strength and arousal. The values are shown in Table 2. With regard to arousal, the value shown is taken to be a baseline that varies experimentally as described below.

Table 2: Unit Parameters

Parameter	Description	Value
Φ_g	Fatigue Growth	0.007
Φ_d	Fatigue Decline	0.0001
σ_g	STCS Growth	0.1
σ_d	STCS Decline	0.001
A_r	Arousal	0.95

In terms of the architecture selected, a two-tiered system was used similar to the one shown in Figure 2. The reasoning is that nonsense syllables were selected specifically to be unlike normal words. This means there should be no perceptual competition between the nonsense syllables and the associated digits. Further, connections between the layers should be relatively uniform and weak. One difference between the layers is that nonsense syllables represent "unlearned" or weakly connected cell assemblies. Digits, on the other hand, are frequently used, and thus the cell assemblies that represent them should be tightly connected. In the original TRACE paper it was proposed that unlearned cell assemblies have an internal long-term connection strength (LTCS) of 0.2, while well-learned ones have an LTCS of 0.5. In the experiments that follow, this is the only difference between the two layers.

The basic format of a given run of a simulation is simple: a unit was randomly selected from the first layer and was stimulated. Meanwhile, a starting arousal level was also generated. Then two seconds later a random arousal deflection was generated. This was categorized as either "high" or "low" by comparing it against the median arousal level. Two seconds after that the original unit was again stimulated along with a randomly selected unit from the second layer. In the short-term case the network was then allowed to run for two minutes of simulated time. Then arousal was again set randomly (it was found that there was no correlation between arousal at presentation and arousal at recall

(Kaplan & Kaplan, 1970)) and the first unit was re-stimulated. If the second unit became active due to stimulation from the first, it was categorized as successful recall. The procedure was the same in the long-term case except that factors such as fatigue and STCS were simply reset to normal levels to simulate the passage of time. In addition, the network learned (in the form of changes in inter-unit LTCS) in the long-term, but not the short-term case. This is because consolidation data shows that the physical process of long-term memory generally takes at least 20 minutes Miller & Marlin, 1984).

The 20-minute recall case was dropped for a variety of reasons. One reason is that the simulation essentially runs in real time. This means it is not possible to run a significant number of trials as compared to the 1000 trials in the other cases. A second reason is that the 20-minute case mixes short-term and long-term memory processes. There is no a priori way of knowing how much the recall rates at those times are due to either factor. The tested cases are purely short-term memory on the one hand, and purely long-term memory on the other, and therefore can generate more meaningful results. There is no obvious way to set the learning rate in the 20-minute case and trying to determine one experimentally would be extremely difficult due to the time-scale involved.

The architectural parameters relating specifically to MultiTrace are given in Table 3.

Table 3: MultiTrace Parameter Settings

Parameter	Description	Value
R	Regional Inhibition	0.1
L	Lateral Inhibition	3
Λ	Learning Rate	0.03
P	Fatigue Power	3

To test the robustness of the model, in each simulation run a new "subject" was generated by randomly perturbing all of the relevant model parameters. For example, if the parameter's ideal value was set at 0.1 a new one was generated for each run by using a Gaussian distribution with a mean of 0.1 and a standard deviation of 0.01. In all, 1000 runs were done for both the long and the short-term cases. The results are shown in Table 4.

Table 4: Recall Rates

Arousal	2-minute	1-week
Low	0.55	0.25
High	0.11	0.50

The results are similar to those obtained by Kleinsmith and Kaplan, though exact comparisons cannot be drawn since they published curves, but not the actual numbers. Further, pursuit of an exact match is probably a fruitless enterprise as the results were later reinterpreted with different scoring methodologies (Kaplan & Kaplan, 1970). The conclusion of that study was that the *trends* in the results were consistent across experiments and scoring, but with a fair degree of variation. The important trends being that initial recall is high in the low arousal case and then declines, and that the reverse is true in the high arousal case.

It is important to emphasize that the parameters for this work were drawn from previous sources. The TRACE parameters in Table 2 were taken directly, without modification, from an extension to TRACE where arousal was added (Chown, 1994). These parameters are different than those used in (Forbell & Chown, 2000) due to the fact that different levels of cognition were being modeled. In the previous case the individual units represented phonemes; in the current case the cell assemblies represent entire syllables. Since syllables are comprised of sequences of phonemes the time course of activity must be longer. In terms of the MultiTrace architectural parameters in Table 3, the only parameter that varied from (Forbell & Chown, 2000) was the learning rate. It is reasonable to believe that the learning rate for lexical material is different than the learning rate at higher levels of cognition, such as at the word level.

Discussion

This work serves two purposes. First, an existing biologically grounded model was used to address a theoretical problem in the memory and arousal literature. The results support Kleinsmith and Kaplan's theory that their data can be explained in terms of reverberating neural circuits. Second, their data provides further constraints in exploring the temporal dynamics of neural processing in cognition and the development of cell assembly-based models.

TRACE and MultiTrace were developed as modern versions of Hebb's cell assembly construct. The individual parts of each model were theoretically motivated as part of a general cognitive theory. The fact that MultiTrace was able to model the Kleinsmith and Kaplan results with only minor parameter changes lends credence to Kleinsmith and Kaplan's original supposition about the underlying cause of their results. In turn, their data, which has defied other explanations for four decades, shows that the physiology of the brain, including architectural factors such as wiring, are critical in fully understanding human learning.

Hebbian learning has become increasingly popular in connectionist models in recent years and the paradigm explored in this paper provides evidence of how it

might be useful in a general theory of learning. The postulate that learning comes as the result of correlated neural activity implies that variation in learning can largely be explained in terms of factors that impact that activity. Arousal is a primary example of such a factor. Cell assembly models such as MultiTrace have the potential to explore this idea by modeling the dynamics of neural activity at a fairly high level while still incorporating physiological constraints. These constraints are often architectural, depending on factors that are not easily modeled in many other types of neural networks where, for example, the temporal dynamics of the individual units are essentially irrelevant.

Dynamic models are gaining in popularity, and control issues, such as those involved in the TRACE and MultiTrace models, will be central in their continuing development. Time dependent data, such as the Kleinsmith and Kaplan data is crucial to the development of these models as it provides a source of constraints on how the models must operate. This work is part of an ongoing process of collecting such constraints and using them to develop and calibrate the model. The diversity of the tasks involved – e.g. modeling priming and interference effects in phoneme processing versus modeling paired-associate learning – is crucial in showing the generality and power of the model.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0092605. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

References

Amit, D.J. (1995). The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences*, 18(4): 617-657.

Artola, A., & Singer, W. (1993). Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends in Neurosciences*, Vol. 16, No. 11, pp. 480-487.

Chown, E. (1994). Consolidation and learning: A connectionist model of human credit assignment. Doctoral dissertation. The University of Michigan.

Eysenck, M.J. (1977). *Human Memory: Theory, Research and Individual Differences*. Pergamon Press.

Forbell, E. & Chown, E. (2000). Lexical contact during speech perception: A connectionist model. In *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society*.

Hebb, D.O. (1949). The Organization of Behavior. John Wiley.

Hetherington, P.A., & Shapiro, M.L. (1993). Simulating Hebb cell assemblies: the necessity for partitioned dendritic trees and a post-not-pre LTD rule. *Network*, 4, 135-153.

Horn, D., Levy, N., Meilijson, I., & Ruppin, E. (2000). Distributed synchrony of spiking neurons in a Hebbian cell assembly. In S.A. Solla, T.K. Leen, and K.R. Muller (eds.), *Advances in Neural Information Processing Systems*, 12, 129-135. MIT Press.

Ito, M. (1992). Posttetanic depression. In L.R. Squire (ed.) *Encyclopedia of Learning and Memory*. New York: Macmillan Pub. Co.

Kaplan, S. & Kaplan, R. (1970). The interaction of arousal and retention interval: Ipsative vs normative scoring. *Psychonomic Science*, 19, 115-117.

Kaplan, S., Sonntag, M. & Chown, E. (1991). Tracing recurrent activity in cognitive elements (TRACE): A model of temporal dynamics in a cell assembly. *Connection Science*, 3, 179-206.

Kinsbourne, M. (1982). Hemispheric specialization and the growth of human understanding. *American Psychologist*, 37(4), 411-420.

Kleinsmith, L.J., & Kaplan, S. (1963). Paired-associate learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology*, 65, 190-193.

Kleinsmith, L.J. & Kaplan, S. (1964). Interaction of arousal and recall interval in nonsense syllable paired-associate learning. *Journal of Experimental Psychology*, 67, 124-126.

Miller, R.R., & Marlin, N.A. (1984). The physiology and semantics of consolidation. In H. Weingartner, & E.S. Parker (Eds.), *Memory Consolidation: Psychobiology of Cognition*, Hillsdale, NJ: Lawrence Erlbaum.

Oades, R.D. (1985). The role of noradrenaline in tuning and dopamine in switching between signals in the CNS, *Neuroscience & Behavioral Reviews*, Vol. 9

Revelle, W. & Loftus, D.A. (1990). Individual differences and arousal: Implications for the study of mood and memory. *Cognition and Emotion*, 4, 209-237.

Rochester, N., Holland, J.H., Haibt, L.H., & Duda, W.L. (1956). Tests on a cell assembly theory of the action of the brain, using a large digital computer. *IRE Transactions on Information Theory*, IT-2:80—93.

Sonntag, M.. (1991). Learning sequences in an associative network: A step towards cognitive structure. Doctoral Dissertation. The University of Michigan.

Weingartner, H., & Parker, E.S. (Eds.) (1984). *Memory Consolidation: Psychobiology of Cognition*. Hillsdale, NJ: Lawrence Erlbaum.

How Conceptual Metaphors are Productive of Spatial-Graphical Expressions

Timothy C. Clausner (Clausner@HRL.Com)
HRL Laboratories, LLC
Human Centered Systems Department, 3011 Malibu Canyon Road
Malibu, CA 90049 USA

Abstract

The theory of conceptual metaphors is adopted in which conceptual relations are productive of linguistic metaphorical expressions. Conceptual metaphors vary in their degree of productivity according to semantic principles. Spatial-graphical expressions of non-spatial concepts are investigated providing evidence that they are instantiations of metaphors. For three cases of differing productivity it is argued that the same semantic principles which result in metaphor productivity for linguistic expressions also result in spatial-graphical expressions.[1]

Background

Language gives us words, and constructions made of words, to talk about abstract concepts. We find in space, conventional shapes and organizations of shapes which also convey abstract concepts. These representations in space are typically experienced visually, but not exclusive of other experiential modalities. This paper addresses the problem of how spatial-graphical representations convey abstract meanings by means of metaphors, which allow us to understand or express abstract concepts in terms of concrete expressions, particularly ordinary, relatively static, conventional devices (e.g., map legends, key pads, and clocks).

Fourceville's (1996) analysis of abstract concepts conveyed by creative images and language in advertising, aims toward a theory of 'pictorial metaphor'. Tversky (2001) treats depictions, such as maps, graphics, and icons as involving spatial metaphor derived from concrete world experience, across languages and cultures. Zacks, & Tversky (1999) argue that systematic correspondences between graph forms and interpretation are naturally derived, not due to knowledge of explicit conventions. This paper takes a similar treatment of metaphor, adopting cognitive semantic theory (Clausner, 1993, 1994; Clausner & Croft, 1997; Grady; 1997; Lakoff, 1993; Lakoff & Johnson, 1980; Lakoff & Turner, 1987), which treats metaphor as conventional schemas expressive of ordinary conventional language.

In this theory of metaphor, knowledge is organized into experientially based domains; e.g., SPACE, TIME, LIVING THINGS (see Clausner & Croft, 1999, for an overview of the theory of domains in cognitive semantics). A conventional metaphor is a stored relation between two domains. Concepts from an abstract (target) domain are systematically comprehended or expressed in terms of concepts from a different, often concrete, (source) domain. For example, MORE IS UP AND LESS IS DOWN is a conventional metaphor whose source domain UP-DOWN stands in relation to the target domain MORE-LESS. This metaphor is a semantic structure which can be instantiated as linguistic expressions; e.g., *rising prices, fell ill, high esteem, fell unconscious*.

Language expresses abstract concepts metaphorically by means of spatial and other basic perceptual concepts (Grady, 1997). Metaphors that relate spatial source domains to non-spatial target domains can be productive of linguistic expressions about non-spatial abstract meanings by using words having spatial meanings. The metaphor MORE IS UP is strongly implicated by investigations of graphs as expressions in space. Tversky, Kugelmass & Winter (1991) found that subjects assigned interpretations to the axes of graphs, such that increasing quantity was preferentially assigned to the vertical axis, and temporal concepts were preferentially assigned to the horizontal axis. Gatis & Holyoak (1996) investigated subjects' interpretation of graphs, finding a significant advantage when the variable being queried was assigned to the vertical axis. They argue that graphing increasing quantity in terms of vertical spatial increase is based on the metaphor MORE IS UP. Given that there is evidence for conventional metaphor being expressed in the construal of spatial graphs, this paper proposes the following hypothesis: The same cognitive principles which determine metaphor productivity for linguistic expressions also determine metaphor productivity for spatial expressions. This hypothesis will be tested with respect to a specific technical characterization of metaphor productivity.

Productivity in Metaphors

Clausner & Croft (1997) argue that just as phonological schemas vary in their productivity of base-derived relations, so semantic schemas (i.e., meta-

[1] The author thanks the three anonymous reviewers of this paper for their helpful comments. An earlier version of this research was presented at the Seventh International Cognitive Linguistics Conference, 2001, Santa Barbara, California.

phor source-target domains) also vary in their productivity of metaphorical expressions. Schematicity is the range of source (or target) domain concepts consistent with the schema. Productivity is the proportion of a schema's range which can be instantiated as expressions. This translates into strength of stored representations, called degrees of entrenchment. Relative entrenchment between a schema and its instantiations characterizes its productivity. High productivity of a metaphor is a configuration of a strongly entrenched schema relative to a wide range of weakly entrenched instantiations. Whereas, low productivity of a metaphor is a weakly entrenched schema relative to a narrow range of strongly entrenched instantiations.

In the following sections three degrees of metaphor productivity are considered: High productivity, Semi-productivity, and Nonproductivity. The three cases are presented in separate sections. In each case, the section begins with the conceptual principles by which Clausner and Croft (1997) account for how metaphors vary in their degree of productivity for linguistic expressions. Then evidence that spatial-graphic expressions are metaphorical expressions is presented. Measures from human subjects or semantic analysis are used to argue that in each case the pattern of results is attributable to conceptual principles of metaphor productivity.

High Productivity

The case of a metaphor having high productivity of linguistic expressions is characterized by Clausner & Croft (1997) as follows. The source-target domain relation [S → T], is a schema that produces instantiations. Each instantiation [e → c] is a metaphorical expression *e* whose source domain words are about target domain concepts *c* (Figure 1). For example the metaphor schema [UP/DOWN → MORE/LESS] (i.e., MORE IS UP AND LESS IS DOWN) can be instantiated as any number of expressions in which words about verticality (e.g., *rising/falling,…*, *up/down*) express non-spatial concepts. Metaphor productivity is a semantic configuration of a metaphor and its instantiations which includes the relation of schematicity between schema and instantiations (descending arrows). High productivity is high schema entrenchment (bold box) relative to a wide range of *i* weakly entrenched instantiations (lighter boxes).

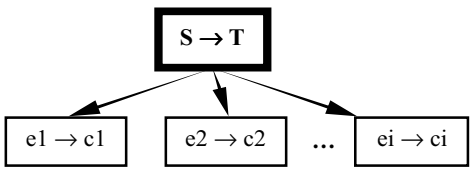


Figure 1: Semantic characterization for high productivity of a metaphor.

A highly productive metaphor expresses a large proportion of concepts consistent with the metaphor schema. Nearly any concept about quantity or quality can be expressed in terms of spatial verticality (e.g., *rising/falling prices,…, feeling up/.down*). This characterization of high metaphor productivity for linguistic expressions will be applied to the investigation of metaphor productivity for spatial-graphic expressions.

High productivity (Spatial-graphical metaphorical expressions)

The vertical spatial axis of graphs can be employed to convey quantities and this can be attributed to the metaphor MORE IS UP (e.g., Gatis & Holyoak, 1996). If this metaphor indeed has high productivity and the theory applies to both linguistic and spatial-graphical expressions, then not only numerical quantities, but also non-spatial qualities (e.g., severity) should be found instantiated in space. Specifically, a wide range of concepts about severity is predicted to be expressed such that great severity is expressed as high vertical space. This was measured using assignments of severity to a vertical map legend. The subjects were 34 graduate and undergraduate student interns at HRL Labs. Each subject was given Figure 2 with the instructions, "This is a legend to be used for a weather map of storm severity. As you can see no colors are yet assigned. Assign colors to the two extremes (two boxes) of the legend. Next, (in the blank lines) label which extreme of storm severity is which."

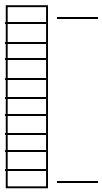


Figure 2: Vertical map legend, uncolored with blanks for labeling the two extreme values.

The aim of color assignment was to first get subjects to commit to a particular orientation of severity (target domain) to verticality (source domain), without having to verbalize target domain concepts. While color names were not expected to be sufficient evidence of any systematic conceptualization of the legend, the labeling of storm severity was expected to evoke words that would demonstrate a preferential orientation consistent with the MORE IS UP metaphor.

Table 1 summarizes the range of types of responses subjects gave as labels for the top and bottom of the legend. The left column lists response types which assign extreme storm severity to the top of the legend (e.g., *severe* and *calm* written in for the top and bottom

labels, respectively). The right column of response types are those which assign extreme storm severity to the bottom of the legend.

Table 1: Response types and percent consistent with MORE IS UP or LESS IS UP, N=34.

MORE SEVERE IS UP / LESS SEVER IS DOWN	LESS SEVERE IS UP / MORE SEVERE IS DOWN
high / low	low / high
heavy / light	lightest / heaviest
severe / calm	calm / severe
bad / good	
most / least	
severe / less severe	
very severe / not severe	
extreme / clear	
hurricane / balmy	
misery / balmy	
stormy / fair	
91%	9%

Of 34 subject responses, 31 responses assigned greatest severity to the top of the legend and 3 responses assigned the least severity to the bottom. As expected, the assignment of severity to verticality does not occur at a chance rate, but is significantly biased toward the assignment of greater severity to higher verticality, $\chi^2(1, N = 34) = 46.12$, $p < 0.001$. It can be concluded that something in the cognitive process of doing the task biases the responses, and this bias is consistent with the conventional metaphor MORE IS UP.

Color assignment is not the central task and the results were not expected to bear on the investigation; indeed they are varied. Nonetheless, it is noteworthy that of 34 total color assignment responses, the end of the legend that was also labeled as the most severe (regardless of vertical assignment), was most frequently designated "red" for 68% of color assignments. "Blue" was the most frequent color assigned to the least severe end of the legend for 38% of all color assignments, regardless of vertical assignment.

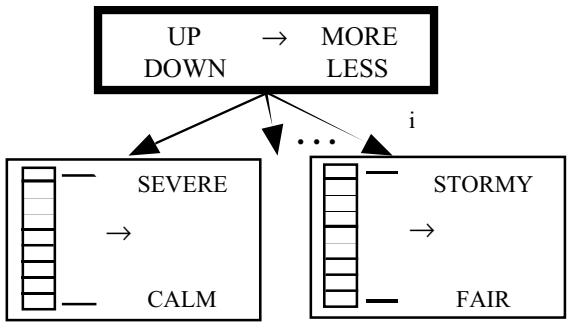


Figure 3: Graphic-spatial expressions of a highly productive metaphor.

These results suggest that understanding the vertical spatial graphic as a legend of storm severity systematically yields interpretations of the top as an expression of greatest severity, contra the bottom. That is, the subjects read the spatial verticality as an expression of concepts about the abstract quality of severity, such that a significantly large proportion of these concepts are consistent with the *up* part of UP/DOWN SPACE expressing the *more* part of the MORE/LESS quality scale. This characterization of the results is depicted in Figure 3 as a case of many concepts about storm severity (e.g., *severe/calm*, *stormy/fair*, ... and others not shown) as the meanings of the extreme ends of the legend. Each instance of the legend having a particular conceptualization is an expression of a metaphor, as depicted by the arrows between the metaphor MORE/LESS IS UP/DOWN (upper bold box) and each of the instantiations (lower lighter boxes) in Figure 3.

This is the same relation between schema and instantiation illustrated in Figure 1 for linguistic expressions of a metaphor. Just as there are many non-spatial concepts *c* expressed as linguistic expressions *e* using words about vertical space, there are also many non-spatial concepts expressed as spatial-graphs. A large range of non-spatial concepts are consistent with the target domain MORE/LESS and these are expressed by means of a vertical graphical legend consistent with the source domain UP/DOWN, respectively.

It can be concluded that since a large proportion of the schema's range can be instantiated as graphical expressions the metaphor schema is highly entrenched relative to a wide range of weakly entrenched instantiations. The metaphor is highly productive, and of more than the legend colorings and labelings investigated here, but of any graphic vertical axis (e.g., mercury thermometers). Having argued for a case of high metaphor productivity expressed in space, we will next consider cases of lesser productivity.

Semi-Productivity

This section presents an analysis of spatial expressions and argues that they systematically instantiate a semi-productive metaphor. First, the characterization of semi-productive metaphors for linguistic expressions given by Clausner & Croft (1997) is summarized. Then, the case of spatial expressions being instantiations of a semi-productive metaphor is made.

Clausner & Croft (1997) argue that semi-productivity of a metaphor is the case of relatively few linguistic expressions of a metaphor, compared with the wide range of concepts potentially consistent with that metaphor. For example, the five idioms, *spill the beans, let the cat out of the bag, loose lips, blow the whistle,* and *blow the lid off* are all about revealing a secret. Lakoff (1987) and Gibbs & O'Brien's (1990) conclude that

these idioms are consistent with the metaphor schema THE MIND IS A CONTAINER and IDEAS ARE ENTITIES. The idiom expressions are transparent idioms, because most people have some awareness of a relationship between specific word meanings and the idiom meaning. They know that the words for physical things (e.g., cat) escaping a container (e.g., bag) are related to an idiomatic meaning of ideas coming out of the mind. Clausner & Croft's (1997) analysis of these transparent idioms argues that the metaphor schema [ENTITIES OUT OF A CONTAINER → IDEAS OUT OF THE MIND] is semi-productive. There are only five instantiations, each a transparent idiom (e.g., cat out of the bag → secret revealed). The metaphor is only partly productive because the transparent idioms express only a limited proportion of possible instantiations that are consistent with the metaphor. For example, **spill the peas* and **let the weasel out of the cage* are expressions which are consistent with the metaphor schema, but nonetheless do not mean, reveal a secret[2]. The five idioms that do have this meaning are highly conventionalized (i.e., highly entrenched) expressions.

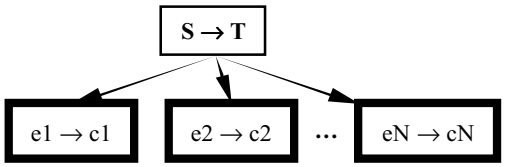


Figure 4: Semantic characterization for semi- productivity of a metaphor.

They conclude that a countable number *N* of highly entrenched instantiations of a relatively less entrenched metaphor schema is characteristic of metaphor semi-productivity. Figure 4 illustrates the cognitive structure of a lightly entrenched metaphor (lighter box) that is consistent with a limited range of highly entrenched instantiations (bold boxes). Given this account of semi-productive metaphors, semi-productivity for graphical spatial expressions is investigated.

Semi-productivity (Spatial-graphical transparent idioms)

The data considered are the spatial-graphic arrangements of digits for entering numerical values, such as those found on telephones, calculators, alphanumeric keyboards, and rotary telephones (left side of Figures 5-8, respectively). Each of these four digit arrangements is highly conventionalized. It is transparent, however, that the digits are arranged in counting order. That is, any specific spatial arrangement of digits is understood to be a meaningful ordering of incrementally successive values. That the spatial arrangements vary widely, but have the same interpretation as ordered values, suggests a systematic relation between their spatial expression and their conceptual meaning.

	*
1 2 3	0
4 5 6	1 2 3
7 8 9	4 5 6
0	7 8 9

Figure 5: Phone, television, and bank ATM digit pad.

	*
7 8 9	0
4 5 6	7 8 9
1 2 3	4 5 6
0	1 2 3

Figure 6: Calculator key digit pad.

1234567890

* 0123456789

Figure 7: Typing keyboard number order.

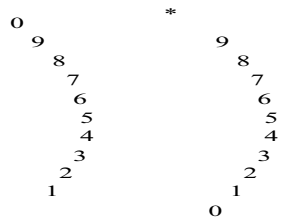


Figure 8: Rotary Telephone digit order.

The investigation of a common semantic relation that is consistent among these four spatial arrangements can proceed as one would a linguistic case. Analyzing the range of spatial expressions characterizes the schema that might be instantiating them, just as it does for linguistic expressions. Among these four digit arrangements, the "0" numeral occurs in two positions, either near the "9" numeral, or near the "1" numeral. This suggests that the shared meaning of the digit arrangements is the concept of counting, expressed with the "0" numeral representing either the 0th or 10th value. In fact, the first form of the rotary phone had numbered finger holes, except the "0" was marked with the Roman numeral "X" (Hill, 1953).

If the counting order of digits is indeed the common meaning among the four spatial arrangements, then that order is independent of whether the spatial configuration of the digits is rotational, horizontal, vertical, or

[2] An asterisk, "*", which begins a linguistic expression or graphic expression denotes its infelicity as a conventional expression of the particular semantic meaning in discussion.

some combination (e.g., reading "1" to "9" successively left-to-right and top-to-bottom, as on a phone pad). These configurations (or combinations, of them) are expressed, but it is the counting order, not the spatial order, that is common across the expressions. The only variation among digit orderings is the relative position of "0" in the counting order.

In laying out a metaphorical system of how humans understand abstract mathematical concepts Lakoff & Núñez (2000) propose metaphors which fit the above analysis. They first establish that MODULAR ARITHMETIC IS ALGEBRAIC GROUPS and GEOMETRIC ROTATIONS ARE ALGEBRAIC GROUPS. That is, the digits for counting in base 10 modular math are conceptualized as rotations. Successive rotations form the basis for counting, which they argue is the metaphor THE INFINITE CLASS OF NUMERALS FOR THE NATURAL NUMBERS IS AN ITERATIVE PROCESS THAT GOES ON AND ON. For the purpose of this paper, the simpler characterization is [CYCLIC PROCESS → COUNTING WITH DIGITS]. The spatial expression of the CYCLE in this case is the spatial traversal of numerals in one of four conventional arrangements. The traversal is cyclic by means of the "0" indicating modulo 10.

Evidence that the metaphor does not express the full range of spatial expressions consistent with its specifications comes from the absence of specific digit arrangements. The right side of Figures 5-8, depict spatial arrangements of digits that are unexpressed as conventional digit orders (as denoted by "*"). These spatial arrangements are consistent with base 10 counting order, but they are not conventionalized.

Figure 9 depicts the metaphor (upper light box) which is semi-productive of *N* highly conventionalized instantiations (lower bold boxes). The *N*=4 spatial arrangements of numerals are understood as ordered digits. The metaphor is semi-productive, because the four instantiations are consistent with the metaphor, which is not productive of other expressions.

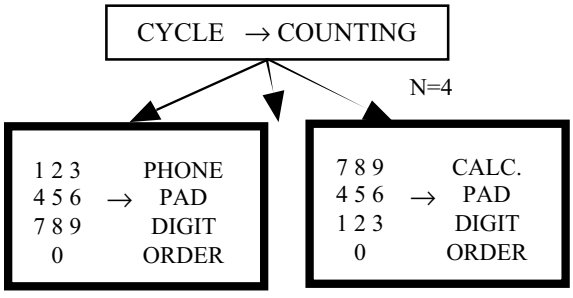


Figure 9: Graphic-spatial expressions of a semi- productive metaphor.

The case of spatial expressions of semi-productive metaphor can be called, "transparent spatial-graphic idioms". Transparent linguistic idioms are highly entrenched expressions of a comparatively lightly entrenched metaphor (Figure 4). This same principle of semi-productivity applies to spatial expressions. Only four of the possible spatial arrangements consistent with the metaphor domain relation are expressed. The expressed spatial arrangements are highly conventional and their meaning is semantically transparent.

Nonproductivity

This section investigates productivity less in degree than semi-productive metaphor, specifically the case of nonproductivity. Opaque linguistic idioms are very few. For example, *kick the bucket*, meaning to die, and *by and large*, meaning something in general. They are semantically opaque in that most English speakers know these linguistic expressions and know what they mean, but do not know why they make sense. Clausner & Croft (1997) analyze *save face* and *lose face* as opaque linguistic expressions of an absent metaphor. The expressions were borrowed from Mandarin Chinese. English speakers know their meaning is about avoiding disgrace and incurring disgrace, but they do not know why they mean this. English lacks the relevant metaphor [HAVING FACE → SOCIAL RESPECT] that would be productive of other expressions, such as those not borrowed into English, e.g., **give face*, does not mean, to show due respect for someone's feelings, as it does in Mandarin.

Figure 10 depicts the absence of an entrenched metaphor schema (light dashed box). The expression is extremely conventional, that is, highly entrenched (bold box), and it is opaque since there is no source-target domain relation that would be productive of the expression *e* in relation to concept *c*. Also absent are vertical arrows, indicative of the nonproductivity.

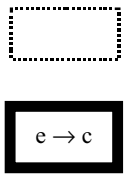


Figure 10: Semantic characterization for metaphor nonproductivity.

Now we turn to graphical-spatial expressions and argue that we find the same semantic structure as for opaque linguistic idioms.

Nonproductivity (Spatial-graphical opaque idioms)

An extremely conventional spatial-graphical expression is the analog clock. There are clockwise clocks, but not counter-clockwise clocks (although they exist for

amusement, but not conventional use). Most people do not know why clocks run in the conventional clockwise direction. It is opaque, in that there is no conventional semantic relation between the source and target which makes the spatial pattern make sense. That is, there is no widespread knowledge that the sidereal rotational direction of shadows cast by sundials in the northern hemisphere is the basis for the rotational direction of clocks. There is no conventional entrenched metaphor schema. If there were a metaphor it might be something like, PASSAGE OF TIME IS SHADOW ROTATION.

Just as there are few opaque linguistic idioms, it should follow that there are few "opaque spatial idioms", precisely because there is no conventional metaphor to produce them. Another example of an opaque spatial idiom would be the QWERTY keyboard[3]. The arrangement of alphanumeric keys is widely known, but why they have this specific arrangement, among all the possible unattested ones, is largely opaque. It is not widely known that early mechanical typewriters arranged keys in order to slow typing speed, thus reducing the likelihood of two keys striking together.

Discussion and Conclusion

These results point to issues of representation and conceptual structure common to linguistic semantics (spoken and signed), psychology, computational modeling and the role of metaphor in human-computer interaction. Further work is required to distinguish the theory of metaphor productivity from alternate interpretations of the results, e.g., treating conventional correlations between the top of a spatial axis and one pole of a semantic scale as due to salience. The top of a legend or vertical mercury thermometer, or keypad digit ordering may be interactions of perception and conventionality; however, conventionalized form-meaning pairs do not obviate a conventional metaphor schema. The theory of productivity is about contemporary conceptual structure (not historical origins of the metaphors) and predicts the above results.

Spatial-graphical expressions are argued to be instantiations of conceptual metaphors which vary in productivity according to the same principles which determine productivity for linguistic expressions. The relative entrenchment of a metaphor to its instantiations is argued to result in varying ranges of expressions. Three degrees of metaphor productivity were investigated. In each case the principles which are held to explain ranges of linguistic expressions are argued to explain evidence about spatial graphical expressions. These are, in decreasing order: High productivity is the case of spatial-graphical metaphorical expressions. Semi-productivity is the case of spatial-graphical transparent idioms. Nonproductivity is the case of spatial-graphical opaque idioms. These cases are concluded to represent three points on a continuum of metaphor productivity for spatial-graphical expressions.

[3] The author thanks Sarah Taub for suggesting this example.

References

Clausner, T.C. (1993). *Cognitive structures in metaphor comprehension.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Clausner, T.C. (1994). Commonsense knowledge and conceptual structure in container metaphors. *Proceedings of Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Clausner, T.C., & Croft, W. (1997). Productivity and schematicity in metaphors. *Cognitive Science, 21*, 247-282.

Clausner, T.C. & Croft, W. (1999). Domains and Image Schemas. *Cognitive Linguistics, 10*, 1-31, New York: Mouton de Gruyter.

Forceville, C. (1996). *Pictorial Metaphor in Advertising*. London: Routledge.

Gatis, M. & Holyoak, K.J. (1996). Mapping conceptual to spatial relations in visual reasoning. *J. Experimental Psychology: Learning, Memory, and Cognition, 22*, 231-239.

Gibbs, R.W. & O'Brien, J. (1990). Idioms and mental imagery: The metaphorical motivation of idiomatic meaning. *Cognition, 36*, 35-68.

Grady, J. (1997). *Foundations of Meaning: Primary Metaphors and Primary Scenes*. Unpublished doctoral dissertation, University of California Berkeley.

Hill, R.B. (1953). The Early Years of the Strowger System. Bell Laboratories Record. Volume XXXI No. 3, March, 1953.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed.). Cambridge: Cambridge University Press.

Lakoff, G. & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science, 4*, 195-208.

Lakoff, G. & Núñez, R.E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Lakoff, G. & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor.* Chicago: University of Chicago Press.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology, 23*, 515-557.

Tversky, B (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas in abstract thought*. Cambridge, Mass.: MIT Press.

Zacks, J. & Tversky, B. (1999). Bars and lines: a study of graphic communication. *Memory and Cognition, 27*, 1073-1079.

What makes a word?

Eliana Colunga (ecolunga@cs.indiana.edu)
Department of Psychology; 1101 East 10th Street
Bloomington, IN 47405-7007 USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychology; 1101 East 10th Street
Bloomington, IN 47405-7007 USA

Abstract

Words seem to have a special status among perceptual signals. The developmental evidence, however, suggests that words *become* special. Woodward and Hoyne (1999) showed that 13-month-olds readily associate both words coming from the experimenter's mouth and non-linguistic sounds coming from a hand-held noisemaker, with object categories. In contrast, 20-month-olds associate words but not non-linguistic sounds with object categories. Woodward and Hoyne suggest that words become privileged as possible names; that the forms a name can take are open at the beginning and become more restricted with development. Are children learning what forms count as words? If so, just what defining features are they learning? This paper presents an associationist account of this developmental trend and tests this explanation in two experiments with 20-26-month-old children.

Introduction

Words seem to have a special status among perceptual signals. Having a label for an object changes the way it is categorized for both adults and children. For example, when asked to generalize an object name to new instances, children and adults generalize by shape. However, when asked to find an object that "goes with" another, they choose by overall similarity (Landau, Smith & Jones, 1988; Imai & Gentner, 1997). A label also makes children's choices shift from thematic to taxonomic (Waxman, 1997) and from surface to more conceptual similarities (Keil, 1989) . As Waxman said, words work like invitations to form categories; words are category names. But what makes a word? How do children know whether a particular sound is a category label?

One finding critical to this issues was reported by Woodward and Hoyne (1999) . They presented children with two novel objects and labeled one of them (the target object). In the Word condition they paired the target object with a word ("this is a toma"); in the Sound condition they paired the target object with a non-linguistic sound, such as a tone. Children were then asked to "get the toma" or "get the $<$ *tone* $>$" to test whether they had associated the "label" (toma or $<$ *tone* $>$) with the object. They asked: Do children treat only words as possible names or do they also accept tones as possible names? Their results indicate that the answer to this question depends on the developmental level of the child. Thirteen month-old infants will associate both a word and a non-linguistic sound with a target object. In contrast, 20-month-old children will associate a word to a target object, but not a non-linguistic sound. Namy & Waxman (1998) have similar results for 18- and 26-month-olds contrasting words and gestures. While the younger children will associate both a novel word and a novel gesture with a target object (object category) the older children will only associate the word to the object, and not the gesture.

Both teams of researchers suggest that older children do not associate non-words with the objects because older children know that non-words are not possible names. The idea is that words become privileged as possible names; that the forms a name can take are open at the beginning and become more restricted with development. But how do words become names and thus privileged? What determines what counts as a name?

In this paper we attempt to answer these questions. First we offer a mechanistic explanation of this developmental trend. Then we present two experiments that test our explanation.

An associationist account

We propose that words become privileged as category names because of the special way in which they correlate with object categories. In the experience of a child, many events may co-occur with attention to objects. For example, objects may co-occur with expressions such as "look!", gestures such as pointing, words related or unrelated to the object, noises, actions related or unrelated to the nature of the object, and so on. However, of all these events, words (as object names) correlate in a way that makes them especially good predictors for category membership.

By our account, there are two properties that make words good candidates for becoming privileged as names. The first property is predictiveness or cue validity. There is one name (more or less) that goes with one category (more or less). Thus, the name of

a category is a feature that all members of the category have in common, while at the same time the name is a feature that distinguishes instances of the named category from members of other categories. This is illustrated in Figure 1. The word "ball" typically co-occurs with members of the category BALL, but not with members of the category DOG. Similarly, the word "dog" co-occurs with members of the category DOG, but not with members of the category BALL. In contrast, events like pointing and hearing "look!" will just as likely co-occur with both balls and dogs. Thus, it is the object names that are predictive of object category, and not events like pointing or the word "look!".

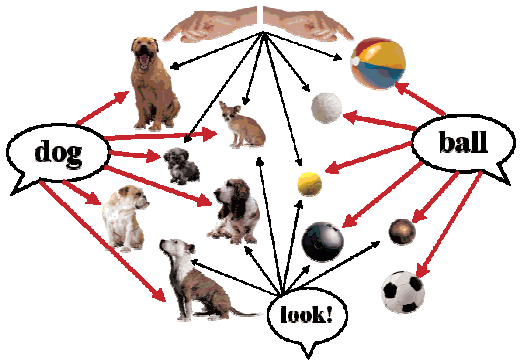


Figure 1: Object names systematically correlate with object categories

Predictiveness, however, is not enough by itself. After all, even 3-4 month-old infants can distinguish pictures of dogs from, for example, cars. Whether infants have the category or concept of DOG or CAR or not, the fact remains that there is something about dogs and cars perceptually that allows them to distinguish members of the two categories. Why isn't the feature "dogness"(or "carness") which is predictive of category membership for the category DOG (or CAR) not something that can be taken as a name, like a word? We propose that the answer to this question lies in a second statistical property: systematicity. That is, words *as a domain* are predictive of object category membership. Put another way, if there were just one word that correlated with a category, words in general would not get an advantage. The fact that there are many words that co-occur with many object categories is what helps children generalize this expectation to novel words.

The mechanism we propose also says something about the nature of words as names. According to our account, a name is simply the bundle of signals that systematically co-occurs with categories. These could be properties such as being a speech sound with particular spectral and prosodic forms, being produced by people, coming out of mouths, or co-occurring with pointing and eye gaze to the object. Thus, these may be the properties that, through language learning, come to define what counts as a name for children.

In sum, in our account what makes words privileged as names is that they co-occur systematically with object categories. Conversely, a name is whatever features systematically co-occur with object categories, even beyond what we usually think of as words (or names).

Thus, we make the following two predictions:

1. Events that co-occur systematically with object categories come to refer — to be usable as names. According to our account, any event domain that systematically predicts category membership will be taken as a name as well. Fortunately for the experiment we report here, children's experiences include a domain in which something other than words co-occurs systematically with categories — the domain of animals. Animal category correlates with animal sound: dogs bark, cats meow, elephants trumpet and so on. Thus our first prediction is that animal sounds should be taken as names for animals.

2. What defines a name is the cluster of features that systematically co-occurs with categories. This means that any strongly correlated feature of a name, even beyond what we think of as a word, will become an integral part of what is a name. For young children who rely on spoken language, words emanating from mouths is a highly systematic property of names. Therefore children should take coming-from-a-mouth as one of the defining features of being a name. Thus, our second prediction is that if a word comes out of a place other than a mouth, young children will not take it as a name. Conversely, young children may take a non-word as a name if it emerges from a mouth.

In the next experiment we tested these predictions in the domain of animals. We selected children to participate who by Woodward & Hoyne's and Namy and Waxman's studies should already treat words as the only privileged naming events. To test the first prediction (that animal sounds can be used as names for animals) we labeled animal toys with different kinds of sounds: a word, an animal sound, and a motor sound. To test the second prediction (that emanating from the mouth was a defining feature of being a name) we made the names emanate from different sources: from the experimenter's mouth or from a nearby object.

The design for Experiment 1 is shown in Figure 2. Note that from Woodward and Hoyne's study we know what will happen in the Word-Mouth cell and in the ArbitrarySound-Noisemaker cell. Children should take the word as a name in the first case and reject the sound as a name in the second case. The questions are: will they accept the animal sound as a

name? Will they accept any kind of sound emanating from the mouth as a name? Will they accept the word as a name regardless of where it comes from, or will the source matter?

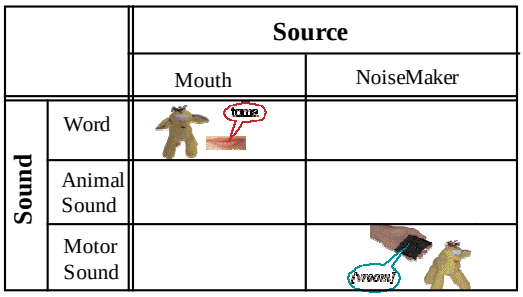


Figure 2: The experiment had three different kinds of sound (Word, Animal Sound, Arbitrary Sound) as within-subject conditions and two different sources (Mouth, Noisemaker) as between-subject conditions

Experiment 1

Methods

Subjects. 24 20-26 month-old children participated in the experiment.

Design. We used a 2x3 mixed design with the two different sources (mouth, noisemaker) as a between-subject variable and the three different sounds (word, animal sound, arbitrary sound) as a within-subject variable.

Stimuli. The stimuli consisted of two sets of six novel toy animals. The animals in the two sets were the same in all respects except in color, and one was used as the generalization of the other. The sounds used as names were the word "toma", a frog sound as the animal sound, and a motor sound as the arbitrary sound.

Procedure. The experiment was preceded by a training phase. The goal of the training phase was to make sure that the child understood the task and could make clear choices. In this phase we presented the child with a familiar object (a ball, a spoon, a flower) and asked the child to "get the ball" (or spoon or flower). Once the child had done this, we put two familiar objects on the tray and asked the child to get one of them. The training was considered successful if the child retrieved the correct object twice from the tray with a distracter.

Each child heard three different kinds of names (Word, Animal Sound, Motor Sound) in three blocks. Each block consisted of a Familiarization phase and a Test phase.

In the Familiarization phase the child was shown two different toy animals and a name was supplied for one of them – the target object. The two objects were presented twice, one toy animal at a time. First the target animal was presented and named, and then the distracter animal was presented with the same phrases but without a name. Then the target animal was presented and named again followed by the distracter animal.

In the Mouth condition, children heard the three kinds of label coming from the experimenter's mouth. When presenting the target object, the experimenter named it saying, "Look at this toma. Wow! See this toma? Look! Toma." in the Word condition, imitated the animal sound in the Animal Sound condition ("look at this $< frog - likeclucking >$") and imitated the mechanical sound in the Motor Sound condition ("look at this $< motor - likesound >$"). In the Noisemaker condition the three kinds of sounds came from cloth-covered recorders that were held close to the toy animal being named. The distracter objects were always presented with the same phrases as the target objects, but without the name: "Look at this! Wow! See this? Look!"

In the Test phase the child was presented with two choices on a tray and asked to retrieve the target object. The test question was asked in the same manner as the naming in each condition – from the mouth in the Mouth condition and from the Noisemaker in the Noisemaker condition. There were four test trials for each kind of sound; two test trials used the same animals as the ones used in the Familiarization phase and the other two were generalization trials, using the animals that matched the familiar ones in all aspects except for their color. Each child got a total of 12 trials. The toys were randomly assigned to each condition for each child. The order of the two sound type conditions was counterbalanced.

Results

We coded children's choices as the first object they touched or took from the tray. Figures 3 and 4 show the number of children who successfully mapped the name to the object category in the Word, Animal Sound and Motor Sound conditions. We classified children as Successfully Mapping if they picked the target object when asked on three or more of the four trials.

Figure 3 shows children's performance. In the Word condition, most children (9 out of 12) successfully mapped the word to the animal category when the word came from the experimenter's mouth. The number of children that successfully mapped the word to the animal category in the Mouth condition was reliably more than would be expected by chance ($p < .01$). In contrast, the number of Successfully Mapping children in the Noisemaker condition was almost reliably below what would be expected by

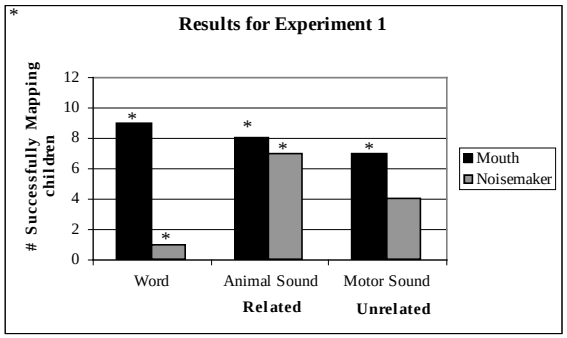


Figure 3: Results of Experiment 1. Any sound emanating from the mouth is taken as a name and animal sounds are taken as names regardless of its source

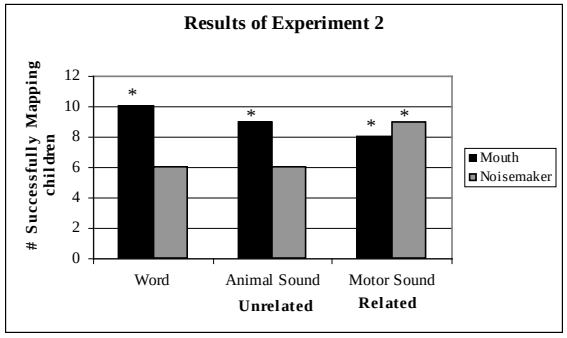


Figure 4: Results of Experiment 2. Any sound emanating from the mouth is taken as a name and motor sounds are taken as names regardless of its source

chance ($p = .06$). In fact, only one child consistently retrieved the target object when the name was a word coming from a handheld noisemaker. Thus, it appears that children *only* accept the word as the name of an animal category when the word emanates from the mouth of the experimenter.

In the Animal Sound condition, the number of the Successfully Mapping children exceeded what would be expected by chance in both source conditions (Mouth: $p = .01$, Noisemaker: $p = .03$). That is, children successfully mapped the animal sound to the animal category regardless of the source of the sound. They did so whether the sound came from the experimenter's mouth or from a handheld noisemaker. So, children always accept the animal sound as the name of an animal category, regardless of the source from which the animal sound emanates.

In the Motor Sound condition – when the sound used was an unrelated sound – the number of children successfully mapping the sound to the animal category only exceeded what would be expected by chance in the Mouth condition ($p < .05$). The number of children that successfully mapped the mechanical sound with the animal category when the sound came from the Noisemaker did not reliably exceed chance ($p > .2$). That is, motor sounds are only accepted as names when they emanate from a mouth.

In short, any sound emanating from the mouth is taken as a name and animal sounds are taken as names regardless of its source.

Discussion

Our results replicate Woodard and Hoyne's study: words emanating from mouths are associated with object categories and arbitrary sounds emanating from handheld noisemakers are not associated with object categories. In addition to that, however, we have shown two things. First, that words are only accepted as names when they come from the speaker's mouth, and not when they come from other sources, such as a hand-held noisemaker. Second, that even non-linguistic sounds, such as the buzz of a motor or the croak of a frog, will be taken as a name if they are produced by a human mouth. Therefore, our results suggest that for children at this age it is not words that are taken as the privileged form of naming, but rather sounds produced by a human mouth, that is, source matters.

One difference in our results between words and animal sounds is that the source of the name matters for the word, but not for the animal sound. This also fits with our associationist account: in children's experience, animal sounds are not specific to a source. They emanate from the mouths of real animals, from the inside of stuffed animal toys, and from mouths of people imitating animals. In contrast, words – as object names – are typically produced by human mouths. Therefore the source will be part of what defines a word as a name, but not of what defines an animal sound as a name for an animal category.

Why are sounds emanating from mouths always taken as names? According to our account, this is because emanating from a mouth is one of the most systematically correlating features of naming situations. Another possibility that needs to be explored is that perhaps when produced by a human mouth even an imitation of a mechanical sound stops being arbitrary. Being made by a mouth may make any sound word-like (or animal sound-like).

More importantly for our proposal, however, our results showed that animal sounds – which are systematically correlated with animal categories in the real world – will be accepted as labels for animal categories regardless of their source. Thus, our predictions were confirmed.

Why are animal sounds taken as names for animal categories? According to our proposal, this is because animal sounds correlate with animal categories in much the same way as words correlate with object categories in general: one animal sound corresponds to one animal category, and animals typically

are associated with a sound they make. However, an alternative explanation is that there is something in the acoustic features of the animal sound used that makes it word-like. That is, it may be that it is not the special way in which animal sounds correlate with animal categories which makes animal sounds good potential labels *for animal categories*, but that there is something about animal sounds (perhaps they are closer to linguistic sounds in some similarity space), that makes them good potential labels *for any category*. Conversely, it could be that motor sounds are just not word-like enough to serve as a label. Thus, to support our propsal we have to show that animal sounds are good *only* for animal categories (and not vehicle categories), and that motor sounds are good *only* for vehicle categories (and not animal categories). Accordingly, in the next experiment we test this alternative explanation by replicating Experiment 1 using toy vehicles instead of toy animals as stimuli.

We reasoned that vehicle sounds correlate with vehicle categories much in the same way as animal sounds correlate with animal categories. Thus, if our account is right, and it is the systematicity of correlations that makes a word, we predict that contrary to what was found in Experiment 1, the motor sound (now related) will be accepted as a name for toy vehicles, but the animal sound (now unrelated) will not. However, if it is something specific about the animal sound we used that made it work as a name, the same animal sound should be accepted as a name for vehicle categories as well.

Experiment 2

Subjects. 24 20-26 month-old children participated in the experiment.

Design. As in Experiment 1, we used a 2x3 mixed design with the two different sources (mouth, noisemaker) as a between-subject variable and the three different sounds (word, animal sound, motor sound) as a within-subject variable.

Stimuli. The stimuli consisted of two sets of six novel toy vehicles. We used the same sounds as in Experiment 1, that is the word "toma", a frog sound, and a motor sound.

Procedure. The procedure is the same as in Experiment 1.

Results

Children's choices were coded as in Experiment 1. Children were classified as Successfully Mapping according to the same criterion as in Experiment 1 – when they chose the target object correctly in at least 3 of the 4 trials. The results of Experiment 2 are analogous to the results of Experiment 1. Children map the motor sound to vehicle categories, but fail to map the animal sound – the unrelated sound.

Figure 4 shows children's performance in Experiment 2. In the Word condition, most children (10 out of 12) successfully mapped the word to the animal category when the word came from the experimenter's mouth. The number of children that successfully mapped the word to the animal category in the Mouth condition was reliably more than would be expected by chance ($p < .01$). In contrast, the number of Successfully Mapping children in the Noisemaker condition was not different from chance ($p > .2$). Thus, as in Experiment 1, children accept the word as the name of a vehicle category only when the word emanates from the mouth of the experimenter.

In the Animal Sound condition, when the sound used as name was a unrelated to the vehicle categories, the number of children successfully mapping the sound to the animal category only exceeded what would be expected by chance in the Mouth condition ($p < .05$). The number of children that successfully mapped the animal sound with the vehicle category when the sound came from the Noisemaker did not reliably exceed chance ($p > .2$). That is, animal sounds are only accepted as names for vehicle categories when they emanate from a mouth.

In the Motor Sound condition, when the sound was systematically related to vehicle categories, the number of the Successfully Mapping children exceeded what would be expected by chance in both conditions (Mouth: $p = .01$, Noisemaker: $p = .03$). That is, children successfully associated the motor sound to the vehicle category regardless of the source of the sound. So, children seem to accept the motor sound as the name of an vehicle category, regardless of the source from which the motor sound emanates.

In short, as in Experiment 1, any sound emanating from the mouth is taken as a name and related sounds (in this case motor sounds) are taken as names regardless of their source.

Discussion

In Experiment 1, the pattern of results could be explained away by suggesting that there was something special about the animal sound used in the experiment, or animal sounds in general, that made it more word-like. By using the same sounds as in Experiment 1, but showing the opposite pattern of results (Animal sound not taken as a label; Motor sound taken as a label), we showed this is not the case. Which non-linguistic sound will be more readily associated with an object category depends on the kind of object categories being associated: if the categories are from the domain of animals, then the animal sound will have the advantage; if the categories are from the domain of vehicles, then the motor sound will have the advantage.

Furthermore, Experiment 2 provided converging evidence from a different domain for the idea that systematically correlating cues become good candidates for label-hood. The results of this experiment agree with the results of Experiment 1. That is, children in Experiment 2, like children in Experiment 1, were likely to map the related non-linguistic sound to the object categories, but not the unrelated non-linguistic sound.

Conclusions

The results of the two experiments showed the same pattern: Words, as well as non-linguistic sounds that systematically correlate with the relevant domain of categories, are accepted as labels. In contrast, non-linguistic sounds that are unrelated to the domain in question are not accepted as labels. Furthermore, words are only accepted as labels when they are produced by a mouth. Why this pattern? We believe that this pattern reflects the systematicity with which events correlate with categories in the world. Sounds from mouths typically name things, so they are taken as names even when they have unusual properties such as the imitation of a mechanical sound does. Animal sounds systematically correlate with animal categories, so these kinds of sound – from mouths or from noisemakers – are accepted as names for animal categories. Analogously, motor sounds systematically correlate with vehicle categories, so these kinds of sound – regardless of their source – are accepted as names for vehicle categories.

Perhaps, before language learning, there is nothing special about words as names and there is nothing special about reference. All that a word is is a bundle of highly correlating features. All that reference is, is the association between a name –the bundle of highly correlating features – and a category. Maybe children learn what is reference as they learn names, and they learn names as they experience words referring to object categories.

With more learning, what counts as a name should get more and more abstract, to the point in which emanating from a mouth may no longer be a crucial feature. However, this may be where it starts; in the systematicity with which events, such as spoken words or animal sounds, refer to categories.

References

Imai, M. and Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, 62:169–200.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press, Cambridge, MA.

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3:299–321.

Namy, L. L. and Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. *Child Development*, 69:295–308.

Woodward, A. L. and Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, 70:65–72.

Sequential Learning by Touch, Vision, and Audition

Christopher M. Conway (cmc82@cornell.edu)
Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology
Cornell University
Ithaca, NY 14853, USA

Abstract

We investigated the extent to which touch, vision, and audition are similar in the ways they mediate the processing of statistical regularities within sequential input. While previous research has examined statistical/sequential learning in the visual and auditory domains, few researchers have conducted rigorous comparisons across sensory modalities; in particular, the sense of touch has been virtually ignored in such research. Our data reveal commonalities between the ways in which these three modalities afford the learning of sequential information. However, the data also suggest that in terms of sequential learning, audition is superior to the other two senses. We discuss these findings in terms of whether statistical/sequential learning is likely to consist of a single, unitary mechanism or multiple, modality-constrained ones.

Introduction

The acquisition of statistical/sequential information from the environment appears to be involved in many learning situations, ranging from speech segmentation (Saffran, Newport, & Aslin, 1996), to learning orthographic regularities of written words (Pacton, Perruchet, Fayol, & Cleeremans, 2001) to processing visual scenes (Fiser & Aslin, 2002). However, previous research, focusing exclusively on visual and auditory domains, has failed to investigate whether such learning can occur via touch. Perhaps more importantly, few studies have attempted directly to compare sequential learning as it occurs across the various sensory modalities.

There are important reasons to pursue such avenues of study. First, a common assumption is that statistical/sequential learning is a broad, domain-general ability (e.g., Kirkham, Slemmer, & Johnson, 2002). But in order to adequately assess this hypothesis, systematic experimentation across the modalities is necessary. If differences exist between sequential learning in the various senses, it may reflect the operation of multiple mechanisms, rather than a single process. Second, in regards to the touch modality in particular, prior research has generally focused on low-level perception; discovering that the sense of touch can accommodate complex sequential learning may have important implications for tactile communication systems.

This paper describes three experiments conducted with the aim to assess sequential learning in three sensory modalities: touch, vision, and audition. Experiment 1 provides the first direct evidence for a fairly complex tactile sequential learning capability. Experiment 2 provides a visual analogue of Experiment 1 and suggests commonalities between visual and tactile sequential learning. Finally, Experiment 3 assesses the auditory domain, revealing an auditory advantage for sequential processing. We conclude by discussing these results in relation to basic issues of cognitive and neural organization—namely, to what extent sequential learning consists of a single or multiple mechanisms.

Sequential Learning

We define sequential learning as an ability to encode and represent the order of discrete elements occurring in a sequence (Conway & Christiansen, 2001). Importantly, we consider a crucial aspect of sequential learning to be the acquisition of statistical regularities occurring among sequence elements. Artificial grammar learning (AGL; Reber, 1967) is a widely used paradigm for studying such sequential learning[1]. AGL experiments typically use finite-state grammars to generate the stimuli; in such grammars, a transition from one state to the next produces an element of the sequence. For example, in the grammar of Figure 1, the path begins at the left-most node, labeled *S1*. The next transition can lead to either *S2* or *S3*. Every time a number is encountered in the transition between states, it is added as the next element of the sequence, producing a sequence corresponding to the rules of the grammar. For example, by passing through the nodes *S1*, *S2*, *S2*, *S4*, *S3*, *S5*, the "legal" sequence 4-1-3-5-2 is generated.

During a training phase, participants typically are exposed to a subset of legal sequences—often under the guise of a "memory experiment" or some other such task—with the intent that they will incidentally encode structural aspects of the stimuli. Next, they are tested on whether they can classify novel sequences as

[1]In the typical AGL task, the stimulus elements are presented simultaneously (e.g., letter strings)—rather than sequentially (i.e., one element at a time). We consider even the former case to be a sequential learning task because scanning strings of letters generally occurs in a left-to-right, sequential manner. However, our aim here is to create a truly sequential learning environment using temporally-distributed input.

incorporating the same regularities they had observed in the training input. Participants commonly achieve levels of correct classification that are significantly greater than chance. Although there has been disagreement as to what types of information participants use to make correct classification judgments, it is likely that statistical information is an essential piece of the puzzle (e.g., Redington & Chater, 1996). Participants appear to become sensitive to the statistical regularities in the training items—i.e., the frequency with which certain "chunks" of information co-occur—allowing them to generalize their knowledge to novel sequences. It is such statistical sensitivity that we consider to be vital for complex sequential learning tasks.

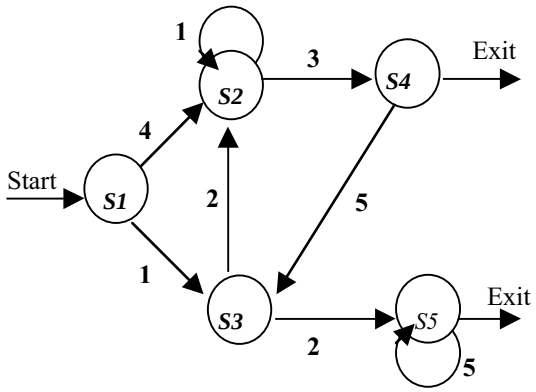


Figure 1: The finite-state grammar used to generate the stimuli for the three experiments.

The standard AGL paradigm has been used extensively to assess visual, as well as auditory learning, suggesting that sequential learning can occur in both modalities. However, two issues remain unexplored: can sequential learning occur in other modalities, such as touch? And, what differences in sequential learning, if any, exist between different sensory modalities?

Experiment 1: Tactile Sequential Learning

The touch sense has been studied extensively in terms of its perceptual and psychophysical attributes (for a review, see Craig & Rollman, 1999), yet only a few studies have hinted that complex sequential learning is possible. For instance, evidence suggests that tactile temporal processing and pattern learning is better than visual, but worse than auditory processing (e.g., Handel & Buffardi, 1969; Manning, Pasquali, & Smith, 1975; Sherrick & Cholewiak, 1986). These studies suggest that touch supports a powerful learning mechanism, which perhaps may be sufficient to allow for successful performance on an AGL task. Experiment 1 attempted to verify this hypothesis.

Method

Participants A total of 20 undergraduates (10 in each condition) from introductory Psychology classes at Southern Illinois University, Carbondale, participated in the experiment. Subjects earned course credit for their participation. The data from an additional five participants were excluded for the following reasons: prior participation in AGL tasks in our laboratory (n=4); did not adequately follow the instructions (n=1).

Apparatus The experiment was conducted using the PsyScope presentation software (Cohen, MacWhinney, Flatt, & Provost, 1993) run on an Apple G3 PowerPC computer. Participants made their responses using an input/output button box (New Micros, Inc., Dallas, TX). Five small motors, normally used in hand-held paging devices, generated the vibrotactile pulses. Each of these motors was less than 18 mm long and 5 mm wide, making them small enough to be easily attached to the participants' fingers with velcro straps. When activated, the motors produced minor vibrations (rated at 150 Hz) at a magnitude equal to that found in hand-held pagers.

The motors were controlled by output signals originating from the New Micros button box. These control signals were in turn determined by the PsyScope program, allowing precise control over the timing and duration of each vibration stimulus.

Materials The stimuli used for Experiment 1 were taken from Gomez and Gerken's (1999) Experiment 2. This grammar (see Figure 1) can generate up to 23 sequences between 3 and 6 elements in length. The grammar generates sequences of elements (numbers) with each number being mapped onto a particular finger (1 is the thumb and 5 is the pinky finger). Each tactile stimulus consisted of a sequence of vibration pulses (pulse duration of 250 ms) delivered to the fingers, one finger at a time (250 ms occurring between pulses). For example, the legal sequence 1-2-5-5 corresponds to one vibration pulse delivered to the thumb, then a pulse to the second finger, and lastly two pulses to the fifth finger.

A total of 12 legal sequences, arranged into pairs, were used for training. Six pairs consisted of one training sequence presented twice (matched pairs) whereas the remaining six pairs consisted of two sequences that differed slightly from one another (mismatched pairs). A 2 s pause occurred between the two sequences of each pair.[2]

The test set consisted of ten legal and ten illegal sequences, all of which were novel to the participants. Illegal sequences were produced by beginning each with a legal element, followed by a series of illegal

[2] An example of a matched pair is 4-1-3, 4-1-3; an example of a mismatched pair is 1-2-5-5, 1-2-1-3.

transitions, and ending with a legal element once more. An illegal transition denotes that a particular pair of elements does not occur together during training. For example, the illegal sequence 4-2-1-5-3 begins and ends with legal elements (4 and 3, respectively) but contains several illegal interior transitions (4-2, 1-5, and 5-3 do not occur during training). In this manner, the legal and illegal sequences differ from one another in terms of the statistical relationships of adjacent elements.[3]

Procedure Participants were assigned randomly to either a control group or an experimental group. The experimental group participated in both a training and a testing phase, whereas the control group only participated in the testing phase. Before beginning the experiment, participants were assessed by the Edinburgh Handedness Inventory (Oldfield, 1971) to determine their preferred hand. Then, using velcro straps, the experimenter placed a vibration device onto each of the five fingers of the preferred hand.

At the beginning of the training phase, the experimental participants were instructed that they were participating in a sensory experiment in which they would feel pairs of vibration sequences. For each pair of sequences, they had to decide whether the two sequences were the same or not, and indicate their decision by pressing a button marked "YES" or "NO". This match-mismatch paradigm used the twelve training pairs described earlier. It was our intention that this paradigm would encourage participants to pay attention to the stimuli while still allowing incidental learning of the statistical structure to occur.

After the last sequence of each pair, a 1 s pause occurred, followed by a prompt on the screen asking for the participant's response. After the participant made a response, there was a 2 s inter-trial interval before the next pair began.

Each pair was presented six times in random order for a total of 72 exposures, the entire training phase lasting roughly ten minutes. A recording of white noise was played during training to mask the sounds of the vibrators. In addition, the participants' hands were covered by a cardboard box so that they could not visually observe their fingers. These precautions were taken to ensure that tactile information alone, without help from auditory or visual senses, contributed to task performance. As mentioned previously, the experimental group—but not the control group—participated in the training phase.

Before the beginning of the testing phase, the experimental participants were told that the vibration sequences they had just felt had been generated by a computer program that, using a complex set of rules, determined the order of the pulses. They were told that they would now be presented with new vibration sequences. Some of these would be generated by the same program while others would not. It was the participant's task to classify each new sequence accordingly (i.e., whether or not the sequence was generated by the same program) by pressing a button marked either "YES" or "NO." The control participants received the same instructions and task except that there was no reference made to a previous training phase.

The twenty test sequences were presented one at a time, in random order, to each participant. The timing of the test sequences was the same as that used for the training sequences.

Results

The training performance for each experimental participant was assessed by calculating the mean percentage of correctly classified pairs. This calculation revealed that participants, on average, made correct match-mismatch decisions for 74% of the trials.

Results from the testing phase revealed that the control group correctly classified 45% of the test sequences while the experimental group correctly classified 62% of the test sequences. Following Redington and Chater's (1996) suggestions, two analyses were conducted on the data. The first was a one-way analysis of variance (ANOVA; experimental vs. control group) to determine whether any differences existed between the two groups. The second compared performance for each group to chance performance (50%) using single group t-tests.[4]

The ANOVA revealed that the main effect of group was significant, $F(1, 18) = 3.16$, $p < .01$, indicating that the experimental group performed significantly better than the control group. Single group t-tests confirmed the ANOVA's finding. The control group's performance was not significantly different from chance, $t(9) = -1.43$, $p = .186$, whereas the experimental group's performance was significantly above chance, $t(9) = 2.97$, $p < .05$.

The results show that the experimental group significantly outperformed the control group. This suggests that the experimental participants learned aspects of the adjacent element statistics inherent in the training sequences, allowing them to classify novel test sequences appropriately. This is the first empirical evidence of a tactile sequential learning system of such complexity to enable participants to make judgments regarding the legality of artificial grammar-generated sequences.

[3] In addition, Gomez and Gerken (1999) matched the legal and illegal sequences in terms of element frequencies and length so that these factors could not influence performance.

[4] Ideally, the control group should perform at chance levels while the experimental group should perform significantly better than both chance and the control group.

Experiment 2: Visual Sequential Learning

Experiment 2 assessed sequential learning in the visual domain. This experiment was identical to Experiment 1 in terms of the general procedure and the timing of the stimuli; however, instead of vibrotactile pulses, the sequences consisted of flashing squares occurring at different spatial locations. The reason for using such stimuli, as opposed to letters, for example, was to provide as close a match as possible to the tactile stimuli used in the first experiment. Importantly, unlike sequences of letters, the vibrotactile sequences consisted of non-linguistic, spatially-distinct elements that were presented one at a time (sequentially). The visual stimuli used for this second experiment shared these same characteristics; therefore, the resulting data should provide a meaningful basis for comparison with the first experiment. Like Experiment 1, there was an experimental group, undergoing training and testing phases, and a control group, undergoing the testing phase only.

Only a handful of statistical learning studies have used non-linguistic visual stimuli in a truly sequential manner (e.g., Fiser & Aslin, 2002; Kirkham et al, 2002). The data suggest that such a presentation does not hamper sequential learning by vision. However, other studies (e.g., Handel & Buffardi, 1969) indicate that for certain temporal processing and pattern learning tasks, vision may be inferior to touch. This experiment aimed to investigate whether such differences would be observed.

Method

Participants An additional 20 undergraduates (10 in each condition) were recruited from introductory Psychology classes at Cornell University. Subjects received extra credit for their participation. The data from three additional participants were excluded for the following reasons: did not adequately follow the instructions (n=2); equipment malfunction (n=1).

Apparatus The apparatus was the same as Experiment 1, except for the exclusion of the vibration devices.

Materials The sequences were identical to those of Experiment 1 except that instead of vibrotactile pulses, they were composed of flashing black squares displayed on the computer monitor (1 was the leftmost location and 5 was the rightmost). Each flashing square appeared for 250 ms and was separated by 250 ms. Thus, 1-2-5-5 represents a sequence consisting of a flash appearing in the first location, then in the second location, followed by two flashes in the fifth location.

Procedure The procedure was the same as that of Experiment 1, the only differences relating to the nature of the stimuli presentations, as described above. The timing of the stimuli were identical to those of Experiment 1.

Results

The same statistical analyses as used in Experiment 1 were performed. During the training phase, the experimental group participants made correct match-mismatch decisions on 86% of the trials. A comparison of means across the two experiments revealed a significantly higher training performance in Experiment 2, $F(1, 18) = 14.21$, $p < .01$.

Results for the testing phase revealed that the control group correctly classified 47% of the test sequences while the experimental group correctly classified 63% of the test sequences. An ANOVA (experimental vs. control group) indicated that the main effect of group was significant: $F(1, 18) = 3.15$, $p < .01$. Single group t-tests revealed that the control group's performance was not significantly different from chance, $t(9) = -1.11$, $p = .3$, whereas the experimental group's performance was significantly different from chance, $t(9) = 3.03$, $p < .05$.

The results indicate that the experimental group significantly outperformed the control group. In addition, overall experimental and control group performance at test was very similar to that observed in Experiment 1, suggesting commonalities between tactile and visual sequential learning.

Experiment 3: Auditory Sequential Learning

Experiment 3 assessed sequential learning in the auditory domain. This experiment was identical to Experiments 1 and 2 except that it used sequences of auditory tones. Like the previous experiments, Experiment 3 had an experimental group, undergoing training and testing phases, and a control group, undergoing the testing phase only. Although previous research has found similar statistical learning performance in vision and audition (Fiser & Aslin, 2002), other data suggest that audition excels at sequential processing tasks (Handel & Buffardi, 1969; Sherrick & Cholewiak, 1986); therefore, we might expect to see a difference in auditory compared to visual and tactile sequential learning.

Method

Participants An additional 20 undergraduates (10 in each condition) were recruited from introductory Psychology classes at Cornell University.

Apparatus The apparatus was the same as Experiment 2. The auditory tones were generated using the SoundEdit 16 version 2 software for the Macintosh.

Materials The sequences were identical to those used in the previous experiments except that instead of vibrotactile pulses or flashing black squares, they consisted of musical tones beginning at middle C (1 = C, 2 = D flat, 3 = F, 4 = G flat, and 5 = B).[5] Each tone lasted 250 ms and was separated by 250 ms. Thus, the sequence 1-2-5-5 consists of a C, then a D flat, and lastly two B's.

Procedure The overall procedure was the same as that of the previous experiments.

Results

During the training phase, the experimental group participants made correct match-mismatch decisions on 96% of the trials. This training performance was significantly higher than that of Experiment 2, $F(1, 18) = 10.20$, $p < .01$.

Results for the testing phase revealed that the control group correctly classified 44% of the test sequences while the experimental group correctly classified 75% of the test sequences. An ANOVA (experimental vs. control group) indicated that the main effect of group was significant: $F(1, 18) = 7.08$, $p < .001$. Single group t-tests revealed that the control group's performance was marginally worse than chance, $t(9) = -2.25$, $p = .051$, indicating that our test stimuli were biased against a positive effect of learning. The experimental group's performance was significantly different from chance, $t(9) = 7.45$, $p < .001$.

Like the previous experiments, the data indicate that the experimental group significantly outperformed the control group; hence, participants appeared to learn aspects of the statistical structure of the input. In fact, the experimental group test performance appears to be substantially greater compared to those of Experiments 1 and 2 (75% vs. 62% and 63%).

General Discussion

Assessing first the training results, we found that performance was significantly different across all three experiments (audition = 96%; vision = 86%; touch = 74%). Because the training task essentially involves remembering and comparing sequences within pairs, the results may elucidate possible differences between the three modalities in representing and maintaining sequential information (Penney, 1989). It is also possible that these results are due to factors such as differential discriminability or perceptibility of sequence elements in different sensory domains.

[5] This particular set of notes was used because it avoids familiar melodies.

The testing results for all three experiments are summarized in Figure 2. All three experiments are similar in that the experimental group test performances were significantly different from both chance and their respective control groups. From these results, it appears that participants learned aspects of the adjacent element statistical structure inherent in the training input, allowing them to classify novel stimuli. In this manner, tactile, visual, and auditory sequential learning display commonalities. It is especially interesting to note that sequential learning is not limited to the visual and auditory modalities, but extends to touch as well.

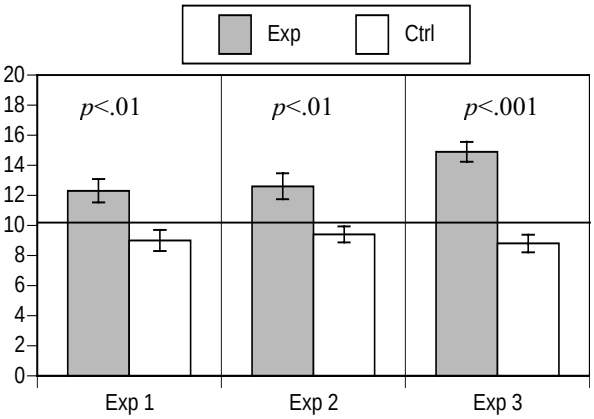


Figure 2: Summary of test results (# of correct responses out of 20).

Despite this overall similarity across modalities, it is also apparent that the Experiment 3 (auditory) results are somewhat different from the other two experiments. Specifically, the auditory experimental group performed better at test as compared to the tactile and visual experimental groups (75% vs. 62% and 63%). This difference is in fact significant [$F(1, 54) = 6.03$, $p < .05$].[6] Thus, it appears that in this task, auditory sequential learning was more successful than both tactile and visual learning. While previous research has suggested that audition excels at relatively low-level temporal processing tasks (Mahar et al., 1994; Sherrick & Cholewiak, 1986), our results appear to be the first evidence that such an advantage extends to complex temporal processing, namely statistical/sequential learning. This auditory advantage perhaps is related to the finding that adults process tone sequences by representing relative, as opposed to absolute, pitch (Saffran & Griepentrog, 2001); such a strategy may allow for more efficient encoding of adjacent element statistics.

[6] This was computed by contrasting the means of the experimental and control groups, as illustrated by the equation: E3-C3 = .5(E1-C1)+.5(E2-C2), where E and C refer to experimental and control group means, respectively.

Conclusion

It has been argued that statistical learning is subserved by a single, domain-general mechanism (e.g., Kirkham et al., 2002). Although a single-mechanism view may be theoretically attractive, our results point toward another possibility: that sequential learning may involve multiple, modality-constrained processes. This idea is supported by a recent multivariate meta-analysis of 35 PET experiments (Lloyd, 2000), which suggested that computations in the different "sensory streams" (i.e., representations of tactile, visual, and auditory information) rely on entirely different cortical areas altogether, at all levels of processing. Additionally, neuroimaging evidence specifically related to sequential learning is consistent with a multiple-mechanism view (see Clegg, DiGirolamo, & Keele, 1998). Thus, we propose that sequential learning is best understood as a functional "suite", composed of multiple, modality-constrained mechanisms. Each mechanism is instantiated in largely non-overlapping brain areas but some degree of interaction is likely to occur between them. We further suggest that each modality-constrained mechanism shares similar computational properties with one another, including the ability to extract adjacent element statistics from incidental exposure to input. However, because each learning mechanism is largely tied to specific sensory areas, each is constrained by the global properties of that sensory system. These properties presumably relate to the types of information that each sensory modality is specialized to process, such as temporal, spatiotemporal, or spatial configurations (Mahar et al., 1994). Our experimental data illustrate one example of such specialization: the auditory system encoded statistical information of temporal input more effectively than did the other senses. Important targets for future research include further substantiating this multiple mechanism view of sequential learning and to discover how such modality-constrained systems might interact with each other, as well as how each relates to human cognition in general. We anticipate that such future research, especially that involving neurophysiological experimentation, will further elucidate the nature of sequential learning by touch, vision, and audition.

Acknowledgments

We thank Dick Darlington, David Gilbert, Erin Hannon, Scott Johnson, Natasha Kirkham, and Michael Young for their feedback on parts of this research.

References

Clegg, B.A., DiGirolamo, G.J., & Keele, S. (1998). Sequence learning. *Trends in Cognitive Sciences, 2*, 275-281.

Cohen J.D., MacWhinney B., Flatt M., & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, *25*, 257-271.

Conway, C.M., & Christiansen, M.H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*, 539-546.

Craig, J.C., & Rollman, G.B. (1999). Somesthesis. *Annual Review of Psychology*, *50*, 305-331.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458-467.

Gomez, R.L., & Gerken, L.A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.

Handel, S., & Buffardi, L. (1969). Using several modalities to perceive one temporal pattern. *Quarterly Journal of Experimental Psychology*, *21*, 256-266.

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: Evidence for a domain-general learning mechanism. *Cognition, 83*, B35-B42.

Lloyd, D. (2000). Terra cognita: From functional neuroimaging to the map of the mind. *Brain & Mind, 1*, 93-116.

Mahar, D., Mackenzie, B., & McNicol, D. (1994). Modality-specific differences in the processing of spatially, temporally, and spatiotemporally distributed information. *Perception*, *23*, 1369-1386.

Manning, S.K., Pasquali, P.E., & Smith, C.A. (1975). Effects of visual and tactual stimulus presentation on learning two-choice patterned and semi-random sequences. *Journal of Experimental Psychology: Human Learning and Memory*, *1*, 736-744.

Oldfield, R. L. (1971). The assessment of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97-113.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*, 401-426.

Penney, C.G. (1989). Modality effects and the structure of short-term verbal memory. *Memory and Cognition, 17*, 398-422.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior, 6*, 855-863.

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123-138.

Saffran, J.R. & Griepentrog, G.J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*, 74-85.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.

Sherrick, C.E., & Cholewiak, R.W. (1986). Cutaneous sensitivity. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance, Vol. I: Sensory processes and perception*). New York: Wiley & Sons.

Feedback Effects in the Acquisition of a Hierarchical Skill

Andrew Corrigan-Halpern (ahalpe1@uic.edu)
Stellan Ohlsson (stellan@uic.edu)
UIC Department of Psychology, 1007 W. Harrison
Chicago, IL 60607

Abstract

Complex cognitive skills cannot be learned without feedback. There are reasons to believe that positive and negative feedback function differently. For skills that are represented hierarchically, feedback can provide information locally, concerning individual actions, or more globally, referring to higher level goals. A 2-by-2 factorial experiment showed an interaction between type of feedback (positive or negative) and scope of feedback (global or local). Contrary to the wide spread belief in the effectiveness of positive feedback, local negative was most effective. Global negative feedback was least effective. Positive feedback fell between these two conditions and was less affected by scope. The results are discussed in terms of the information content of the feedback.

Introduction

A large part of learning consists of gradually fine-tuning our efforts until we have succeeded in the task at hand. Learners may be able to evaluate progress without aid in simple tasks, but for complex tasks they rely on feedback (Kluger & DeNisi, 1996). For complex tasks, performance will often involve the execution of many goals, each involving the application of component skills (Kotovsky & Simon, 1973).

Many of the studies that employ feedback use it in the service of other research interests and are less concerned with comparing one feedback condition to another. Feedback is used to establish the levels of learning that the purpose of the experiment requires, but is not itself the object of study. Consequently, we know less about the cognitive function of feedback than one might expect. Studies that do focus on the feedback itself report contradictory findings and no current theory can fully explain why feedback is effective.

Some contrary findings may be explained by the fact that studies take one of two approaches, either investigating the informational or the motivational impact of feedback (Kluger & DeNisi, 1996). However, inconsistencies are not eliminated when we limit our survey to one or the other approach. In this article, we focus on the information aspect of feedback.

Background and Rationale

It seems obvious that if feedback is helpful, then more feedback should be more beneficial than less feedback. Many investigations have studied the impact of feedback rate (Salmoni, Schmidt, & Walter, 1984). In some cases, increasing the rate improves performance (Kulik & Kulik, 1988; Salmoni et al., 1984; Schmidt et al., 1989; Thorndike, 1927). In other studies, increasing the feedback rate is found to hinder performance (Bourne, 1957; Bourne & Bundeson, 1963; Schroth, 1997). This indicates that the effect of feedback is strongly mediated by other variables.

We would expect the content of the feedback itself to be one such variable. Studies in social psychology suggest that feedback is less effective when it draws attention to general performance or to self-efficacy and it is most effective when it draws attention to the task itself (Kluger & DeNisi, 1998). Another relevant variable is the type of task the learner is acquiring knowledge about. For complex tasks, feedback might foster more active processing, enhancing performance. For simple tasks, constant feedback may interfere with deeper evaluation, reducing performance. (Schmidt & Wulf, 1997). A third variable that affects the outcome reported in a study is the manner in which feedback is measured. The most effective feedback condition as measured during training might not be the most effective condition as measured by transfer measures (Schmidt, Young, Swinnen & Shapiro, 1989; Scroth, 1997).

In the study reported in this paper, we focus on yet another dimension of feedback: whether it provides information about appropriate, correct and useful actions, which we refer to as *positive feedback*, or whether it provides information about errors and mistakes, which we call *negative feedback*. Empirical work supports the notion that negative and positive feedback have qualitatively different effects (Taylor, 1991).

There is widespread belief that positive feedback is more effective than negative feedback. A simple argument about *information content* supports this belief: When a learner receives information that an action is appropriate, correct or useful, this requires neither change nor interpretation. The straightforward implication is that he or she should repeat that action when a similar situation arises in the future. When a learner receives information that his or her action was an error, he or she merely knows to avoid that particular action in the future, but not which action to perform instead. Positive feedback provides definite information, while negative feedback requires interpretation of the cause of the error and the selection of an alternative action. The interpretation process for negative feedback can be quite complex (Ohlsson, 1996).

The argument from information content is even stronger if we factor in the *scope* of the feedback. Many situations in real life do not provide feedback immediately after every elementary action. The effects of one's behavior might be delayed until after a series of actions has been completed

(Salmoni, Schmidt, & Walter, 1984). We call feedback that refers to a single action as being *local* in scope. Feedback that refers back to a series of actions is *global* in scope. The interpretation problem for global negative feedback is even more difficult than for local feedback. If a series of actions ends in an undesirable outcome, which part of the underlying cognitive skill should be affected by this feedback? If the sequence of actions is controlled by a subgoal, which in turn is dominated by a superordinate goal, which goal or goals should be affected by the negative feedback? During problem solving, a correct higher goal can dominate an incorrect lower goal, and vice versa. In prior work with a simulation model, we have shown that situations can arise in which a hierarchical, feedback driven system fails to learn because a superordinate and correct goal cannot recover from the negative feedback generated by incorrect lower level goals or actions (Corrigan-Halpern & Ohlsson, 2001). Ohlsson & Halpern, 1998). In contrast, global positive feedback does not seem to pose a more complex interpretation problem than local positive feedback. Learning that an entire sequence of actions was correct should be more powerful than receiving similar information about a single action.

However, empirical and theoretical research is not fully in accord with the implications of the argument from information content. Some researchers have indeed found that positive feedback is more effective than negative feedback (Greeno, 1974). However, in other cases, negative feedback was found to produce better performance (Mesch, Farh & Podsakoff, 1994). In prior work using a computer simulation model, we showed that some combinations of learning mechanisms imply that negative feedback will have a stronger impact on learning rate than positive feedback (Ohlsson & Jewett, 1997). The interaction with scope appears not to have been systematically investigated.

Closer scrutiny of the information content argument itself shows that it overlooks potential interactions with some of the factors mentioned above. The difference between practicing a simple task – drill – and solving an unfamiliar problem might be important. During problem solving, many actions are taken tentatively, with little or no rationale. When such an action fortuitously generates positive feedback, the learner only gains knowledge about how to solve that particular problem. To learn something that transfers to a related but different problem, he or she must figure out *why* the action worked. Hence, in this case, positive feedback requires interpretation. On the other hand, when the learner has acquired a rationale for his or her responses, then the positive feedback arrives after learning.

In summary, a straightforward application of the information content argument suggests that positive feedback should be more effective than negative feedback and that the advantage should be greater for large than for small scope, but this argument overlooks potential complicating factors. In the present study, we aimed to increase available information about these issues by systematically varying both feedback type and feedback scope, and by assessing the outcome during learning, after learning, and with transfer tasks.

Method

Participants. Ninety-four undergraduates participated in return for class credit.

Task. The subjects mastered a version of the sequence extrapolation task studied by, among others, Simon and Kotovsky (1963). The subject is shown a series of letters that follow a specifiable pattern (e.g., MABMCDM ...). Then he or she is asked to extrapolate the pattern to N additional places (EFMGHM...). We used a type of extrapolation problem that incorporated the hierarchical organization typical of the related problems studied by Restle (1970).

To make letter extrapolation into a task with multiple opportunities to receive feedback, we presented the given sequence via several short presentations and asked for a response after each one. The subjects viewed the given sequence for 20 seconds, then attempted to extrapolate it. They were asked to reproduce as much of the pattern as they could, guessing the letters for which they were uncertain. They received feedback on their extrapolation as described in detail below. Then the next trial (20 second study period, plus extrapolation attempt) began. The subjects went through 12 such trials.

To make the problem more difficult, a different sequence of letters was presented on each trial. For example, suppose a subject studied the sequence C A D F F D A C D B E G G E B D. He or she was prompted with the letter M and the correct extrapolation was M K N P P N K M N L O Q Q O L N. On the following trial, he or she saw a new instantiation of the pattern, e.g., F D G I I G D F G E H J J H E G. He or she was again prompted with M and the correct extrapolation was once again M K N P P N K M N L O Q Q O L N. The extrapolation prompt (M), and hence the correct extrapolation, was the same on each trial.

The tasks were presented on a 15 inch computer monitor with help of the PsyScope experimental control software. The given sequences were presented in black letters on a white background. When the 20 second study period ended, the given sequence was erased and the prompt M appeared on the screen to the left of a horizontal row of 15 answer boxes. The subjects gave their answer by clicking on the answer boxes in any order and typing in a letter. When done, they clicked on a 'Done' button, and the next trial commenced.

Design. In a 2-by-2 between-subjects design, we varied feedback with respect to type (positive vs. negative) and scope (local vs. global). *Negative feedback* consisted of the word "wrong" appearing in red text underneath an incorrect response. *Positive feedback* consisted of the word "correct" appearing in green text underneath a correct response. Each subject saw only one type of feedback.

Because feedback is intermittent in most real learning scenarios, we provided feedback probabilistically. A subject in a negative feedback condition received feedback on a random selection of 75% of his or her errors. For the remaining 25% of erroneous responses, plus all his or her correct responses, the subject saw the word "none" in white letters on black background below his or her response. The instructions emphasized that "none" meant that the response was *either* incorrect *or* correct. This feature prevented subjects from inferring that a response that did not receive negative feedback was correct (and vice versa). The presentation of positive feedback was analogous.

By *local feedback* we mean information about the correctness of a single letter. In contrast, *global feedback* referred to the natural chunks of the extrapolated sequence. Consider the given sequence CADFFDACDBEGGEBD. It consists of two parts, CADFFDAC and DBEGGEBD, which have the same structure but are separated by one position in the alphabet. The first part consists in turn of two subparts, CADF and FDAC, which are mirror images of each other; similarly for DBEG GEBD. Hence, the correct extrapolation MKNPPNKMNLOQQOLN consists of the four chunks MKNP, PNKM, NLOQ, and QOLN. In the global conditions, feedback was given with respect to these chunks by drawing a line underneath each group of four answer boxes; the word "correct", "wrong" or "none" appeared under the center of that line. The instructions emphasized that the feedback referred to the entire group of for letters. Thus, negative feedback meant that *at least one* of the responses in that group was wrong. In order to be get global positive feedback, all the responses in a chunk had to be correct.

In both the local and global conditions, the feedback remained on the screen for 45 seconds before the next trial commenced.

Procedure. The experiment began with an *Instruction* stage, in which subjects were given general information about the experiment. They were also taught the three relevant pattern construction operations (displace a letter or a group of letters one position forwards or backwards in the alphabet; repeat a letter or group of letters, and extend a sequence with its own reversal) with both verbal descriptions and examples. In the *Training* stage, the subjects went through twelve trials with respect to the target problem, as described above. In the *Assessment* stage, the subject solved the target problem twice without prior presentation of the given sequence and without feedback. That is, the prompt letter M appeared together with the answer boxes, and the subject attempted to fill them in; then M appeared again and the subject filled in the answer boxes once more. In the *Transfer* stage, the subject tried to solve a letter sequence problem with the same pattern as the pattern in the target problem. However, the prompt letter was T instead of M. In this case, the correct extrapolation consisted of a completely different sequence of letters. No feedback was given on the transfer problem.

Results

Analysis of Learning Stage

A Mixed ANOVA was performed to assess performance during the learning stage. All 12 Training trials were entered as a within-subjects factor. Feedback Scope (local or global) and feedback Type (positive or negative) were entered as between-subjects factor. The dependent measure was the number of letters correct per trial. There was a significant learning effect as shown in Figure 1, F (990, 11) = 42.21, $p < .001$. There was no effect of feedback Scope, F (90 , 1) = 1.33, $p > .05$. There was no effect of feedback Type, F (90 , 1) = 0.31, $p > .05$. There was an interaction between Scope and Type, F (90 , 1) = 8.45, $p = .005$, indicating that the effectiveness of either feedback type is mediated by scope.

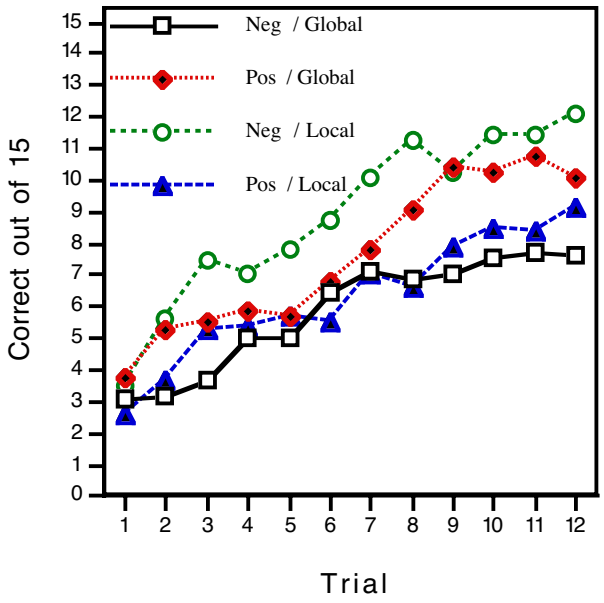


Figure 1. Performance on the training trials.

We hoped that by training subjects on the three construction rules, and by providing problems with a chunking structure, we would maximize the likelihood that they would represent the pattern hierarchically. To test whether this manipulation worked, we examined the inter response time for each position.

We computed solution times for each of the positions of the pattern. There was no time for the first letter of the pattern since this was given to subjects in the form of the prompt "M". We analyzed the last four trials of training, so that we could capture the final product of learning, the point where subjects would have come to represent the pattern hierarchically. Four subjects were removed from the analysis because they omitted responses on one or more of the trials. Figure 2 shows the result.

Subjects spent the longest times at the beginning of the pattern. The first three responses "K" "N" and "P" correspond to the first chunk. The high latency for the first 'K" probably reflected initial time planning to reproduce the pattern. The next four positions (5-8) correspond to the second chunk, "PNKM". Since subjects could reproduce these letters by reflecting the first chunk, response times were much quicker and there was a trend for these positions to form a horizontal line. The next chunk "NLOQ" involves the translation of the first chunk by one letter in the alphabet. This required more effort as indicated by the increase in latency from position 8 to position 9. The last chunk, "QOLN", could also be completed by reflection. Latencies match those of the second chunk and the horizontal trend is again present.

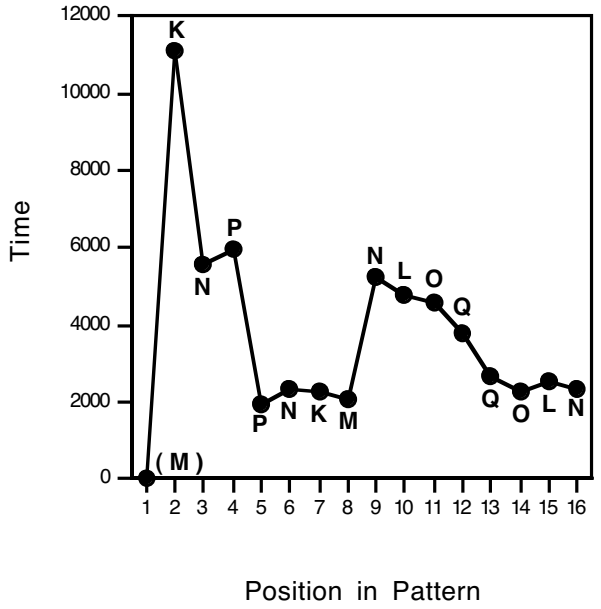


Figure 2. Time to complete each position averaged across the last 4 learning trials.

Analysis of Assessment and Transfer Stages

A mixed within-subjects ANOVA was performed for the Assessment trials. Both trials of the Assessment stage were entered as within-subjects measure. Type and Scope were entered as between-subjects measures. The dependent measure was the number of letters correct per trial.

There was no effect of feedback Type, $\underline{F}$ (90 , 1) = 0.10, $\underline{p}$ > .05 and no effect of feedback Scope $\underline{F}$ (90 , 1) = 3.23, $\underline{p}$ > .05. The interaction between Type and Scope was once again significant, $\underline{F}$ (90 , 1) = 7.70, $\underline{p}$ < .01, indicating that the effect of feedback type depends on scope. There were no effects involving the repeated measure. Figure 3 shows the results for the Assessment trials.

The two Transfer trials were entered as a within-subjects measure in an ANOVA. Feedback Type and Scope were entered as between-subjects measures. The dependent measure was the number of letters correct per trial. There was neither an effect of Type nor of Scope, $\underline{F}$ (90 , 1) = 0.05, $\underline{F}$ (90 , 0) = 2.15, p > .05. The interaction between Type and Scope was once again significant, $\underline{F}$ (90 , 1) = 7.58, p < .01. Figure 4 shows the results for the Transfer trials.

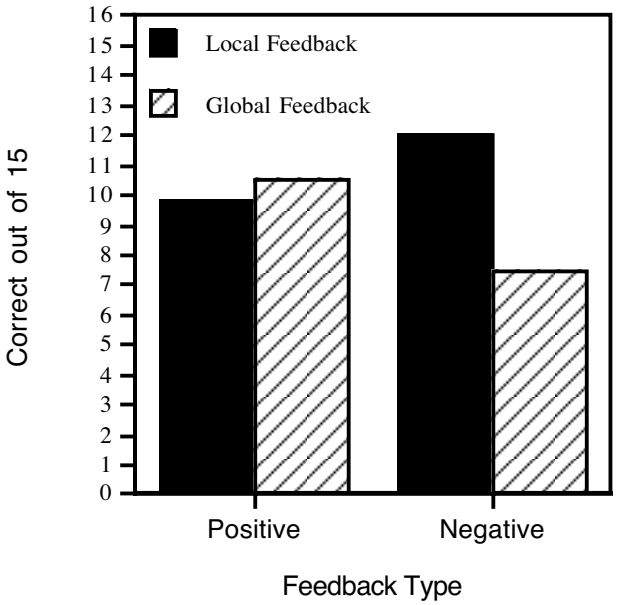


Figure 3. Accuracy for the two Assessment trials.

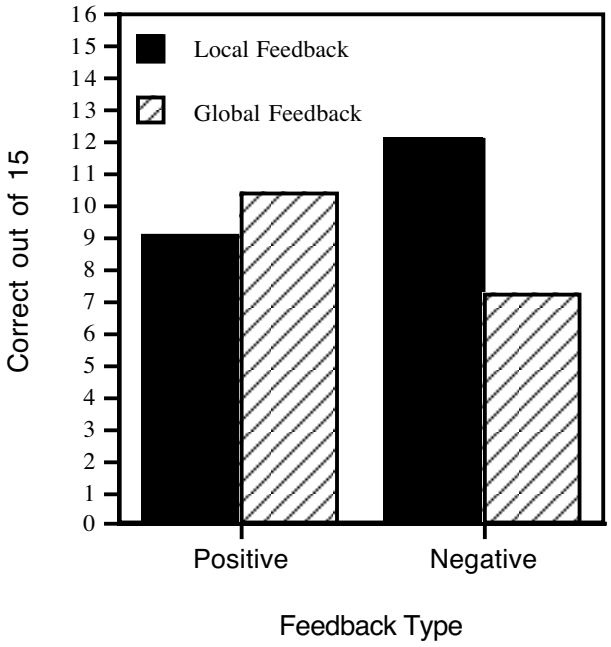


Figure 4. Accuracy for the two Transfer trials.

In summary, we first showed that learning occurred during the Training stage of the experiment and we verified that subjects were representing the pattern hierarchically We showed that during all of the stages of the experiment there was an interaction between Scope and Type. This interaction is driven by the fact that negative feedback is maximized when provided locally, $F(90, 1) = 9.88$, $p < .005$, while for positive feedback, there is no significant effect of Scope, $F(90, 1) = 1.36$, $p > .05$.

Discussion

We presented data to show an interaction between feedback Type and feedback Scope. Local negative feedback was the most effective condition, while global negative feedback was the least effective one. Positive feedback fell between these two in effectiveness. Performance was similar for the two positive feedback conditions, suggesting that for this type of feedback Scope plays a limited role. This exact pattern of outcomes recurred during training (see Figure 1), in the assessment tasks (see Figure 3), and in the transfer tasks (see Figure 4).

This is not the pattern that is predicted by a straightforward formulation of the information content argument, since it predicts that the two positive feedback conditions should be best. According to the latter, negative feedback should provide less information regardless of scope, but this is not what we found. Furthermore, it is unclear what the information content argument predicts with respect to a transfer task in which the exact solution acquired during training is no longer sufficient. Nevertheless, the same pattern was observed in the transfer tasks.

We propose an alternative to the information content view, one where the informational content plays a role, but where it is not the only factor. In this study, negative local feedback was superior to negative global. Since negative local feedback provides direct information concerning individual responses, it is easier to interpret and hence can more directly be used to change subsequent responses. This explains why local negative feedback was more effective than global, but it does not explain our findings with respect to positive feedback.

Positive feedback applies in either of two cases. During problem solving, responses are sometimes made tentatively, without a rationale or reason. In the case of such fortuitously correct responses, positive feedback requires no less interpretation than negative feedback: *why* was the response correct? When the rationale has been worked out, the feedback arrives after learning has already occurred. The argument from information content ignores the dynamics of learning. Negative feedback is naturally received before learning occurs, and hence can influence and support learning. If the learner already knows how to generate a correct answer, the information in the resulting positive feedback is not novel and hence might not contribute strongly to learning.

This study has multiple limitations. The learning phase was short, only twelve trials. The feedback messages were limited to "correct" versus "incorrect" with none of the explanatory content that would be likely to accompany a feedback message in a realistic instructional situations (but not always in realistic situations where the feedback is 'delivered' by physical reality itself, e.g., in the form of a malfunctioning device). The transfer tasks were only moderately distant from the training tasks. The generalisability of the interaction between scope and type of feedback cannot be determined at this time. Future work will address these limitations.

What emerged clearly in this course of this study is that, contrary to the wide spread impression that feedback during problem solving practice is a topic that has been studied to death, we do not, in fact, know very much about the function of feedback. In particular, we do not understand the space of relevant parameters, and we do not know how currently identified parameters interact. It is plausible that the contradictory findings mentioned in the introduction are due to aggregation over overlooked interactions. A systematic experimental attack on the determinants of feedback effectiveness is warranted.

References

Bourne, L. E. (1957). Effects of delay of information feedback and task complexity on the identification of concepts. *Journal of Experimental Psychology, 54*, 201-207.

Bourne, L. E., & Bunderson, C. V. (1963). Effects of delay of informative feedback and length of postfeedback interval on concept identification. *Journal of Experimental Psychology, 65*, 1-5.

Chi, M. T. H., Feltovich, P. J., & Moser, J. M. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Corrigan-Halpern, A., & Ohlsson, S. (2001). Failure to Learn from Negative Feedback in a Hierarchical Adaptive System. In E. M. Altman & A. Cleermans & C. D. Schunn & W. D. Gray (Eds.), *Fourth International Conference on Computer Modling*. Fairfax, VA: Laurence Erlbaum.

Greeno, J. G. (1974). Hobbits and orcs: Acquisition of a sequential concept. *Cognitive Science, 6*, 270-292.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome feedback in multiple cue probability learning. *Organizational Behavior and Human Performance, 9*, 30-34.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double edged sword. *Current Directions in Psychological Science, 7*(3), 67-72.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4(3)*, 399-424.

Kulik, J. A., & Kulik, C.-l. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research., 58*(1), 79-97.

Mesch, D. J., Farh, J.-L., & Podsakoff, P. M. (1994). Effects of feedback sign on group goal setting, strategies, and performance. *Group & Organization Management, 19(3)*, 309-333.

Ohlsson, S., (1996). Learning from performance errors. *Psychological Review*, *103*, 241-262.

Ohlsson, S., & Halpern, A. (1998). *Strength adjustment in hierarchical learning.* Paper presented at the 20th Annual Conference of the Cognitive Science Society, Madison, WI.

Ohlsson, S., & Jewett, J. J. (1997). Ideal adaptive agents and the learning curve. In J. Brzezinski & B. Krause & T. Maruszewski (Eds.), *Idealization VIII: Modeling in psychology*. Amsterdam, The Netherlands: Rodopi.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77(6)*, 481-495.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: a review and critical reappraisal. *Psychological Bulletin, 95*(3), 355-386.

Schmidt, R. A., & Wulf, G. (1997). Continuous concurrent feedback degrades skill learning: Implications for training and simulation. *Human Factors, 39(4)*, 509-525.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: support for the guidance hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition, 13*(2), 352-359.

Schroth, M. L. (1997). Effects of frequency of feedback on transfer in concept identification. *American Journal of Psychology, 110*(1), 71-79.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70(6)*, 534-546.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The

mobilization-minimization hypothesis. *Psychological Bulletin, 110*(1), 67-85.

Thorndike, E. L. (1927). The Law of effect. *American Journal of Psychology, 39*, 212-222.

Investigating Creative Language: People's Choice of Words in the Production of Novel Noun-Noun Compounds

Fintan Costello (fintan@compapp.dcu.ie)
School of Computer Applications, Dublin City University,
Glasnevin, Dublin 9, Ireland.

Abstract

The production of novel noun-noun compounds is a prime example of everyday linguistic creativity. What cognitive processes guide people's choice of words when they make up a new noun-noun compound? An experiment examined people's production of noun-noun compounds as names for novel objects. The results showed that people's choice of words in these novel compounds was influenced by the diagnosticity of properties in those objects. By contrast, people's choice of words did not seem to be influenced by the communicative precision of the resulting compounds. These results suggest that, in constructing novel compounds, people are guided by conceptual representation rather than communicative task.

Introduction

The production of noun-noun compounds is a prime example of everyday linguistic creativity. Compounds such as *soccer mom* (a middle-class suburban mother) or *alpha geek* (the person in a workplace who knows most about computers) convey a lot of information in a concise and inventive way. How do people produce novel compounds such as these? What cognitive processes guide people's choice of words when they make up a new noun-noun compound? There has been much recent research on the cognitive processes of conceptual combination, which allow people to understand novel noun-noun compounds by combining their constituent words in meaningful ways (Costello & Keane, 2000, 2001; Gagné & Shoben, 1997; Hampton, 1987; Murphy, 1988; Wisniewski & Gentner, 1991). There has also been much work on the situations in which noun-noun compound production occurs, especially in child language (asking whether children create compounds to fill gaps in their lexicon, to mark contrasts, or to allow more precise communication; see Clark, 1987, Clark & Berman, 1984, Windsor, 1993). However, there has been little research on the specific cognitive processes involved in people's creation of novel noun-noun compounds. This paper attempts to address this gap.

This paper describes an experiment examining people's choice of words in novel noun-noun compounds. In this experiment participants are given a description of a novel object and asked to make up a noun-noun compound as a name for that object. The experiment examined the influence of property diagnosticity on people's compound production. Diagnostic properties for a concept are those which serve to identify members of that concept: a diagnostic property is one that most members of a concept have, but most non-members do not have. Previous research has shown that property diagnosticity is important in people's interpretation of compound phrases (Costello & Keane, 2001). In the current experiment, the novel object descriptions presented to participants are controlled for diagnosticity: some containing diagnostic properties for a given concept, others containing non-diagnostic properties. If diagnosticity also plays a role in compound production, then there should be a relationship between the diagnosticity of properties in a novel object description, and people's choice of words when producing compound names for that object.

The current experiment uses materials derived from Costello & Keane's (2001) study of diagnosticity in compound phrase interpretation. The first part of this paper describes this earlier study. The second part describes the current experiment examining the production of novel compounds. To foreshadow the results, this experiment found that diagnosticity was an accurate predictor of compound production: in the experiment the more diagnostic the property in an object description was for a given word, the more likely that word was to be used in generating a compound to name that object. An alternative factor, that of communicative precision (Clark, 1987, 1990), was not a reliable predictor of compound production. The final part of the paper links these findings to other research on concept combination and compound production.

Diagnosticity in the Interpretation of Noun-Noun Compounds

How are people able to understand and grasp the meaning of a noun-noun compound which they have never seen before? When confronted with a novel noun-noun compound, people interpret that compound by combining the compound's modifier concept (the first word in the compound) with the compound's head concept (the second word). People can combine these two parts in a variety of different ways. Three main combination types have been recognised: conjunctive, relational, and property-transfer interpretations (Hampton, 1987; Murphy, 1988; Wisniewski & Gentner, 1991). In conjunctive interpretations people produce a combined concept that is an instance of both concepts being combined (e.g., "a *pet bird* is a bird which is also a pet"). In relational interpretations people assert a relation between the two concepts being combined ("an *apartment dog* is a small dog which lives in city apartments"). In property-transfer interpretations people create a new combined concept by transferring a property from the modifier concept to the head. For example, the compound "elephant pig" might be interpreted as " an *elephant pig* is a pig that has tusks": the transfer of a property from the modifier concept ("elephant") to the head concept ("pig"). These property-transfer combinations have been the focus of much recent research (Costello & Keane, 2001; Gagné, 2000; Wisniewski & Love, 1998). This focus in this paper is on property-transfer combinations.

Costello & Keane (2001) describe an experiment examining people's interpretation of property-transfer combinations. The

experiment was designed to test two differing predictions about these property-transfer combinations. One prediction was that the transferred property would be a structurally aligned difference between the modifier and head concepts; that is, a property in the modifier concept that structurally corresponds to a different property in the head (Wisniewski, 1996). The competing prediction was that the transferred property would be a diagnostic property of the modifier (source) concept; that is, a property that helps identify members of that concept and distinguish it from other concepts (Tversky, 1977; Costello & Keane, 2000). In this experiment, participants were shown 16 novel compound phrases, with 4 different property-transfer interpretations as possible meanings for each phrase. Using two pre-tests, the 4 interpretations for each compound were controlled and crossed for structural alignment and diagnosticity. For example, for the phrase "elephant pig" the 4 interpretations were

Elephant pigs are

- pigs that are big (diagnostic, aligned)
- pigs that are grey (non-diagnostic, aligned)
- pigs that have tusks (diagnostic, non-aligned)
- pigs that are an endangered species (non-diagnostic, non-aligned)

Participants were asked to judge how good or bad they thought each property-transfer interpretation was as a meaning for the compound phrase in question, and to rate the acceptability of each interpretation on a scale going from -3 to +3.

The results of this experiment showed that people's interpretation of property-transfer compound phrases was strongly influenced by the diagnosticity of the transferred property for the modifier concept in the combination. People reliably rated interpretations using diagnostic properties as acceptable, and those using non-diagnostic properties as unacceptable. There was a significant correlation between the acceptability of interpretations and the rated diagnosticity of the properties they contained. Structural alignment had no influence on people's interpretation acceptability ratings in the experiment. In a subsequent test participants were shown the same phrases and simply asked to write down their own interpretations. Again, diagnosticity, but not structural alignment, influenced people's interpretation of the phrases.

Costello & Keane's (2001) materials provide the basis for the current experiment on role of diagnosticity in people's production of novel compound phrases. Costello & Keane's materials were constructed as follows. First, 16 participants in a property-generation task produced lists of properties for the modifier and head concepts of the compounds used in the experiment. For the modifier concept in each compound, 4 frequently-listed properties were selected. The diagnosticity of these 4 properties was then obtained in a diagnosticity-rating task. Another set of participants were shown the selected properties, each property being paired with the modifier concept of the relevant compound. Participants were asked to imagine they were playing the game in which they had to help the other player to guess the concept shown. They were asked to rate how helpful each property would be in helping their partner identify the concept in question. For example, to assess the diagnosticity of properties for the concept "elephant", participants were asked to rate how helpful each of properties:

- are big
- are grey
- have tusks
- are an endangered species

would be in allowing their partner in the game to identify the concept "elephant". Participants rated the helpfulness of the properties on a 7 point scale going from -3 (not at all helpful) to +3 (very helpful). Diagnostic properties were those whose average rating was above 0 on this 7-point scale. Two diagnostic and two non-diagnostic properties were obtained for each modifier concept, and these properties were used to construct the 4 interpretations for the compound phrase in question. In the main experiment, participants rated the acceptability of these interpretations for those compound phrases. The next section describes how these materials were used to investigate the role of diagnosticity in the production of compound phrases.

Diagnosticity in the Production of Noun-Noun Compounds

The previous section described an experiment showing that, in interpreting compound phrases, the property diagnosticity plays an important part. Is diagnosticity also important when people are producing, rather than interpreting, compound phrases? This section describes an experiment addressing this question.

In this compound production experiment, participants were shown descriptions of unusual objects and asked to generate compound names for those objects. The object descriptions used were selected from the interpretations used in the comprehension study. Each of the 16 phrases in the earlier experiment had 4 object descriptions as interpretations, two using diagnostic properties and two using non-diagnostic properties. In the current experiment one diagnostic and one non-diagnostic object description was selected for each phrase. Participants were given the object descriptions alone (and not given the compound phrases). They were then be asked to generate a two-word noun-noun phrase to name that object. For example, some participants were given the object description

- "a special type of pig that has tusks"

and asked to write down a two-word noun-noun phrase to name that special type of pig. Other participants were given the object description

- " special type of pig that is grey",

and asked to come up with a noun-noun phrase for that object. The question of interest was whether participants would produce the phrase which corresponded to that object in the earlier experiment; that is, whether participants would produce the phrase "elephant pig" as a name for "pigs that have tusks". Notice that in the object description the head word for the phrase in question is already given (participants already know that the object described is a type of pig, and so would most

likely use that word as the head of whatever phrase they produce). The focus of analysis in the experiment is therefore on people's choice of modifier word for the phrases they generate. Will people be more likely to produce the expected modifier for object descriptions containing diagnostic properties ("pigs that have tusks") than for those containing non-diagnostic properties ("pigs that are grey")?

Method

Participants. The participants were 18 Dublin City University undergraduates who took part for course credit.

Materials. The materials were 16 sets of object descriptions. Each set contained two object descriptions both of which had the same target phrase consisting of a modifier and a head. (Appendix A shows all 16 sets of object descriptions and target phrases). One object description contained a diagnostic property for the modifier concept in that target phrase; the other object description contained a non-diagnostic property for that concept. Participants in the experiment saw one object description from each set. Participants did not see the modifier concept for the target phrase for that set. The factor of interest was whether participants would produce a phrase containing the target modifier concept. Object descriptions were taken from the materials of Costello & Keane's (2001) experiment on compound phrase comprehension. The framing phrase "a special type of.." was added to each object description (so that, for example, participants were asked to produce a noun-noun compound to name "a special type of pig that has tusks").

Design. All participants saw 16 object descriptions, one from each of the 16 object-description sets. Participants were randomly divided into two groups. The first group of participants obtained the diagnostic object descriptions from one half of the object-description sets and the non-diagnostic descriptions from the other half. The second group of participants obtained non-diagnostic object descriptions from the first half of the object-description sets and the diagnostic object descriptions from the other half.

Procedure. Each participant received a booklet consisting of an instruction sheet followed by 16 object-description sheets in random order. Each object description sheet had an object description at the top of the page, and three slots in which participants were asked to write down three noun-noun compound names for the object described at the top of the page. Each slot provided space for a modifier and a head concept Participants had 40 minutes to complete the task.

The instruction sheet explained to participants that they would be asked to read a set of object descriptions and to respond by writing down some noun-noun phrases which they thought would be good names for the objects described. Four examples of familiar noun-noun phrases were given, to illustrate the type of response required. These examples were marked with a tick, to show that they were the correct type of response (see Table 1). These examples involved various different relationships between the nouns in the phrase.

A pre-test showed that, when asked to produce compound phrases as names for object descriptions, participants sometimes simply reproduced a word from the description rather than coming up with a new noun-noun phrase (producing the compound name "tusked pig" for the object description "a special type pig that has tusks", for example). To avoid uninteresting responses of this type, instruction sheet stressed that, in producing a compound name for an object, participants should not simply reproduce the words used in the description of that object. Participants were shown four examples of compounds which simply reproduced terms from an object description (Table 2, below). These were marked by crosses to show they were incorrect responses.

Table 1. Examples of correct responses in instruction sheet.

Special type of thing	Correct compound
a special bed with hot lightbulbs above it, giving the user a tan.	a "sun bed" ✓
a special bike with strong rugged tyres and frame.	a "mountain bike" ✓
a special padded glove which protects against heat.	an "oven glove" ✓
a special lamp containing moving bubbles of hot, coloured oil.	a "lava lamp" ✓

Results

The noun-noun phrases produced by participants for the object descriptions were analysed by counting, for each object description, the number of participants who produced that description's target phrase as a name for that object description. For the object description "a special type of pig that has tusks", for example, this would mean counting how many participants produced the target phrase "elephant pig" for that object description. Variations of the target phrase were allowed if they clearly included the target modifier; for example, if a participant produced a blending such as "ele-pig" for "a special type of pig that has tusks", that would be taken as an occurrence of the target phrase.

For each target phrase there were two object descriptions; one using a diagnostic property of the modifier concept, the other using a non-diagnostic property of the modifier. For each target phrase, one group of 9 participants saw one object description, the other group saw the other object description; thus each object description was seen by 9 participants. For the object descriptions containing diagnostic properties of the modifier concept, the target phrase was produced by 3 out of 9 participants, on average. For the non-diagnostic object descriptions, the target phrase was never produced.

Table 2. Examples of incorrect responses in instruction sheet.

Special type of thing	Incorrect compound
a special bed with hot lightbulbs above it, giving the user a tan.	a "lightbulb bed" ✗
a special bike with strong rugged tyres and frame.	a "rugged bike" ✗
a special padded glove which protects against heat.	a "protective glove" ✗
a special lamp containing moving bubbles of hot, coloured oil.	a "coloured lamp" ✗

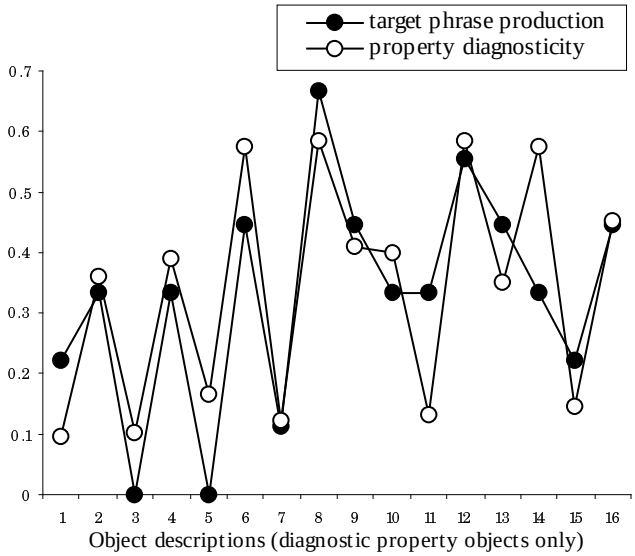


Figure 1. Probability of target phrase production for the 16 diagnostic-property object descriptions, and diagnosticity of properties for modifier in target phrase (diagnosticity mapped to probability scale).

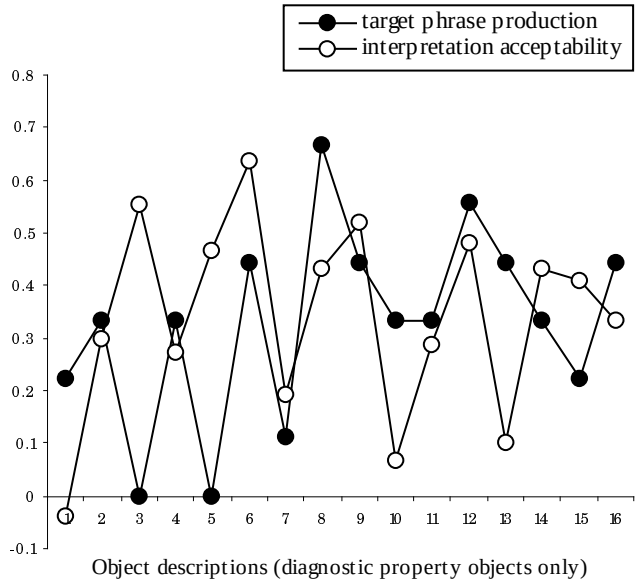


Figure 2. Probability of target phrase production for the 16 diagnostic-property object descriptions, and acceptability of descriptions as interpretations for phrases (acceptability mapped to probability scale).

Results: diagnosticity. The previous analysis showed that participants were clearly more likely to produce the target phrase for descriptions with diagnostic properties than for descriptions with non-diagnostic properties. A more detailed examination of the role of diagnosticity in compound phrase production was performed by analysing the relationship between the diagnosticity of properties in descriptions and the rate of target phrase production for those descriptions. This analysis was carried out for diagnostic-property descriptions only (the only descriptions for which the target phrase was produced). There was a significant correlation between the average rated diagnosticity of properties in these descriptions for the modifier concept (obtained from the pre-test for Costello & Keane's (2001) experiment), and the probability of target phrase production for those descriptions. The more diagnostic the property in an object description was for the modifier, the more frequently participants produced the target phrase ($r = 0.81$, $p < 0.001$, %var = 0.66). Figure 1 shows a graph of probability of target phrase production versus average property diagnosticity for the 16 diagnostic object descriptions. (In this graph property diagnosticity is mapped from its original rating scale of -3 to +3 onto the 0 to 1 interval on which probability of target phrase production is shown).

This result suggests that the production of novel noun-noun phrases for property-based object descriptions can be predicted from the diagnosticity of properties in those object descriptions. It might be argued, however, that the target phrase production in the experiment was not influenced by property diagnosticity per se, but by a more general association between the properties and concepts in question. For example, it could be that the phrase *elephant pig* was produced frequently for the object description "a special type of pig with tusks", not because having tusks is specifically diagnostic in identifying the concept *elephant*, but because there is a general association between elephants and tusks.

To address this possibility a general measure of association between concepts and properties was obtained from Costello & Keane's (2001) original property-generation task. In that task, 16 participants listed properties for the concepts used in the experiment. For each property and concept, the number of participants who listed that property as belonging to that concept was obtained. This number was taken to be a measure of the general association between the property and the concept. The correlation between this property-concept association and the rate of target-phrase production did not reach significance ($r = 0.41$, $p > .1$, %var = 0.17). This suggests that it is diagnosticity specifically, rather than a more general association between properties and concepts, that is important for compound production.

Results: Communicative accuracy. In addition to examining the role of diagnosticity in compound phrase production, the current experiment allows us to examine the role of communicative precision in people's production of compound phrases. One view of compound production is that people produce novel compounds to allow them to precisely communicate their intended meaning to the listener (Clark, 1987, 1990; Clark & Berman, 1984). In this view, compound production is a listener-centered process: a person producing a compound as a name for a given object will choose a compound that the listener would easily interpret as referring to that object. If communicative precision is important in compound production, there should be a relationship between the rate at which people produce a given compound phrase as name for a particular object, and the degree to which people, when given that compound phrase, describe that object as the correct interpretation for the phrase.

The role of communicative precision in compound phrase production was analysed by comparing the rate at which people produced the target phrase for the given object descriptions in the current experiment against the degree to

which participants in Costello & Keane's (2001) study rated those object descriptions as good interpretations for those phrases. The correlation between the phrase production for a given description and description acceptability as an interpretation for that phrase was not significant ($r = 0.1$, $p > .1$, %var= 0.01). Figure 2 shows a graph of the rated description acceptability as interpretation and target phrase production. This finding of no relation between target phrase production and interpretation acceptability casts some doubt on the role of communicative precision in compound phrase production. This finding is in line with other results (e.g. Windsor, 1993; Elbers, 1988) which also call into question the communicative precision view of compound production.

Conclusion

This paper has examined people's production of novel noun-noun compounds. Compound production is quite a creative linguistic task, with compounds often conveying a lot of information in an inventive way (as in the original *soccer mom* and *alpha geek* examples). The experiment reported here found that compound production is influenced by the diagnosticity of properties for concepts in compounds, but not by the communicative precision of those compounds. This is in line with other findings (Windsor, 1993; Eberts, 1988), suggesting that people's production of novel compounds is influenced more by their conceptual representations than by the communicative task that they are carrying out.

Why should diagnosticity play a role in people's novel compound production? Diagnostic properties are properties which help identify items and classify them as members of particular categories. Perhaps diagnosticity is important in compound production because in compound production the choice of words depends on judgements of category membership. According to this suggestion, when producing the phrase "elephant pig" as a name for the object description "a special type of pig with tusks", people use the diagnostic property "tusks" to classify the object described as to some extent falling into the category "elephant", thus allowing the modifier "elephant" to be used in forming the phrase. This is in line with Costello & Keane's (2000, 2001) proposal on the role of diagnosticity in compound interpretation. According to this proposal, when people interpret the phrase "elephant pig" as "a pig with tusks" they are asserting the diagnostic property "tusks" because it allows the newly described object to be classified under the category "elephant", hence justifying the compound (see Costello & Keane, 2000, for a computational model of implementing this approach).

How do the results reported here fit with other work on language production? The main stream of research on language production tends to use a "picture-naming" task to focus on people's production of words in response to a single concept. In this task people are presented with a picture of a single concrete object (e.g. a picture of a rabbit) and are asked to name that object. Models of language production in this framework typically have 5 stages: (i) Preliminary Analysis, (ii) Feature Recognition, (iii) Categorisation, (iv) Lexical Access, and (v) Decision (Bock & Griffin, 2000). The suggestion that diagnosticity is important in compound production because it related to judgements of category membership, fits nicely into the third Categorisation stage in this framework. Indeed, the findings of the current experiment suggest that property diagnosticity may play an important role in the Categorisation stage in this language production framework.

References

Bock, K, & Griffin, Z. M. (2000). Producing words: how mind meets mouth. In Linda Wheeldon, ed. "Aspects of language production", pp 7-49. Psychology Press.

Costello, F. J., & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. *Nineteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint guided conceptual combination. *Cognitive Science,24(2).*

Costello, F. J., & Keane, M. T. (2001). Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27(1),* 255-271.

Clark, E. V. (1987). The principle of contrast: a constraint on acquisition. In B. MacWhinney (ed.) *Mechanisms of language acquisition.* Hillsdale, NJ. Erlbaum.

Clark, E. V. & Berman, R. A. (1984). Structure and use in the acquisition of word formation. *Language*, **60**, 542-590.

Elbers, L. (1988). New names from old words: related aspects of children's metaphors and word compounds. *Journal of Child Language* **15**, 591-617.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. J*ournal of Experimental Psychology: Learning, Memory & Cognition, 23(1),* 71-87.

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition, 15*(1), 55-71.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science, 12*(4), 529-562.

Tversky, A. (1977). Features of similarity. *Psychological Review,* 84, 327-352.

Windsor, J. (1993). The functions of novel word compounds. *Journal of Child Language* 20, 119-138.

Wisniewski, E. J., & Love, B. C. (1998). Properties versus relations in conceptual combination. *Journal of Memory and Language.(38),* 177-202.

Wisniewski E. J. (1998). Property instantiation in conceptual combination. *Memory & Cognition, 26(6),* 1330-1347.

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G. B. Simpson (Ed.) *Understanding word and sentence.* Amsterdam: North Holland.

APPENDIX 1. Object descriptions and target phrases from Experiment

Description Set	Object description		Target phrase (modifier head)
	Property is diagnostic for modifier	**Property is not diagnostic for modifier**	
1	A special type of chair that has blankets	A special type of chair that has wheels	Bed chair
2	A special type of moth that stings	A special type of moth that fertilises plants	Bumblebee moth
3	A special type of oak that stores water	A special type of oak that is completely green	Cactus oak
4	A special type of pig that has tusks	A special type of pig that is grey	Elephant pig
5	A special type of beetle that jumps	A special type of beetle that eats insects	Frog beetle
6	A special type of antelope that has a long neck	A special type of antelope that has a long tongue	Giraffe antelope
7	A special type of airplane that has horizontal rotors	A special type of airplane that is manuverable	Helicopter airplane
8	A special type of monkey that has a pouch to carry young	A special type of monkey that doesn't climb trees	Kangaroo monkey
9	A special type of lobster that has eight limbs	A special type of lobster found in warm waters	Octopus lobster
10	A special type of robin that can talk	A special type of robin that can be a pet	Parrot robin
11	A special type of horse that has a horn	A special type of horse that is very dangerous	Rhinoceros horse
12	A special type of squirrel that smells bad	A special type of squirrel that lives on the ground	Skunk squirrel
13	A special type of spider that has a shell	A special type of spider that eats plants	Snail spider
14	A special type of iguana that slithers	A special type of iguana that is used to make handbags	Snake iguana
15	A special type of slug that is poisonous	A special type of slug that is fast	Viper slug
16	A special type of seal that has a blowhole	A special type of seal that is an endangered species	Whale seal

Do Expression and Identity Need Separate Representations?

Garrison W. Cottrell (gary@cs.ucsd.edu)
Kristin M. Branson (kbranson@cs.ucsd.edu)
Computer Science and Engineering; 9500 Gilman Drive
La Jolla, CA 92093-0114 USA
Andrew J. Calder (andy.calder@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit
15, Chaucer Road
Cambridge, CB2 2EF
UK

Abstract

Recent work has shown that expression recognition shows holistic processing effects much like face recognition (Calder et al., ress). We extend our previous model of facial expression recognition (Dailey et al., 2000) to account for these results. We show that our model, with small modifications to the training procedure, can account for the systematic biases between upper and lower facial expression recognition, and the holistic/configural processing effect. Finally, we show that results that seem to support the idea that separate representations are necessary for emotion and identity processing can be accounted for by a single representation model. This latter effect is demonstrated in subjects by constructing chimeric faces by taking the top half of one face and the bottom half of another.

Background

In recent years a consensus has emerged that face processing is "holistic" in nature (Tanaka and Farah, 1993; Farah et al., 1998). Here "holistic" means "configural": that is, there is an effect of the whole in recognition of the parts of a face. One way to show this is to construct composite faces by combining the upper and lower halves of different faces. If the subject's task is to recognize the identity of the upper half of the face, there is interference if the lower half of the face is from a different person. However, if the lower half of the face is misaligned with the upper half, there is no difference in subjects' responses when the lower half is the same person or a different person (Young et al., 1987).

Recently, it has been shown that this effect extends to facial expression recognition (Calder et al., ress). Calder et al. first replicate a well-known effect that certain expressions are more easily recognized from their top or bottom halves. A summary of the data is shown in Table 1. Based on this data, Calder et al. constructed composite faces from the same subject using a top-biased emotion in the top half and a bottom-biased emotion in the bottom half. When subjects were asked to identify the emotion in one half, their reaction times were slower when the two halves were aligned versus when they were misaligned.

In further experiments, Calder et al. present data that they interpret as showing that there must be separate representations for facial identity and facial expression processing. The experiment in this case was to show several kinds of composite images to the subjects, and to give them two tasks: identity judgements (after training on the identities of the subjects), and expression judgements. There were three kinds of composite images: 1) same identity, different emotion; 2) different identity, same emotion; and 3) different identity, different emotion. Consider the case of identity judgements. The subjects are asked to judge the identity of the subject in the bottom half of the image. If the top of the image is the same subject, but a different emotion, the reaction times are faster than if the top half of the image is a different subject. This is expected based on the above results on configural processing. However, when the top was a different subject, their reaction times did not differ between the case where the expression of the different subject was the same or different. The interpretation is that identity processing is not affected by affect processing. Identical results in the emotion identification task support the idea that expression processing is not affected by identity processing. The conclusion that two representations must therefore be in play makes intuitive sense, but should be tested in a model. In the following, we show that a single representation suffices to obtain these results.

The Model

We performed three experiments that paralleled as closely as possibly three of the experiments reported in (Calder et al., ress). In each experiment, we used the same images from the Pictures of Facial Affect (POFA) dataset (Ekman and Friesen, 1976), although normalized as described below. Also, we constructed our own versions of Calder's hand-constructed split images. All experiments used a similar model and data. The details of these experiments are described in this section.

Classification Model

In all experiments, our classification model employs image filtering, principal components analysis

Expression	Human Top	Human Bottom	Net Top	Net Bottom
Happy	0.20 (.09)	0.01 (.01)	0.40	0.00
Sad	0.19 (.05)	0.34 (.08)	0.28	0.40
Afraid	0.33 (.08)	0.56 (.09)	0.28	0.70
Angry	0.28 (.06)	0.49 (.09)	0.29	0.65
Surprised	0.06 (.21)	0.33 (.07)	0.00	0.21
Disgusted	0.62 (.10)	0.04 (.14)	0.20	0.00

Table 1: Fraction of test examples incorrectly identified for each expression. The Human Top and Human Bottom results correspond to the results reported for expression recognition by (Calder et al., ress). The Net Top and Net Bottom results correspond to the results achieved by our classification model. The number in parentheses is the standard error for the humans.

(PCA), and a single-layer neural network to classify the expression and identity of an input pixel image of an actor posing an expression (Dailey et al., 2000).

Preprocessing of these images begins by aligning the images so that the eyes and mouth of all images are in the same location, then cropping the images to eliminate the background. After each image is aligned, it is convolved with a grid of two-dimensional Gabor jets. Each jet is composed of 40 Gabor filters of five different sizes and eight different orientations. Each jet is centered on a pixel of the aligned image. This image filtering was chosen because it is similar to filtering done in the striate cortex of cats and has previously been shown to improve expression recognition in neural networks (Dailey et al., 2000). Applying Gabor filters to a subsampled 240 x 292 pixel image results in a 40,600 component vector. These vectors are then z-scored (transformed to 0 mean, unit std. dev.) on an individual filter basis, resulting in the Gabor pattern.

As our experiments required a method of directing the classification model's attention to just one half (bottom or top) of the face stimulus, the other half of the face stimulus is attenuated. Each component in the half of the Gabor pattern to be attenuated is multiplied by 0.125. The factor 0.125 was chosen after comparing the results of attenuating by different fractions. An attenuation factor of 0.125 resulted in the error of the model's recognition of expression in half face training data (described below) most closely resembling the error of the human's recognition of expression in half face images. However, we found little variance in the model's error with attenuation factors between 0.5 and 0.125. We define a Gabor filter as being within the half the image to be attenuated if the pixel it is centered on is in that half, or if it is within two times the standard deviation of the Gaussian of the Gabor filter. This way, even filters in the attended half will be attenuated if their receptive field overlaps the other half of the image.

50 principal components of each Gabor pattern are extracted from the training data. These are also z-scored. Finally, a soft-maxed single-layer neural network is trained using the cross-entropy error criterion.

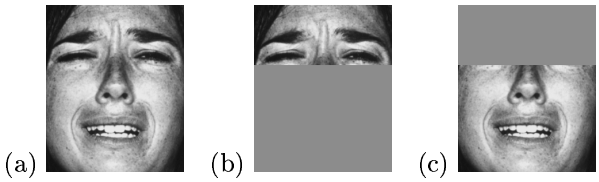


Figure 1: Whole and Half Face Training Data: (a) Aligned and cropped pixel image; (b) Pixels corresponding to the top half; (c) Pixels corresponding to the bottom half.

Training Data

All stimuli are derived from pixel images from the POFA database. This data set includes images of 14 actors posing 6 expressions: happiness, sadness, fear, anger, surprise, and disgust.

In all experiments, the principal components of the Gabor patterns are extracted from whole face and half face Gabor patterns, and the neural network is trained upon those. We assume that during the learning of the emotions, subjects also attend to just one half of the face at different times. We can think of this as either an attentional process or as a crude simulation of eye movements. In addition, splitting the images between the top half and bottom half, as defined by Calder, resulted in the top half of the image being smaller in area than the bottom half of the image. Extracting principal components while attenuating one half of the image allowed more equal representation of each half in the principal components. Whole face Gabor patterns are created by convolving the original pixel image with the Gabor filters, as described above, then z-scoring these patterns. Half face Gabor patterns are created in a similar manner: the original pixel image is convolved with the Gabor filters and z-scored, then one half of the pattern is attenuated to a fraction of 0.125. Thus, when creating a top half face Gabor

pattern, components corresponding to the bottom half of the face are multiplied by the fraction 0.125. An alternative method would have been to start with a pixel image of just one half of the face, and then convolve the image with Gabor filters. We did not do this because the Gabor filters would have given a strong response at the edge in the image between the zeroed half and the non-zeroed half, and we were concerned that this signal might confuse the network (this may have been an unnecessary worry). An example of an aligned and cropped pixel image and bottom and top half images are shown in Figure 1.

Experiments

Experiment 1

The goal of the first experiment was to determine whether our network gave the same results as Calder et al.'s subjects in terms of which expressions are top-biased and which are bottom-biased. An expression is *top-biased* if it is more accurately identified by our model from the top half of the face than the bottom. An expression is *bottom-biased* if the opposite is true. Calder et al. used 10 of the POFA actors for this experiment. We used the same 10 actors.

The general procedure for this experiment was to classify half face examples using the classification model described above, then compare the classification error for top and bottom test half face examples for each expression. We used "hold one actor out" cross validation, so we trained on nine and tested on the tenth. Each of these were repeated ten times using different initial random weights. The test half-face examples classified by the network are the result of convolving a whole face image with the Gabor filters, z-scoring, attenuating one half using the multiplier 0.0 (thus there is no output from that half), and projecting onto the principal components. Training is stopped when error on the training set most closely corresponds to the human confusion matrix reported by Ekman.

Results: The average results of classifying each test example 10 times are shown in Table 1. The network responses do not vary much over networks. These results show that for our classification model, happy and disgusted are bottom-biased while sad, afraid, angry and surprised are top-biased. These results are very similar to the human results: happy and disgusted are bottom-biased and sad, afraid, and angry are top-biased. The only difference is that our classification model finds surprised to be top-biased while the human results find surprised to be unbiased, due to subject variance. In addition, large differences between the network's classification error fractions of the top- and bottom-half stimuli correspond to large differences between humans' classification error fractions of the top and bottom-half images.

Experiment 2

The goal of the second experiment was to determine if incorrect configural information disrupts the model's expression recognition. This involves comparing the model's accuracy on identifying the expression in one half of two different types of stimuli: composite and noncomposite. A composite example is the result of aligning the top half of one face with the bottom half of another (Figure 2(a)). A noncomposite example is the result of misaligning the top half of one face with the bottom half of another. When performance degrades on composite faces compared to non-composites, this is taken to be an indicator of configural processing (Young et al., 1987). That is, when the two halves are aligned, subjects are unable to ignore the information in the other half of the face, even though they are judging only one half.

In this experiment, both halves of the composite and noncomposite examples correspond to the same actor but different expressions. In Calder et al.'s experiment, reaction times were slower for composite images than for non-composite. Composite images are created by aligning the bottom half of one happy, surprised, or disgusted face image with the top half of one sad, afraid, or angry face image. Following Calder et al., images from only four of the actors are used.

Composite Gabor patterns are the result of convolving these composite images with Gabor filters and z-scoring. As the model must identify the expression in one half of the example, one half of the composite Gabor pattern is attenuated by a fraction of 0.125. Noncomposite Gabor patterns are created from the composite Gabor patterns. The half of the pattern that is attenuated is misaligned with the other half by replacing the components of the attenuated half corresponding to the right side of the face with the components corresponding to the left side and zeroing the components corresponding to the left side. These are projected onto the principal components of the same training set as the first experiment. The network is trained as in the previous experiment, and tested on identifying the expression in one half of both composite and noncomposite examples.

Results: The results are shown in Figure 3. These results indicate that our classification performs better on noncomposite test examples than composite test examples. As the top and bottom face halves are aligned in the composite examples, there is incorrect configural information in these examples that is not present in the noncomposite examples. Therefore, this result indicates that incorrect configural information disrupts our model's expression recognition. The trend in classification error for these two types of test examples is similar to the trend reported for human response time for these two types of examples, as shown in Figure 3.

Figure 2: Different stimuli types. (a) Composite/Non-composite images; composites with (b) same identity/different expression; (c) different identity/same expression; and (d) different identity/different expression.

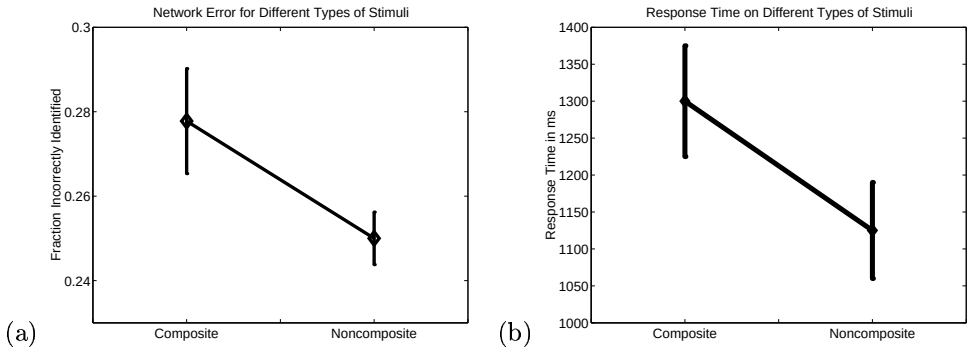


Figure 3: Experiment 2 results. (a) Average error proportion for composite, noncomposite, and half face test examples. Vertical bars indicate standard deviations. (b) Human response times for expression classification of composite and noncomposite test examples

Experiment 3

In Calder et al.'s Experiment 4, they investigated the interference between identity and expression processing. They noticed that when two happy expressions were combined from different individuals, the new image looked like a happy person who was a new individual, different from the two source individuals (see Figure 2(b)). They hypothesized from this that "the configural information used to encode facial identity may be different from that used to code facial expression." They suggested that if the two kinds of processing could be selectively disrupted, then that would be support for this two-representation model. Note that our model has one representation, corresponding to the principal components layer. If we can obtain the same results, then we will show that their conclusion, that there are two representations, is unwarranted.

This experiment involved three types of composite stimuli. Same identity with different expression (SID/DE) composite examples require that the identity of the actor in both halves be the same but the expression in each half be different. Different identity with same expression (DID/SE) composite examples require that the identity of the actor in each half be different but the expression in each half be the same. Different identity with different expression (DID/DE) composite examples require that the identity of the actor in each half be different but the expression in each half be the same. Figure 2(b-d) shows examples of all three types.

The model is trained on all the whole and half face examples for ten actors. Following Calder et al., the model must classify the expression and identity in the bottom half of the composite test examples constructed from three of the actors. In order to test for identity performance, Calder et al. trained their subjects on both identity and expression for the three actors they used to create the composite test stimuli. Hence these stimuli were included in the training set. Three additional localist outputs were included to learn the identity of these three actors. The seven remaining actors were trained to produce all 0's on these units. Following Calder et al., for the three test subjects, the composite examples constructed from them for testing purposes only use the happy, surprised, and disgusted expressions. We stopped training by using a holdout set composed of the remaining four actors from POFA that were not used in any of the Calder et al. experiments. The network was trained until error on the holdout set was minimal, and tested on all three types of composite examples.

Results: The results are shown in Figure 4. While both expression and identity were classified without error on all types of composites, there was a significant difference in a standard measure of reaction time from feed-forward networks: 1 - the maximum output. Figure 4(a) shows the reaction time of our network on the relevant stimuli.

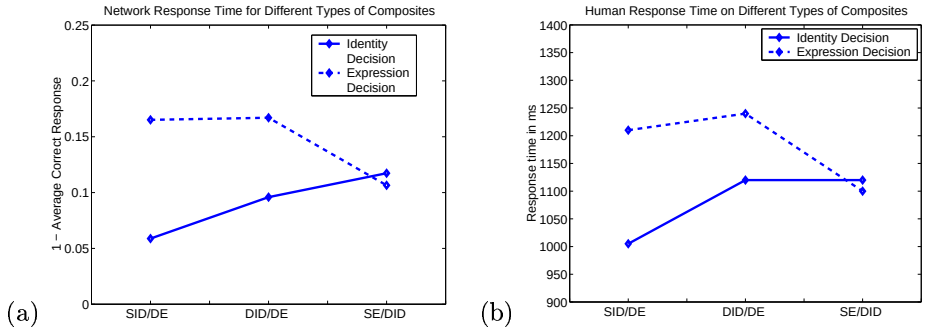


Figure 4: Experiment 3 results. Panel (a) shows the effect of composite type on our network reaction time for identity and expression recognition. Panel (b) is derived from (Calder et al., ress).

When classifying expression (dashed line), the model responds more quickly when the expression is the same in both halves of the composite example. Crucially, the model is not slowed more in the DID/DE case than the SID/DE case. The same result is true for identity classification. When classifying identity, the model is faster when the identity is the same in both halves of the composite example. Crucially, the model is not slowed more in the DID/DE case than the DID/SE case. In fact, it is slightly faster. The pattern of these results are equivalent to the pattern of the response times in the human experiments (Figure 4(b)).

Analysis

In this section we examine the kind of representation that the network is using for this task. In particular, we examine the principal components of the Gabor filters. Calder et al. showed that the principal components of gray scale images taken from the POFA database show a separation in terms of identity and expression. That is, there are components that best separate expressions, and different components that best separate the identities of the models in the POFA.

The principal components that best discriminate a data set, assuming the data samples are normally distributed, are those that minimize the variance of the samples within each class, while maximizing the total variance of all the samples. Therefore, we can rank the approximate discriminating power of each principal component by the Wilk's Λ value: the within-class scatter divided by the total scatter of the samples. The smaller the Wilk's Λ, the greater the discriminating power.

The two Wilk's Λ values for each principal component were computed for the two face recognition tasks compared in this paper: expression recognition and identity recognition. Only two components out of the highest 10 ranked principal components for each task overlapped. The 5th and the 19th principal components were ranked in the top 10 for both expression and identity recognition, but other than that the top ten components for the two tasks were disjoint.

This suggests that the principal components used to encode expression are separate from those used to encode identity. Figure 5(a) and (b) show the projections of the identity classes and expression classes on the most discriminating expression principal components. It is apparent that these components separate expression better than identity. Figure 5(c) shows the identity Wilk's Λ value plotted against the expression Wilk's Λ value, for corresponding principal components.

Discussion

Our results suggest that our simple neural network model can explain a variety of effects in psychological research in expression recognition. We found that our model showed nearly the same pattern of results in discriminability of expressions from half faces. In order to obtain this result, we had to change our model in two ways. First, we modeled the attentional process as an inhibition of irrelevant information, an approach that is well supported in the literature. Second, we had to add training on half faces to our model. We suggest that this modification is independently motivated by the fact that people foveate on different parts of the face when performing such tasks. Future research should concentrate on actually implementing an eye movement mechanism that is modulated by the task. Our previous model of scan paths (Yamada and Cottrell, 1995) only used bottom up information. Top down, task-related information can be incorporated by using the mutual information between the features and the categories – essentially, feature selection.

Second, we showed that our model's error patterns when shown composite versus non-composite faces follow the reaction time pattern in the human subjects data. That is, where our model makes more errors, the humans have longer reaction times. In an unreported experiment, we found the correct pattern

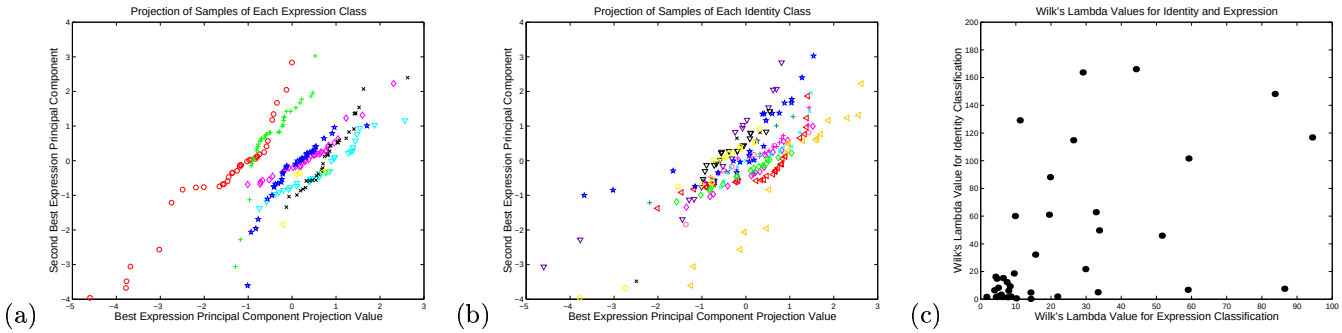


Figure 5: PC Analysis: (a) Plots of each class of expression on the two most discriminating expression principal components; (b) Plots of each class of identity on the two most discriminating expression principal components; (c) Identity and expression Wilk's Λ values for corresponding principal components.

of reaction times when the network is only shown a half face versus a composite face. This suggests that either the human subjects do not look at the other half of the face at all when they are misaligned, or that a greater degree of attenuation of the misaligned half should lead to results more in keeping with the RT data.

Third, our model showed that it is not necessary to posit two independent representations for identity and expression processing. Since the representation at the principal components layer is a set of orthogonal vectors, and the categorizer is a single layer perceptron, this suggests that each output unit is simply cutting off a different corner of the feature hypercube, and the learned hyperplane is simply orthogonal to the non-informative directions of variation. This is exactly what one would expect given this type of model, and so there is no mystery in our results.

This result might also have been expected given previous results examining the principal components of gray scale images of facial expressions directly (Calder et al., 2001), where it was found that expression and identity tended to load on different principal components. The main difference here is that our principal components are performed on a more biologically plausible representation of faces than gray scale images. However, we find that when we do a similar analysis to that carried out by Calder et al. (2001), we also find that the representation loads expression and identity on orthogonal components. Does this mean that we have two representations? If one thinks of each principal component as a linear "neuron," and the projection of an input on that component as its activation, then observing these activities will appear to show units that respond to identity, and other units that respond to expression. One could argue that these are then separate representations.

However, we argue that when viewed ontologically these representations were developed to be orthogonal *with respect to one another*. That is, due to the constrain that the principal components must be orthogonal, the representation of emotion is made in the context of and in competition with the representation of identity, as these are the major directions in which the data vary. On the other hand, one can say that this is a *functional* separation of the representations, and indeed, it is.

Acknowledgments We thank Gary's Unbelievable Research Unit (GURU) for discussion and comments. This research was supported by NIMH grant MH57075 to GWC.

References

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, 41:1179–1208.

Calder, A. J., Young, A., Keane, J., and Dean, M. (In Press). Configural information in facial expression perception. *JEP:HPP*.

Dailey, M., Cottrell, G., and Adolphs, R. (2000). A six-unit network is all you need to discover happiness. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 101–106, Mahwah, New Jersey. Lawrence Erlbaum Associates.

Ekman, P. and Friesen, W. (1976). *Pictures of Facial Affect*. Consulting Psychologists, Palo Alto, CA.

Farah, M. J., Wilson, K. D., Drain, M., and Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, 105(3):482–498.

Tanaka, J. and Farah, M. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology*, 46(2):225–245.

Yamada, K. and Cottrell, G. (1995). A model of scan paths applied to face recognition. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 55–60, Mahwah, New Jersey. Lawrence Erlbaum Associates.

Young, A., Hellawell, D., and Hay, D. (1987). Configural information in face perception. *Perception*, 16:747–759.

Cognitive Precursors to Science Comprehension

Kimberly G. Cottrell (kcottrel@odu.edu)
Danielle S. McNamara (dmcnamar@odu.edu)
Department of Psychology, Old Dominion University
Norfolk, VA 23529 USA

Abstract

This study examined the ability of cognitive factors (i.e., prior domain knowledge, reading ability, and metacognitive reading strategies) to predict students' comprehension of science texts and students' performance in an introductory psychology course. Both prior knowledge and reading ability reliably predicted comprehension of the science text (about sensory memory). Prior knowledge was the best predictor for exam performance. However, greater knowledge provided no benefit for students who did not use certain types of metacognitive reading strategies. Also, the tendency to use previewing strategies only benefited students if they possessed sufficient prior knowledge.

Introduction

What cognitive abilities are most important to a student entering a college level science course? If we assume that science course performance relies on factors related to science comprehension, then cognitive factors such as the students' level of reading ability, their prior knowledge of the domain area, and students' knowledge and use of metacognitive reading strategies should play key roles in students' course performance. Hence, the purpose of this research was to establish whether these cognitive factors were predictive of students' comprehension of science texts as well as their performance in a science course (in this case, introductory psychology).

There is no doubt that better readers better comprehend text (Perfetti, 1985) – because, of course, that is the underlying definition of reading skill. Skilled readers also tend to experience the reading process as more automatic and effortless than less skilled readers (Underwood, 1997). Skilled readers tend to make reading process decisions below the level of consciousness, particularly when reading familiar material. Thus, skilled readers unconsciously, or with very little conscious effort, understand the thoughts communicated through the texts and are reminded of the knowledge they have regarding the topic covered within a text (Underwood, 1997). Furthermore, skilled readers approach confusing sentences or passages by incorporating their prior domain knowledge to help them better understand the text (Collins, 1994). Thus, it would be expected that not only reading skill, but also prior knowledge would provide considerable benefits to science text comprehension and by consequence to students' course performance.

Researchers have established that prior domain knowledge has a strong effect on text comprehension and memory. Bransford and Johnson (1972) first established that prior knowledge improves readers' memory for written information. They showed that when readers were provided with a prior schema via a passage title or a picture, readers recalled twice as much from the passage compared to those who were not provided with prior schematic information. Essentially, the passage title activated the appropriate prior knowledge, or schema, that allows the reader to understand and thus remember the passage. Chiesi, Spilich, and Voss, (1979) demonstrated that readers with greater prior knowledge of baseball better understood and remembered a passage concerning baseball, regardless of the participants' age or reading ability. Further research has demonstrated that prior knowledge has a pronounced effect on comprehension of difficult expository texts, such as those found in science textbooks. Readers with greater prior knowledge exhibit superior comprehension and thus enhanced learning compared to those with less prior knowledge (Alexander, Kulikowich, & Schulze, 1994; Chiesi et al., 1979; McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996). Therefore, we can expect that prior knowledge will have a substantial effect on science course performance, perhaps more so than reading skill.

Whereas prior knowledge is certainly critical for successful text comprehension and course performance, students' metacognitive knowledge, such as their knowledge of metacognitive reading strategies, may also play an important role. Generally, metacognition refers to an individual's ability to think about thinking. More specifically, metacognition can be defined as an individuals' ability to self-monitor, self-assess, and self-evaluate. These processes help a learner determine why a process such as reading a science textbook is difficult, and then potentially overcome the difficulty.

Metacognition when applied to reading refers to the process of monitoring comprehension and the use of strategies to improve comprehension (Forget & Morgan, 1997). Reading strategies such as summarization, mental imagery, mnemonic imagery, question generation, answering self-generated questions, and look-backs have all been shown to enhance text comprehension (Pressley & Woloshyn, 1995). Chi and Bassok (1989) found that successful students tended to employ reading strategies such as generating elaborations and paraphrases, monitoring and creating statements, and producing self-explanations. In turn, these strategies enhanced their understanding of the

text material (see also, Chi, DeLeeuw, Chiu, & LaVandcher, 1994). Chi and Bassok also found that the students were more likely to generate explanations of the material covered within a text when they monitored and detected the points that they did not successfully understand. In addition, more successful students' self-explanations tended to include additional knowledge compared to less successful students who were more likely to simply paraphrase the text material.

Chi, Bassok, Lewis, Reimann, and Glaser (1989) found that successful students showed a tendency to use strategies such as explaining and justifying example-exercises in a science textbook to themselves whereas less successful students were not as likely to show this tendency. When the less successful students explained the exercise to themselves, they did not seem to connect their prior knowledge with the information covered within the science textbook. The study also found a tendency for the successful students to accurately monitor and detect when they understood a concept as well as when they did not fully comprehend or understand a concept. The less successful students did not show this tendency when attempting to detect comprehension failures of concepts covered within a science text.

Reading strategy instruction also improves reading comprehension. Bereiter and Bird (1985) found that when readers were taught strategies such as restatement, backtracking and problem solving, there was a significant increase in reading comprehension. McNamara and Scott (1999) found that when students in a college level science course were trained to use reading strategies to improve self-explanation, they had superior course grades compared to their counterparts who did not receive this training. Moreover, this training was more beneficial to students with less prior knowledge than for those who had greater prior knowledge. In essence, the training helped the students to overcome their knowledge deficits.

Collins (1994) also notes that metacognition affects an individual's ability to integrate prior knowledge with incoming novel information from a textbook, such as those textbooks used in science courses. As indicated earlier, prior knowledge assists an individual's ability to comprehend incoming information. Thus, an individual's metacognitive skills may assist him or her in incorporating their prior knowledge with the new information from the textbook.

In summary, the literature indicates that reading skill, prior knowledge, and the knowledge and use of metacognitive reading strategies are important tools in science comprehension. Thus, the current study examines whether reading ability, prior domain knowledge, and metacognitive reading strategies are related to course performance in an introductory psychology course at Old Dominion University. These factors were assessed at the beginning of a semester to determine their ability to predict students' comprehension of a science text and students' course performance. Our secondary goal was to reveal whether specific metacognitive reading strategies were more or less associated with reading ability, science text comprehension, and prior knowledge.

Reading ability was expected to facilitate readers' comprehension of text material and thus improve course performance. Additionally, it was expected that prior knowledge would have a profound effect on science course performance as well as comprehension of a science text. Those students who have prior knowledge of the domain of a text should perform better on comprehension measures of that material because they have the opportunity to incorporate that prior knowledge with the text material. However, knowledge and use of metacognitive reading strategies should enhance that ability by the reader strategically incorporating his or her prior knowledge with the novel material from the text. However, for those students who lack the adequate knowledge needed to comprehend difficult texts, such as science textbooks, the use of metacognitive reading skills may compensate for the lack of prior knowledge (see e.g., O'Reilly and McNamara, 2002). Metacognitive reading strategies may assist the learner to monitor their comprehension of the text material and thus actively attempt to understand the material.

Method

Participants

The sample consisted of 144 undergraduates enrolled in Introduction to Psychology at Old Dominion University. The participants included 57 males and 87 females with a mean age of 19 years. The majority of the participants were freshman (n=111). The remaining sample consisted of 21 sophomores, 8 juniors, and 4 seniors. They were given extra credits points in the psychology course of their choice for their participation.

Procedure

The experiment involved two sessions, which took place during the regularly scheduled class periods. In the first session, participants were invited to participate and given the Metacognitive Strategies Index (MSI) to complete at home. (There was no instruction of metacognitive reading strategies.) In the second session, participants were administered the prior knowledge test (19 min), the Nelson Denny reading test (15 min), and the sensory memory text with the comprehension questions (8 min). Students' grades were provided by the instructor at the end of the semester.

Materials

Metacognitive Reading Strategies Knowledge and use of metacognitive reading strategies was assessed using the Metacognitive Strategies Index (MSI; Forget, 1999). This was a 25-item multiple-choice questionnaire. Studies conducted by Forget in content areas at the high-school level found validity and test-retest reliability to be high (Forget, 1991; Forget & Morgan, 1997; Forget, 1999). The questions determine what the student does before, during, and after reading a text. For the purpose of this study, four sub-factors were examined (1) predicting and verifying (predicting the content and evaluating predictions and creating new ones), (2) previewing the text, purpose setting, and self-questioning, (3) drawing from background knowledge (activating and incorporating information from

background knowledge), and (4) summarizing the content. The reliability of the MSI computed by Cronbach's Alpha was .66. Reliability of sub-factor (1) predicting and verifying was .30, of sub-factor (2) previewing, purpose setting, and self-questioning was .38, of sub-factor (3) drawing from background knowledge was .46, and for sub-factor (4) summarizing and applying fix-up strategies was .29.

Demographics Demographics, motivation, effort, and education were assessed using a questionnaire consisting of 15 questions. The questions assess how much time and effort the students devote to the course as compared to how much time and effort they devote to other courses (i.e., "How many hours per week do you plan to devote to reading and studying for this course?"). The questions also determine how many science courses the participants have completed and how much they enjoy reading and learning scientific material as well as non-scientific material. Examples are: "How much do you enjoy reading?" and "How much do you enjoy learning information about science?"

Prior Knowledge Prior knowledge was assessed using an unpublished prerequisite knowledge test developed in collaboration with Linda Buyer at Governors State University in Illinois. The 48 multiple-choice questions were developed based on the concepts covered within three textbooks used for introductory psychology courses that were assumed known to the reader by the textbook authors. The questions included psychology specific questions (i.e., "Which person is most closely associated with the concept of the unconscious?"), general knowledge questions (i.e., "Which of the following is a logarithmic scale?"), and research methodology questions (i.e., "How is sample size related to the accuracy of population estimates derived from sample data?"). The questions were presented in random order. Reliability of the prior knowledge test computed by Cronbach's Alpha was .67.

Reading Ability Reading ability was evaluated using form G of the Nelson Denny adult reading comprehension test (Brown, Fishco, & Hanna, 1993). The measure consists of seven passages and 38 questions. The participants were instructed to read a passage and then answer the comprehension questions regarding that particular passage. They were permitted to look back on the passages to answer the comprehension questions. Reliability computed by Cronbach's Alpha was α=.88.

Science Comprehension The text consisted of a 307-word passage on the topic of sensory memory adapted from Lefton (pp. 195-196; 2000). The reading ease was 31.3 and the Flesch-Kincaid Grade level was 12. Twelve open-ended comprehension questions were used to measure comprehension of the passage. Six of the 12 questions were bridging questions, which require the reader to make inferences from two or more sentences in the text (i.e., "What was the dependent variable in Sperling's experiment? That is what did he measure") and the remaining 6 were text-based questions which require the reader to use only one sentence in order to successfully answer the question (i.e., "What is sensory memory?"). The students were allotted 8 minutes to read the passage and answer the questions, but were allowed to refer back to the passage to answer the questions. Therefore, performance on the questions assesses comprehension, but does not necessarily assess memory or learning. The participants were allotted 8 minutes to complete the exercise. Reliability of the science text comprehension questions computed by Cronbach's Alpha was .33.

Results

The alpha level was set at .05; hence, probability values are only reported for marginal results. As might be expected, knowledge and reading skill were highly correlated (R=.63), but were not reliably correlated with performance on the MSI. Text comprehension performance was measured in terms of proportion correct on the open-ended questions. Two raters scored the comprehension questions and discrepancies (12% of the scores) were resolved via discussion, yielding a final set of scores.

Demographics

A standard multiple regression was computed to determine whether variables such as the students' amount of effort in the class, motivation, enjoyment for leaning science and non-science material as well as the number of previous science courses taken by the student were associated with comprehension of the sensory memory text as well as course performance. For comprehension of the sensory memory passage, the overall regression model was significant, $F(15,121)=2.29$ accounting for 12% of the variance. Total points on the SAT in high school was the only significant predictor, $\beta=.35$, $sr_i^2=.09$. For average exam performance, the overall regression model was reliable, $F(15,121)=1.93$ accounting for 9% of the variance. The only reliable predictor was high school grade point average, $\beta=.33$, $sr_i^2=.08$. Hence, neither course effort nor reading enjoyment were significant predictors of performance in this study.

Predicting Comprehension and Exam Performance

Our first question regarded the ability of prior knowledge, reading ability, and the MSI to predict science text comprehension and course performance. Regression analyses were performed for each dependent measure including prior knowledge, reading ability, and metacognitive reading strategies as predictor variables.

Science Text Comprehension For science text comprehension, the overall regression model was reliable, $F(3,140)=30.70$, accounting for 38% of the variance. Performance on the MSI did not predict performance, whereas both prior knowledge, $F(1,140)=9.54$; $\beta=.27$, $sr_i^2=.04$, and reading ability, $F(1,140)=23.68$; $\beta=.41$, $sr_i^2=.10$, reliably predicted comprehension. High-knowledge

students scored significantly higher on the comprehension questions (M=49% correct) than low-knowledge students (M=33% correct); and similarly, skilled readers showed better comprehension (M=50% correct) than less-skilled readers (M=31% correct). Separate analyses did not reveal any interdependencies between prior knowledge and reading ability.

Average Exam Performance Average Exam performance in this course was based on the top five of six exams. Analyses included students who completed at least four of the six exams (n=136). In terms of average exam performance, the overall regression model was reliable, $F(3,132)=13.46$, accounting for 25% of the variance. Neither reading ability, $F(1,132)=2.38$, nor performance on the MSI, $F(1,132)=3.24$, reliably predicted exam performance. Prior knowledge, in contrast, accounted for 21% of the variance, $F(1,132)=30.85$; $\beta=.48$, $sr_i^2=.14$. Students with greater prior knowledge about concepts related to psychology when beginning the course scored significantly higher on the exams (M=0.80) than did low-knowledge students (M=0.72). In addition, the effect of knowledge on exam performance remained stable across exams, thus, prior knowledge affected performance equivalently across all of the exams.

Metacognitive Reading Strategies

Our second question regarded whether the sub-factors of the MSI differentially predicted reading skill, science text comprehension, and course performance. The sub-factors included (1) predicting and verifying, (2) previewing, purpose setting, and self-questioning, (3) drawing from background knowledge, and (4) summarizing and applying fix-up strategies.

Reading Ability None of the four metacognitive sub-factors reliably predicted scores on the Nelson Denny reading test.

Science Comprehension Performance on the science text comprehension questions was reliably predicted by sub-factor 3 (drawing from background knowledge), $F(1,140)=4.37$; $\beta=.21$, $sr_i^2=.03$. The students who were more likely to use prior knowledge while reading scored significantly higher on the comprehension of the science text (M=42%), than those students who were less likely to use knowledge (M=37%). This relationship remained significant when prior knowledge and reading ability were included in the regression equation. Although, one might expect that drawing on background knowledge would depend on the student's prior knowledge, this interaction was not reliable.

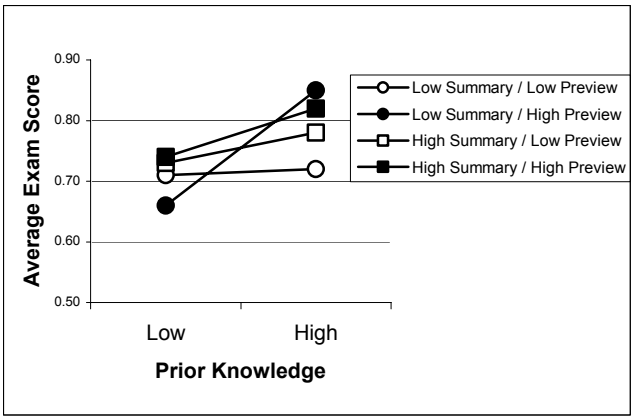


Figure 1. Exam Performance as a Function of Prior Knowledge, Summarization Reading Strategies, and Previewing Reading Strategies

Exam Performance Performance on exams was reliably predicted by sub-factor 2 (previewing, purpose setting, and self-questioning), $F(1,131)=4.16$; $\beta=.19$, $sr_i^2=.03$ (High M=0.77; Low M=0.73) and by sub-factor 4 (summarizing and applying fix-up strategies), $F(1,131)=4.33$; $\beta=.19$, $sr_i^2=.03$ (High M=0.77; Low M=0.73). However, these effects were not reliable when prior knowledge was also included in the regression model indicating that prior knowledge is more predictive of exam performance. Moreover, further analyses revealed that these factors were interdependent. An interaction between knowledge and previewing indicated that students who were more likely to use previewing strategies (i.e., sub-factor 2) only benefited from these strategies on the exams if they possessed sufficient knowledge, $F(1,128)=7.95$. In addition, there was a three-way interaction between knowledge, previewing, and summarization, $F(1,128)=4.65$ (see Figure 1). Students who were less likely to use previewing strategies but more likely to use summarization strategies showed a marginal effect of prior knowledge, $F(1,36)=3.31$, $p=.08$, Students who were more likely to use previewing strategies, regardless of the extent of their use of summarizing strategies significantly benefited from their prior knowledge on exams (High Summary $F(1,46)=14.33$; Low Summary $F(1,19)=21.36$). Finally, greater knowledge provided no benefit for students who did not use either type of reading strategies, $F<2$. Looking at the interaction from a different angle, the effects of strategy use were not reliable for low-knowledge students. In contrast for high-knowledge students, there was a reliable effect of previewing, $F(1,47)=7.84$, and marginal interaction of previewing and summarizing, $F(1,47)=3.61$, $p=.06$.

Discussion

The purpose of this study was to determine which cognitive abilities were precursors to science comprehension, and by consequence to students' performance in a science course (in this case, introductory psychology). Hence, we examined students' prior knowledge, reading ability, and metacognitive reading strategies at the beginning of a semester and related performance on these measures to students' ability to comprehend a science text (about sensory memory) and students' average exam scores. In addition, we determined whether specific metacognitive reading strategies differentially predicted reading ability, science text comprehension, and course performance.

As hypothesized, prior knowledge strongly predicted students' comprehension of the sensory memory passage as well as their exam performance. High-knowledge students showed a 16 percent advantage on comprehension questions in comparison to low-knowledge students. Similarly, students with greater prior knowledge about basic psychological concepts performed about 8 percent better on the course exams than low-knowledge students. Indeed, prior knowledge was the only reliable predictor of exam performance, accounting for 21 percent of the variance. Moreover, there was not a decline in the effect of prior knowledge throughout the semester. Hence, students did not overcome their knowledge deficits as the course proceeded.

Reading ability was a strong predictor of text comprehension, even more so than prior knowledge. The fact that reading ability did not predict exam performance could imply that the exams did not include information covered solely within textbooks rather was primarily based on information covered within the lectures; thus reading comprehension would not be necessary to succeed in the course. In addition, this would mean that our reading ability measure would be a poor predictor of lecture comprehension. Alternatively, the type of reading ability measure could be at fault. That is, perhaps the Nelson Denny reading test does not tap into the same processes involved in comprehending a course textbook. To contradict that argument, however, it was found that reading ability was the strongest predictor of comprehension of the science text, which was derived from an introductory psychology textbook.

In addition, it was found that neither the overall score nor the four metacognitive sub-factors predicted reading ability. This result indicates that the Nelson Denny reading test does not rely heavily on strategy use. This result may be expected because the Nelson Denny reading test primarily covers relatively familiar material, whereas metacognitive reading strategies may be most helpful for less familiar or difficult material.

Students' overall score on the Metacognitive Strategies Index (MSI) did not reliably predict either exam scores or science text comprehension. However, when the sub-factors were considered, it was found that *drawing from background knowledge* was predictive of science text comprehension. The influence of this type of reading strategy was positive regardless of students' prior knowledge and reading ability. This result further supports the importance of teaching both high- and low-knowledge students to integrate prior knowledge with new information when reading difficult texts such as science textbooks (e.g., McNamara & Scott, 1999).

The two sub-factors, *previewing, purpose setting and self-questioning* as well as *summarizing and applying fix-up strategies* predicted exam performance; however, this was dependent on the amount of knowledge the students' possessed. The students' use of previewing strategies was only beneficial on the exams when they possessed sufficient prior knowledge. The purpose of previewing is to activate knowledge schemas. These schemas presumably help the student to prepare for the learning process – just as a story title helps the reader understand a passage (Bransford & Johnson, 1972). However, without the necessary knowledge about the topic, previewing is of little utility.

In addition, it was found that students with low-knowledge did not benefit reliably from strategy use on exams (cf., O'Reilly & McNamara, 2002). Thus, in contrast to the results for text comprehension, it was found for exams that strategy use did not help to compensate for knowledge deficits. High-knowledge participants benefited from strategies, and most importantly, having more prior knowledge did not benefit students who did not use either of the metacognitive reading strategies. Hence, knowledge and strategy use are critically intertwined.

There are several limitations to this study. First, this study was correlational, and thus causal relationships cannot be assumed. Clearly, additional experimental studies are necessary to more completely understand these issues (cf., McNamara & Scott, 1999). Second, the sample was college students all enrolled in the same introductory psychology course and most of the students were freshmen in their second semester of college. Thus, the results of this study may not generalize to other populations. Additionally, this study may not adequately tap into science comprehension per se since the students are enrolled in a psychology course and the sensory memory text was derived from an introductory psychology textbook. The predictors of comprehension of passages within a psychology textbook may differ from those predictors of comprehension of a biology textbook or textbooks used in other hard sciences. Future research in this area could be to examine what factors predict comprehension of texts and course performance in other science courses such as biology or chemistry (e.g., see O'Reilly & McNamara, 2002). Finally, it may be beneficial to examine predictive factors of science comprehension for college students compared to younger students such as those in high school.

In conclusion, these results underline the notion that students should be taught to utilize strategies when reading texts, particularly those found in science courses. It is important, though, to understand which strategies may be more or less helpful under different circumstances. This study is a small step toward better understanding what these circumstances may be.

Acknowledgments

We would like to thank Linda Buyer who developed the prior knowledge test for this study. We would also like to thank the ODU Strategy Lab for helping us administer the tests to the students, and Erin McSherry for helping to score the comprehension questions. This project was partially supported by an NSF IERI grant (REC-0089271) to the second author. For further information about this study, contact Danielle McNamara at the University of Memphis Psychology Department.

References

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal, 31*, 313-337.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and Instruction, 2*, 131-156.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Brown, J., Fishco, V., & Hanna, G. (1993). Manual for Scoring and Interpretation, Forms G&H. Chicago, Illinois: Riverside Press.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 257-273.

Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: essays in honor of Robert Glaser* (pp.251-282). New Jersey: Erlbaun Associates.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glasser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M. T. H., DeLeeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Collins, N. D. (1994). *Metacognition and Reading to Learn*. Bloomington, IN: ERIC Clearinghouse on Reading English and Communication. (ED 376427).

Forget, M. A. (1991). The effects of a content area reading curriculum on senior high school students. Unpulished paper presented in partial fulfillment of the Certification in Reading. Old Dominion University.

Forget, M. A. (1999). Comparative effects of three methods of staff development in content area reading instruction on urban high school teachers. Unpublished doctoral dissertation. Old Dominion University.

Forget, M. A., & Morgan, R. F. (1997). Brain compatible strategies for improving student metacognition. *Reading Improvement, 34*, 161-175.

Lefton, L. A. (2000). *Psychology*. Allyn & Bacon. 195-196.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-288.

McNamara, D. S., Kintsch, E., Songer, N. B., Kintsch, W. (1996). Are good texts always better?: Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

O'Reilly, T. & McNamara, D. S. (2002). What's a science student to do? *Proceedings of the Twenty-fourth Annual Meeting of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

Perfetti, C. A. (1985). *Reading Ability.* NY: Oxford: University Press.

Pressley, M., & Woloshyn, V. (1995). *Cognitive Strategy Instruction that really Improves Children's Academic Performance.* MA: Brookline Books.

Underwood, T. (1997). On knowing what you know: Metacognition and the act of reading. *The Clearing House, 71*, 77.

The Role of Diagrams and Diagrammatic Affordances in Analogy

David Latch Craig (david.craig@arch.gatech.edu)
College of Architecture, Georgia Institute of Technology
Atlanta, GA 30332-0155 USA

Nancy J. Nersessian (nancyn@cc.gatech.edu)
College of Computing, Georgia Institute of Technology
Atlanta, GA 30332-0280 USA

Richard Catrambone (rc7@prism.gatech.edu)
School of Psychology, Georgia Institute of Technology
Atlanta, GA 30332-0170 USA

Abstract

We argue that problem solvers can, in certain cases, solve target problems by transforming perceptual simulations of solutions to analogous source problems. We further argue that source diagrams may facilitate the process, but only if they convey physical affordances consistent with the necessary transformations. We conducted an exploratory study in which participants were asked to solve a source and a target problem. We identified two properties of extemporaneously drawn source diagrams – view and configuration – that were highly correlated with the production of analogous solutions to the target problem. We speculated that view and configuration influenced the ease with which certain simulated transformations were performed. The results of two additional experiments in which the view and configuration of source diagrams were independently controlled further support the claim.

Introduction

In this paper we explore the functioning of diagrams in analogical problem solving. Specifically, we investigate how contextual aspects of diagrams – things ranging from depicted physical details to intrinsic properties like perspective, orientation and scale – might afford the kind of simulated physical transformations needed to convert a solution to one problem into a solution to another. In the next two sections we briefly outline our claims concerning diagrams, simulations and affordances, and how they might relate to analogy. In the remaining sections we present the findings of three experiments designed to both illustrate and test those claims.

Diagrams, Simulations and Affordances

One way external diagrams can function in problem solving is by scaffolding perceptual, or analog, simulations in the perceptual and motor cortices of the brain (Barsalou, 1999; Glenberg, 1997). Perceptual simulations have been found to facilitate spatial reasoning (e.g., Kosslyn, 1994) as well as various forms of conceptual reasoning (e.g., Barsalou, Solomon & Wu, 1999; Glenberg & Robertson, 2000; Stanfield & Zwaan, 2001; Fincher-Keifer, 2001). They could potentially benefit problem solving by facilitating the testing and general exploration of candidate solutions.

We argue that the way a diagram is drawn affects not only what is perceptually simulated but also how the resulting simulation can be perceptually transformed. A long history of findings, dating back to Cooper and Shepard's (1973) chronometric studies of mental rotation, support the basic premise that simulations are transformed through simulated motor activity. More recently, researchers have found that simulated transformations are motorically structured and constrained. The ease with which imagined body parts are mentally rotated, for example, parallels the ease with which those parts can be rotated in actuality (Parsons, 1987). In addition, concurrent motor activity consistent with simulated transformations of imagined objects tends to make those transformations faster and more accurate, while inconsistent activities produce interference (Wexler, Kosslyn & Berthoz, 1998). Generally speaking, simulated transformations appear to be constrained in the same way real interaction with the physical world is constrained. Insofar as contextual aspects of diagrams would help determine the physical properties of simulated objects (e.g., texture, shape, mass, etc.) and the context in which they are perceived (e.g., perspective, orientation, scale, etc.), those aspects act as transformational affordances by facilitating certain simulated transformations and inhibiting others.

A finding that illustrates the idea that diagrams convey transformational affordances comes from a study by Schwartz and Black (1996) in which people were shown a diagram of two gears meshed together, one larger than the other, and asked whether two marks, one on the circumference of each gear, would eventually line up if the gears were rotated. By comparing response times against the initial angular disparity of the marks, Schwartz and Black were able to identify different strategies used to complete the task, one of which appeared to be perceptually simulating the two gears rotating together. Ultimately, Schwartz and Black were able to constrain the strategy people used

by manipulating the gear diagrams. In particular they found that the simulated-rotation strategy was most likely to be used when the contacting surfaces of the gears were depicted as rough rather than smooth, as if roughness made it easier to imagine one gear driving the other. In this case a physical property depicted in the diagram appears to have affected the ease with which associated perceptual simulations were subsequently transformed.

Analogical Problem Solving

In analogical problem solving, problem solvers start with a solved "source" problem that is similar in some way to an unsolved "target" problem. If a problem solver is aware that the two are related, he or she will need to map the source problem onto the target, thus identifying which problem elements and constraints are identical, which are comparable, and which are irrelevant. Ideally, a mapping will be formed that allows the problem solver to transfer additional aspects of the source to the target, producing a target solution. Although most accounts of analogy are based on perceptually neutral representations (e.g., Gentner, 1983; Gick and Holyoak, 1983), we argue that people can, in certain situations, perceptually simulate source solutions and transform the simulations into solutions to target problems. Following the hypothesis presented in the previous section, we further argue that affordances associated with source diagrams might influence the likelihood that an analogical solution is produced by constraining what transformations can be executed.

Experiment 1

To explore how diagrams influence analogical problem solving we devised an open-ended experiment in which participants were given two superficially dissimilar but analogous problems and asked to 1) consider possible links between them, 2) list whatever similarities they found, and 3) try to solve them. The first problem was written to be easier than the second, the hope being that participants would solve it and thus have a source they could apply to the second problem. No independent variables were controlled. Instead, variations in solutions to the easier problem – in particular variations in the contextual aspects of spontaneously produced sketches – were analyzed after the fact. Correlations between various contextual properties and the production of analogous solutions to the harder problem were then sought.

The easier of the two problems – the one written to be a potential source for the harder problem – involved designing a door system for a laboratory that would give workers free access to the lab space while keeping the air outside the lab from contaminating the air inside. It was assumed that most participants would come up with a redundant-door solution, one that involved either two sets of doors on either side of a vestibule or a revolving door. The harder problem – the one written to be the target – involved designing a pole that suspended a device several feet off the side of a truck. The pole was described as sticking out in such a way that it ran into signposts on the side the of the road (Figure 1). The problem was to design the pole so that it could pass through signposts at a right angle. Ideally, if participants came up with a redundant-door solution to the door problem they would use it to come up with a redundant-pole solution to the pole problem. They might, for example, specify two poles, one that moved out of the way while the other stayed in place and vice versa.

To form an analogy between the door problem and the pole problem requires overcoming not only superficial differences (e.g., differences in objects and object attributes) but also a key structural difference in their respective perceptual contexts. In the door problem, passing through the door is natural; the problem is that it lets bad air in and good air out. In the pole problem, by contrast, passing one object through the other is not natural, and the problem is to make it so. Furthermore, the pole problem involves modifying the thing in motion, while the door problem involves modifying the thing being passed through. Thus, to map a simulated redundant-door solution onto the pole problem ultimately requires a shift in one's physical frame of reference. One must either 1) imagine that the sign post in the pole problem is the lab worker in the door problem, or 2) imagine that the lab boundary in the door problem is the pole in the pole problem. The latter means imagining an otherwise rooted lab boundary in motion, while the former means imagining an otherwise rooted sign post in motion.

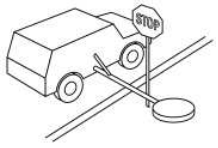


Figure 1 The pole problem: Participants are asked to design a pole that can pass through a signpost.

Transforming the motional context of a simulated redundant-door solution may not be easy. Such a shift might depend on what sort of transformational affordances are present, which might, in turn, depend on the contextual properties of an external diagram. We argue that a diagram of a redundant-door solution might, by scaffolding a perceptual simulation, facilitate the use of such a solution in solving the pole problem, but only if the contextual properties of the diagram afford the shift in motional context required to align the two problems.

Materials and Procedure

The experiment was administered in booklet form. The door problem was printed at the top of the first page

and the pole problem just below it. Instructions on the first page asked participants to write down as many similarities between the two problems as they could in 4 minutes. The second page of the booklet was divided vertically, the top containing instructions asking participants to write down a solution to the door problem, the bottom containing instructions asking participants to write down a solution to the pole problem. The instructions specified that they would have 5 minutes total. As participants were led through the booklet they were reminded to carefully read the instructions before starting each task. When they turned to the second page they were verbally told they could draw pictures if it helped them articulate their solutions.

Participants

Two hundred and nine participants were recruited from an introductory psychology class at Georgia Tech to participate in exchange for class credit. The experiment was administered in one large group during a regularly scheduled class session.

Results and Discussion

We first analyzed similarities participants reported prior to solving the problems. Most participants reported superficial similarities, such as that both problems involved engineers. A few also reported highly abstract similarities, such as that both problems involved an obstacle that prevented a device from working. More interestingly, some participants reported that both problems involved something passing through a solid barrier. Although the requirement that something pass through a barrier is clearly stated in the pole problem, it is not stated at all in the door problem. The objective in the door problem is, in fact, opposite that of the pole problem: Something with a penetrable (as opposed to solid) boundary needs to be redesigned to prevent something from getting through (as opposed to allow something through). Despite the implicit nature of the pass-through similarity, 60 of the 209 participants (29%) reported it.

We next analyzed solutions produced for the two problems. Solutions to the door problem were classified as either redundant-door solutions or non-redundant-door solutions, the former being those that included one or more of the following: 1) a verbal reference to two doorways, 2) a verbal reference to an airlock, 3) a verbal reference to a revolving door, 4) a diagram showing two doorways, 5) a diagram showing an airlock, or 6) a diagram showing a revolving door. As expected most participants (184, or 88%) produced some kind of redundant-door solution.

Fewer students, by contrast, were successful in solving the pole problem. Solutions to the pole problem were first classified as either analogous to the door problem or non-analogous. Analogous solutions were those that made use of redundancy. Specifically, a solution was deemed analogous if one part remained in place while another part moved out of the way and vice versa. Such solutions included those with multiple latches (with one latch opening at a time), multiple poles (with one pole retracting at a time), or rotating devices (with one end swinging out of the way as the other end swung into place). Of the 209 participants, 33 (16%) produced analogous solutions to the pole problem. An example of an analogous solution is shown alongside a non-analogous solution in Figure 2

Not surprisingly, participants were more likely to generate an analogous solution to the pole problem if they generated a redundant-door solution to the door problem. Of the 184 participants who generated redundant-door solutions, 31 (17%) produced analogous solutions to the pole problem, while only 2 of the other 25 participants (8%) produced them. Analogous solutions to the pole problem were also correlated with the reporting of pass-through similarities. Of the 60 participants who reported pass-through similarities, 17 (28%) produced analogous solutions to the pole problem, compared to 16 of the 149 (11%) who did not.

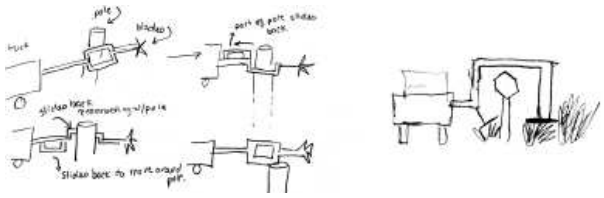


Figure 2 A solution to the pole problem that is analogous to a redundant-door solution to the door problem (left) and one that is not (right).

The remaining analyses concern the diagrams participants drew to illustrate their solutions to the door problem. Of the 184 participants who produced redundant-door solutions, 131 drew at least one diagram. Diagrams alone were not correlated with analogous solutions to the pole problem. Of the 131 who drew diagrams, 22 (17%) produced analogous solutions to the pole problem, while 9 of the remaining 53 participants (17%) also produced them. This is not inconsistent with the argument made earlier about the role of diagrams in problem solving. Of interest is not whether diagrams in general help but whether certain types of diagrams are more highly correlated with the production of analogous solution than others.

To roughly classify diagrams according to transformational affordances we looked at two diagram properties: view and configuration. View was coded as either plan (viewed from above), elevation (viewed from the side), perspective, or ambiguous (either plan or elevation). Configuration was more varied. After reviewing all redundant-door diagrams, 17 distinct configuration types were identified based on the spaces that were depicted and their organization. From these 17 types, two higher-level categories were defined: 1) single-space diagrams, or those in which the only space

depicted was the space between the redundant doors and 2) multiple-space diagrams, or those in which additional spaces were depicted. A space, in this case, was defined as any convex area bounded by at least three walls. An example of each type is shown in Figure 3.

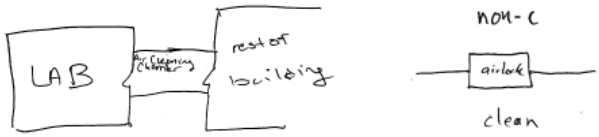


Figure 3 A multiple-space diagram of a redundant-door solution (left) and a single-space diagram (right).

Participants who drew single-space diagrams were significantly more likely to produce analogous solutions to the pole problem than those who drew multiple-space diagrams. Of the 131 participants drawing redundant-door diagrams 26 drew single-space diagrams, and of those 12 (46%) produced analogous solutions to the pole problem. By contrast, only 10 of the 105 (10%) participants who drew multiple-space diagrams produced analogous solutions. The percentage of participants who produced analogous solutions to the pole problem in each of the two main configuration types is listed in Table 1.

Participants who drew diagrams in plan were also much more likely to produce analogous solutions to the pole problem than those who drew diagrams from other views. Of the 131 participants who drew redundant-door diagrams, 58 drew diagrams in plan, 14 (24%) of whom went on to produce an analogous solution to the pole problem. Of the 73 who drew redundant-door diagrams from other views, only 8 (11%) produced analogous solutions. The percentage of participants who produced analogous solutions to the pole problem in each view is listed in Table 2.

Table 1 Number of redundant-door diagrams drawn in each configuration type, and the percentage of those followed by an analogous solution to the pole problem.

	N	Analogous solutions
Single-space	26	46%
Multiple-space	105	10%
N	131	17%

Table 2 Number of redundant-door diagrams drawn in each view, and the percentage of those followed by an analogous solution to the pole problem.

	N	Analogous solutions
Plan	58	24%
Elevation	42	10%
Perspective	22	9%
Ambiguous	9	22%
N	131	17%

There are at least two explanations for why participants were less likely to produce analogous solutions to the pole problem when drawing multiple-space diagrams than when drawing single-space diagrams. One is that additional spaces meant that there were additional unalignable features in the source that could have interfered with a successful mapping. Although this possibility is hard to assess, it should be noted that there were a number of other randomly distributed unalignable features in the diagrams that could have countervailed those associated with multiple spaces.

A second explanation, and one that is more in line with our original hypothesis, is that additional spaces made it more difficult to transform a perceptual simulation of a redundant-door solution into a perceptual simulation of a redundant-pole solution. This explanation rests on three assumptions: first, that the diagrams scaffolded simulations of physical objects with particular transformational affordances; second, that using a redundant-door solution to solve the pole problem required imagining the door system in motion; and third, that a simulated door system might have been rooted via a kinesthetic sense of inertia that would make it difficult to imagine motion. If so, whether a redundant-door diagram facilitated the production of an analogy would have depended on the diagram's affordances. Specifically, the depiction of additional spaces could have caused the door system to seem more physically encumbered and hence harder to simulate in motion as required for a successful mapping.

The fact that view was also correlated with the production of analogous solutions to the pole problem further supports the idea that diagrams both scaffolded and constrained perceptual simulations. Participants, for example, would have likely been able to visualize doors swinging open more easily in plan than in other views (the motion being orthogonal to such a view), making it easier to simulate the actions needed to solve the pole problem. In addition, plan-view simulations may have been more easily transformed because they were not constrained by gravitational affordances, gravity being orthogonal to spatial relations depicted in plan view (Franklin and Tversky, 1990; Rock, 1973). The idea that view and configuration may have influenced perceptually simulated transformations is, of course, speculative. The next two experiments attempt to provide more support for the claim.

Experiment 2

One of the findings from Experiment 1 was that the configuration of source diagrams was correlated with the production of analogous solutions to a target problem. We argued that the diagrams scaffolded perceptual simulations, which could have then been transformed to fit the physical context of the target problem if the diagrams afforded those transformations.

Although Experiment 1 helped illustrate this argument, the analysis was primarily post hoc. The experiment discussed in this section is designed to test the claim in a more controlled way.

A two-condition variation of Experiment 1 was designed. Participants in both conditions were given the door problem with a redundant-door solution already specified and a diagram illustrating it. They were then given the pole problem and asked to solve it, along with the hint that the solution to the door problem might help them. The independent variable was the type of diagram shown with the door problem, while the dependent variable was the type of solution participants produced for the pole problem.

In one condition (the afforded condition) participants were given a redundant-door diagram showing a door vestibule bisecting a wall bounding the lab. In the other condition (the unafforded condition) participants were given a diagram showing the same door vestibule abutting the wall (Figure 4). The number and type of elements were the same in both diagrams, ensuring that differences in performance could not be attributed to differences in the number or type of unalignable objects. Although both diagrams are single-space configurations according to the coding scheme used in Experiment 1, they differ in their physical affordances, particularly in how the vestibule is perceived in relation to the wall. In the afforded condition, the wall and the vestibule are meant to be perceived as overlapping, following the Gestalt law of continuation. In the unafforded condition, by contrast, the vestibule is meant to be seen as resting up against, or attached to, the wall. The diagram in the unafforded condition should thus be harder to imagine moving because it is encumbered by (or anchored to) the lab space, in turn making it harder to align with the pole problem. Following this reasoning, we predicted that participants in the unafforded condition would be less likely to produce an analogous solution to the pole problem than those in the afforded condition.

Figure 4 Diagrams used in the afforded condition (left) and the unafforded condition (right) of Experiment 2.

Materials and Procedure

The door problem and the pole problem used in Experiment 1 were printed on a single sheet of paper. Just below the door problem was written, "Have workers enter a vestibule space before entering the lab," along with one of the two diagrams shown in Figure 4, depending on the condition. Instructions printed at the top of the page and just below the two problems asked participants to carefully read them and write down a solution to the second one. Participants were given 7 minutes to complete the task.

Participants

Twenty-eight students enrolled in undergraduate psychology classes at Georgia Tech participated in groups of 2 to 6 each, 14 in the afforded condition and 14 in the unafforded condition. All received class credit for participating.

Results and Discussion

Solutions to the pole problem were categorized as either analogous or non-analogous to the door problem using the criteria established in Experiment 1. Of the 14 participants in the afforded condition, 10 (71%) produced analogous solutions, compared to only 5 of 14 participants (36%) in the unafforded condition. As predicted, configuration was a significant predictor of whether participants produced analogous solutions (χ^2=3.82, p<.05). The results are shown in Table 3.

Table 3 Participants producing analogous solutions to the pole problem in Experiment 2.

	Analogous solutions	N
Afforded diagram	71%	14
Unafforded diagram	36%	14

Experiment 3

Another notable finding in Experiment 1 was that participants who drew plan diagrams were more likely than those who drew diagrams from other views to produce analogous solutions to the pole problem. We argued that it was easier to perceptually simulate doors opening in plan and hence easier to simulate the action required to solve the pole problem. We also argued that it might be easier to transform a simulation of a redundant-door solution if the simulation was not perceptually structured in relation to gravity, or, in other words, if all spatial relations were orthogonal to gravity. To test this claim, we repeated Experiment 2 with two new redundant-door diagrams: one drawn from the side (unafforded condition) and one drawn from above (afforded condition) (Figure 5). Consistent with the arguments put forth in Experiment 1, we predicted that participants in the afforded condition would be more likely to produce an analogous solution to the pole problem.

Figure 5 Diagrams used in the afforded condition (left) and unafforded condition (right) of Experiment 3.

Materials and Procedure

The same materials and procedure used in Experiment 2 were used except that the diagrams accompanying the redundant-door solution were either elevation or plan diagrams, depending on the condition (Figure 5).

Participants

Twenty-two students enrolled in undergraduate psychology classes at Georgia Tech participated in groups of 2 to 6 each, 11 in the afforded condition and 11 in the unafforded condition. All received class credit for participating.

Results and Discussion

Solutions to the pole problem were categorized as either analogous or non-analogous to the door problem, using the criteria established in Experiment 1. Of the 11 participants in the afforded condition, 7 (64%) produced analogous solutions, compared to only 2 of the 11 participants (18%) in the unafforded condition. Consistent with our prediction, diagram view was thus a significant predictor of whether participants produced analogous solutions (χ^2=5.43, p<.05). The results are shown in Table 4.

Table 4 Participants producing analogous solutions to the pole problem in Experiment 3.

	Analogous solutions	N
Afforded diagram	64%	11
Unafforded diagram	18%	11

Conclusions

The studies reported here begin to shed light on what might make a diagram useful for constructing an analogy. They results strongly suggest that aspects of diagrams like view and configuration can influence the ease with which diagrammed solutions can be used to solve analogous problems, possibly by regulating simulated transformations. The studies also lend support to the more general idea that analogies can be constructed via perceptual simulations, as opposed to predicate-based, or otherwise perceptually neutral, representations. And finally, although just a start, the results reported here help illustrate an expanded role for drawings as cognitive tools. Drawings might now be seen not only as a means for recording ideas for future reference but also as a means for exploring the transformational affordances of problem spaces in search for those that will ultimately lead to more promising solution paths. Problem solvers might, from this point of view, actually learn to manipulate problem spaces via diagrammatic affordances just as they might learn to navigate problem spaces using conventional reasoning strategies.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.

Barsalou, L. W., Solomon, K. O., & Wu, L. L. (1999). Perceptual simulation in conceptual tasks. In M. K. Hiraga, C. Sinha, & S. Wilcox (Eds.), *Cultural, typological, and psychological perspectives in cognitive linguistics: The proceedings of the 4th conference of the International Cognitive Linguistics Association*, (209-228). Amsterdam: John Benjamins.

Cooper, L. A., & Shepard, R. N. (1973). The time required to prepare for a rotated stimulus. *Memory & Cognition*, 1, 246-250.

Fincher-Kiefer, R. (2001). Perceptual components of situation models. *Memory & Cognition*, 29, 336-343.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General*, 119, 63-76.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1-38.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20, 1-19.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, 43, 379-401.

Kosslyn, S. M. (1994) *Image and Brain*. Cambridge, MA: MIT Press.

Parsons, L. M. (1987). Imagined spatial transformation of one's body. *Journal of Experimental Psychology: General*, 116, 172-191.

Rock, I. (1973). *Orientation and Form*. New York: Academic Press.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, 30, 154-219.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 121, 153-156.

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68, 77-94.

A Classification of Cognitive Agents

Mehdi Dastani (mehdi@cs.uu.nl)
Institute of Information and Computer Sciences
P.O.Box 80.089
3508 TB Utrecht, The Netherlands

Leendert van der Torre (torre@cs.vu.nl)
Department of Artificial Intelligence, Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherlands

Abstract

In this paper we discuss a generic component of a cognitive agent architecture that merges beliefs, obligations, intentions and desires into goals. The output of belief, obligation, intention and desire components may conflict and the way the conflicts are resolved determines the type of the agent. For component based cognitive agents, we introduce an alternative classification of agent types based on the order of output generation among components. This ordering determines the type of agents. Given four components, there are 24 distinct total orders and 144 distinct partial orders of output generation. These orders of output generation provide the space of possible types for the suggested component based cognitive agents. Some of these agent types correspond to well-known agent types such as realistic, social, and selfish, but most of them are new characterizing specific types of cognitive agents.

Introduction

Imagine an agent who is obliged to settle his debt, desires to go on holiday, and intends to attend a conference. Suppose that he believes he can only afford to finance one of these activities and decides to pay his checks to settle his debt. Unfortunately, our agent does not earn much money and is in the habit of buying expensive books. Therefore, he runs again into debt after a short while. Despite the fact that he still has the same obligation, desire, and intention and believes that he can only afford to finance one of these activities, he decides this time to attend the conference. Directly after this decision, he hears that the conference is cancelled and he receives a telephone call from his mother telling him that she is willing to pay his checks for this time. The agent is now happy and decides to go on holiday. Our agent has a friend who has the same obligation, desire, and intention, and likewise believes that he can only afford to finance one of these activities. In contrast to our agent, this friend decides to go on holiday. However, he is late with arranging his holiday; all travel agencies are sold out. Therefore, he decides to attend the conference. In a different situation where these two agents are obliged to visit their mothers, desire to go to cinema, and believe they cannot do both simultaneously, the first agent decides to visit his mother while his friend goes to cinema. Yet in another situation where these agents intend to clean up their houses, are obliged to help their friends, and believe they cannot do both, they decide to clean up their houses. Although these agents behave differently, each of them seems to follow a certain behavior pattern under different situations. The first agent seems to be more sensitive to his intentions and obligations than to his desires while the second agent seems to prefer his desires more than his intentions and obligations. Moreover, the first agent seems to be indifferent towards his intentions and obligations while the second agent seems to prefer his intentions above his obligations. These characteristics and principles that govern agent's actions and behavior determine the type of cognitive agents and can be used as the basis for a classification of cognitive agents.

We are motivated by the studies of cognitive agents where the behavior of an agent is defined in terms of rational balance between its mental attitudes [1, 9, 5]. A classification of cognitive agent types specifies possible ways to define the rational balance. Beside the scientific need to study possible definitions of rational balance in a systematic way, a classification of cognitive agent types is important for many applications where it is impossible to specify agent behavior in specific and usually unknown situations. In such applications, it is important to specify the behavior of agents in strategic terms and by means of types of behavior.

In [2] we investigate the design and implementation issues of generic component-based cognitive agents. In the present paper, we propose an alternative classification of cognitive agent types. There has been many formal and informal studies proposing agent types [1, 8, 4]. In these studies, there is a trade-off between the space of possible agent types and their precise and formal definitions. In particular, informal studies provide a rich space of possible types of cognitive agents and ignore their precise definitions, while formal studies provide precise definition of agent types but ignore the richness of the space of possible types. The proposed classification of cognitive agent types in this paper is formal and in terms of a generic component based architecture.

This classification is systematic and provides a large space of possible types for cognitive agents. Some of these agent types such as realistic, social, and selfish are well-known. However, most of these agent types are new and characterize specific types of behavior.

The layout of this paper is as follows. First, we discuss different ways of classifying agent types. Since our classification is based on generic component based agent architecture, we briefly discuss this architecture and explain some of its properties that are relevant for the agent type classification. Possible agent types within this architecture are discussed. An example of a conflict situation is formalized and it is shown how different agent types behave differently in this situation. Finally, we conclude the paper and indicate future research directions.

Classification based on Agent Architecture

Various frameworks with corresponding type classifications for cognitive agents are proposed [9, 5, 3]. Considering different phases in agent oriented software development process such as analysis, design, and implementation phases, most proposed cognitive agent frameworks with corresponding type classifications are provided for the analysis phase. For example, Rao and Georgeff's BDI framework with realism and commitment strategies as agents types [9] have been developed as formal specification tools for the analysis phase. In this framework, the single minded agent type is thought to be the one which maintains its commitments until either it believes it has fulfilled its commitments or it does not believe it can ever fulfill its commitments.

Although these formal tools and concepts are very useful to specify various types of cognitive agents, they are specifically developed for the analysis phase which makes them too abstract for other phases. In fact, to design and to implement various types of cognitive agents, we need to define agent types in terms of tools and concepts available at the design and the implementation phases such that they can be translated into agent architectures and agent implementations. A closer look at the specification formalisms such as Rao and Georgeff's BDICTL formalism shows that the space of theoretically possible cognitive agent types is determined by the expressive power of that formalism. Obviously, other phases of agent development process restrict and narrow down the space of possible agent types since available concepts and tools at those phases should satisfy conditions such as realizability and computability. This implies that each agent architecture allows only a subset of possible agent types that can be specified at the analysis phase. Therefore, it is essential for each agent architecture to indicate which types of agents can be designed in that architecture. The classification of cognitive agent types in this paper is proposed for the design phase and it is thus in terms of agent architecture.

Agent Architecture

In general, agent architectures are defined in terms of knowledge representation (i.e. data) and reasoning mechanism (i.e. control). The agent type classification, which we introduce in the next section, is defined in terms of properties of generic component based architecture called BOID (BOID stands for Beliefs, Obligations, Intention, and Desire). Therefore, we first briefly explain this architecture, which can be seen as a black box with observations as input and intended actions as output. The architecture and the logic of BOID are discussed in more detail elsewhere [2].

A BOID agent observes the environment and reacts to it by means of detectors and effectors, respectively. Each component in the BOID architecture is a process having an input and output behavior. For this reason and to model the input/output behavior of each component, the components are abstracted as a rule-based systems that contains a set of defeasible rules. As these components output mental attitude only for certain inputs, they represent *conditional* mental attitudes. In the BOID architecture two modules are distinguished: the goal generation module and the plan generation module. The goal generation module generates goals based on beliefs, desires, intentions and obligations, and the plan generation module generates sequences of actions based on these goals. In the rest of this paper, we focus only on the goal generation module since the presented classification of the agent types is defined in terms of rational balance between agent's mental attitudes. Possible classification of agent types that can be defined in terms of the plan generation module or in terms of the interaction between the goal or the plan generation modules are out of the scope of this paper.

The BOID architecture differs from the Procedural Reasoning System (PRS) [7], which is developed within the BDI (Beliefs, Desires, and Intention) framework, in several aspects. The first difference is that BOID extends PRS with obligations as an additional component. One reason for this extension is to incorporate elements of the social level, i.e. social commitments, to formalize for example social agents and social rationality. The second difference is related to the conditional nature of mental attitudes in BOID. In fact, each mental attitude is abstracted as a rule-based system containing defeasible rules. This is in contrast with the representation of mental attitudes in PRS which are sets or lists of formula. The third difference is that the BOID components, which represent mental attitudes, are processes having their own control mechanism. Thus, in contrast to the central control mechanism in PRS, in BOID there are two levels of controls. A central control

mechanism at the agent level coordinates activities among components. The control mechanism at the component level determines how and which output is generated by each component when it receives input. Finally, the goals in BOID are generated by the interactions between agent's mental attitudes in contrast to the PRS where goals are given beforehand and become selected by the central control mechanism.

As noticed, each component can be abstracted as a rule-based system specified by propositional logical formulas, in the form of defeasible rules represented as $a \hookrightarrow b$. The reading of a rule depends on the component in which it occurs. For example, a rule in the obligation component, represented as $a \overset{\text{O}}{\hookrightarrow} b$, should be read as follows: if a is derived as a goal and it is not inconsistent to derive b, then b is obliged to be a goal. The input and the output of components are represented by sets of logical formulas, closed under logical consequence. Following Thomason [10] these are called *extensions*. The logic that specifies extensions is based on prioritized default logic that takes an ordering function ρ as parameter. This function constraints the order of derivation steps for different components and characterizes the type of the agent. We first briefly discuss the BOID conflict resolution mechanism and then explain how the ordering function can be used to define various agent types.

Conflict Resolution Mechanism

In the BOID architecture, goals are generated by a calculation mechanism. The calculation starts with a set of observations Obs, which cannot be overridden, and initial sets of default rules for the other components: B, O, I, D. Moreover, it assumes an ordering function ρ on the rules of the different components. The procedure then determines a sequence of sets of extensions $S_0, S_1, \ldots$. The first element in the sequence is the set of observations: $S_0 = \{Obs\}$. A set of extensions S_{i+1} is calculated from a set of extensions S_i by checking for each extension E in S_i whether there are rules that can extend the extension. There can be none, in which case nothing happens. Otherwise each of the consequents of the applicable rules with highest ρ-value are added to the extension separately, to form distinct extensions in S_{i+1}. The operator $Th(S)$ refers to the logical closure of S, and the syntactic operation $Lit(b)$ extracts the set of literals from a conjunction of literals b. In practice not the whole set of extensions is calculated, but only those that are calculated before the agent runs out of resources.

Definition 1 *A tuple $\Delta = \langle Obs, B, O, I, D, \rho\rangle$ is called a BOID theory. Let L be a propositional logic, and an extension E be a set of L literals (an atom or the negation of an atom). We say that:*

- *a rule $(a \hookrightarrow b)$ is strictly applicable to an extension E, iff $a \in Th(E)$, $b \notin Th(E)$ and $\neg b \notin Th(E)$;*
- $\max(E, \Delta) \subseteq B \cup O \cup I \cup D$ *is the set of rules $(a \hookrightarrow b) \in \max(E, \Delta)$ strictly applicable to E such that there does not exists a $(c \hookrightarrow d) \in B \cup O \cup I \cup D$ strictly applicable to E with $\rho(c \hookrightarrow d) > \rho(a \hookrightarrow b)$;*
- *$E \subseteq L$ is an extension for Δ iff $E \in S_n$ and $S_n = S_{n+1}$ for the procedure in Figure 1.*

```
i := 0; S_i := {Obs};
repeat
  S_{i+1} := ∅;
  for all E ∈ S_i do
    if exists (a ↪ b) ∈ B ∪ O ∪ I ∪ D strictly
    applicable to E then
      for all (a ↪ b) ∈ max(E, Δ) do
        S_{i+1} := S_{i+1} ∪ { E ∪ Lit(w)};
      end for
    else
      S_{i+1} := S_{i+1} ∪ {E};
    end if
  end for
  i:=i+1;
until S_i = S_{i-1};
```

Figure 1: Procedure to calculate extensions

In our model, ρ can assign values to the rules, such that all rules from one component receive either larger or smaller values than the rules from another component. This implies that the rules from one components are applied before the rules from another component can be applied. This is the basis of our idea to define agent types. Of course, in many practical applications ρ must be specified further. For example, an agent may prefer some of his O rules to some of his D rules while conversely preferring some other D rules to some other O rules. However, this does not mean that our basic idea has to be dropped. It just means that the number of components has to be further specified and the ρ function has to be defined accordingly. Each component can thus be subdivided in a number of subcomponents such that the ρ can describe the preference of the rules accordingly. Here we do not further describe this division since it is not important for the general idea of agent type classification that we present in this paper.

The parameter ρ may assign unique values to the rules of all components. In such a case, the BOID calculation scheme can apply in each iteration loop only one rule, which implies that the BOID calculation scheme generates only one extension. However, ρ may also assign identical integers to different rules. In this case, ρ imposes a partial ordering among the rules. For such a ρ, the above BOID calculation scheme can apply more than one rule in each iteration loop, which implies that the BOID calculation

scheme may generate a set of extensions. For example, consider a scenario in which an agent believes that he is in a non-smoking area (i.e. $\top \overset{\mathrm{B}}{\hookrightarrow} nsa$). He intends to smoke (i.e. $\top \overset{\mathrm{I}}{\hookrightarrow} s$), but he intends not to smoke when he is in a non-smoking area (i.e. $nsa \overset{\mathrm{I}}{\hookrightarrow} \neg s$). Define ρ as follows:

$$\rho(\top \overset{\mathrm{B}}{\hookrightarrow} nsa) > \rho(nsa \overset{\mathrm{I}}{\hookrightarrow} \neg s) > \rho(\top \overset{\mathrm{I}}{\hookrightarrow} s)$$

For this ρ, the BOID calculation scheme as defined in Definition 1 generates one single extension which is: $\{nsa, \neg s\}$.
Now, suppose ρ is defined as follows:

$$\rho(\top \overset{\mathrm{B}}{\hookrightarrow} nsa) > \rho(nsa \overset{\mathrm{I}}{\hookrightarrow} \neg s) = \rho(\top \overset{\mathrm{I}}{\hookrightarrow} s)$$

This ρ does assign identical integers to the intention rules and the BOID calculation scheme generates the following two extensions: $\{nsa, \neg s\}$ and $\{nsa, s\}$.

Agent Types

Given the presentation of mental attitudes and the BOID calculation scheme, we investigate which type of interactions between mental attitudes can arise within the BOID architecture and how these interactions can be classified. In principle, there are fifteen types of conflicts that can occur between the mentioned four mental attitudes [2]. These conflicts can be solved in different ways. We explain how different ways of resolving conflicts can be modelled by restricting the order of rule application in the BOID calculation scheme. We argue that these restrictions specify different types of the BOID agent and introduce a classification of the types for the BOID agents. Finally, some examples of BOID types and their solutions to one and the same conflict situation is presented.

Conflict resolution and agent types

One of the main tasks of deliberative agents is to solve possible conflicts among their mental attitudes. In principle, there are fifteen different types of conflicts that may arise either within each class or between classes. Dependent on the exact interpretation of these classes, some of the conflict types may be more interesting or important than others. We distinguish two general types of conflicts: internal and external conflicts. *Internal conflicts* are caused within each component while *external conflicts* are caused between them. Internal conflicts can be distinguished into four unary subtypes (B ; O ; I ; D). External conflicts can be distinguished into six binary conflict subtypes (BO ; BI ; BD ; OI; OD ; ID), and four ternary conflict types (BOI; BOD; BID; OID) and one quadruplicate conflict type (BOID). An example of the BOID external conflict type is the following situation: *The agent intends to go to a conference. It is obligatory for the agent not to spend too much money for the conference. In particular, either the agent should pay for a cheap flight ticket and stay in a better hotel, or the agent should pay for an expensive flight ticket and stay in a budget hotel. The agent desires to stay in a better hotel. But, he believes that the secretary has booked an expensive flight ticket for him.* More examples of these conflicts are presented in [2].

A conflict resolution type, which characterizes an agent type, is considered here as an order of overruling. Given four components in the goal generation module of the BOID architecture, there are 24 possible orders of overruling. In this paper, we only consider those orders according to which the belief component overrules any other component. This reduces the number of possible overruling orders to 6. Some examples of conflict resolution with beliefs are as follows. A conflict between a belief and an intention means that an intended action can no longer be executed due to the changing environment. Beliefs therefore overrule the intention, which is retracted. Any derived consequences of this intention are retracted too. Of course, one may allow intentions to overrule beliefs, but this results in unrealistic behavior. Conflicts between beliefs and obligations or desires need to be resolved as well. As observed by Thomason [10], the beliefs must override the desires or otherwise there is wishful thinking. Moreover, a conflict between an intention and an obligation or desire means that you now should or want to do something else than you intended before. Here intentions override the latter because it is exactly this property for which intentions have been introduced: to bring stability [1]. Only in a call for intention reconsideration such conflicts may be resolved otherwise. For example, if I intend to go to cinema but I am obliged to visit my mother, then I go to cinema unless I reconsider my intentions.

Using the order of string letters as the overruling order and thus as representing the agent type, a realistic agent can have any of the following six specific agent types, i.e. BOID, BODI, BDIO, BDOI, BIOD, and BIDO. These specific agent types are not known in the literature and we do not have any name for them. Note that we overloaded the name BOID in this way, because it becomes a specific type of agent as well as the general name for the agent architecture. These six specific agent types, in which beliefs override all other components, can be represented as a constraint on the ρ function resulting in the well-known agent type, called *realistic*.

Definition 2 *Realistic agent type is a constraint on the ρ function formulated as follows:*
$\forall r_b \in B, r_o \in O, r_i \in I, r_d \in D$
$(\rho(r_b) > \rho(r_o) \ \wedge \ \rho(r_b) > \rho(r_i) \ \wedge \ \rho(r_b) > \rho(r_d))$
or simply

$$B \succ O \ \wedge \ B \succ I \ \wedge \ B \succ D$$

Now that we have a specific ρ function that characterizes realistic BOID types, we indicate how the extension is calculated. Following definition 1, a realistic BOID agent starts with the observations and calculates belief extensions by iteratively applying belief rules. When no belief rule is applicable anymore, then either the O, the I, or the D component is chosen from which one applicable rule is selected and applied. When a rule from a chosen component is applied successfully, the belief component is attended again and belief rules are applied. If there is no rule from the chosen component applicable, then another component is chosen again. If there is no rule from any of the components applicable, then the process terminates – a fixed point is reached – and extensions are calculated.

Other agent types can be specified as constraints on the ρ function as well. Since we consider in this paper only realistic agent types, we limit ourselves to those agent types that are subtypes of realistic agent types. Some of well-known agent types can now be represented as follows.

BIDO, BOID, and BIOD are called *stable*, because intentions overrule desires, i.e.

$$B \succ O \ \wedge \ B \succ I \ \wedge \ B \succ D \ \wedge \ I \succ D$$

BDIO, BIDO, and BDOI are called *selfish*, because desires overrule intentions, i.e.

$$B \succ O \ \wedge \ B \succ I \ \wedge \ B \succ D \ \wedge \ D \succ O$$

BOID, BIOD, and BODI are called *social*, because obligations overrule desires, i.e.

$$B \succ O \ \wedge \ B \succ I \ \wedge \ B \succ D \wedge \ O \succ D$$

The six specific realistic agent types mentioned earlier are subtypes of these three well-known more general realistic agent types. Other agent types, for which we do not have any name, are still possible. The relation between these and other realistic agent types forms a lattice illustrated in Figure 2. The level in this hierarchy indicates the generality of agent types. The bottom of this lattice is the realistic agent type that is characterized by the least number of constraints on the ρ function. Each higher layer adds additional constraints resulting in more specific agent types. At the second level, the stable, social, and selfish agent types result, and at the fourth level the mentioned six specific and unknown agent types (BIDO, BIOD, BDIO, BDOI, BOID, and BODI) result. The top of this lattice is the falsum, which indicates that adding any additional constraint to the ρ function results in an inconsistent ordering.

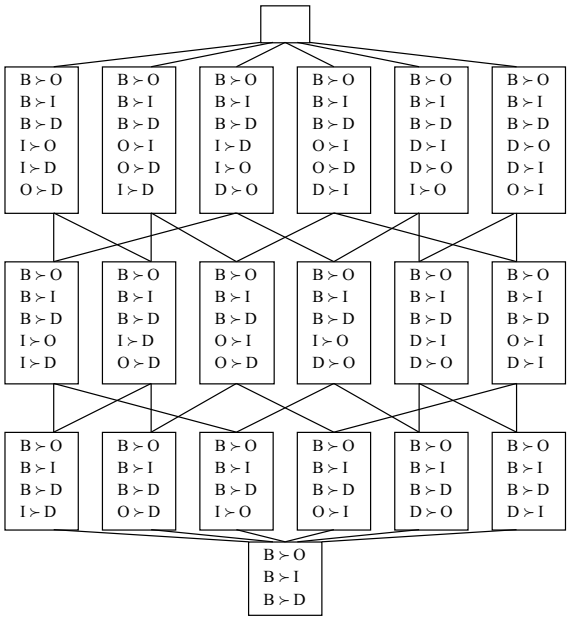


Figure 2: The lattice structure of agent types.

Example

In this section, we illustrate how conflicts between mental attitudes can be solved within the BOID architecture by giving an example that describes the following mental attitudes: *If I go to Washington DC (Go2DC), then I believe that there are no cheap rooms (ChRm) close to the conference site (Close2ConfSite). If I go to Washington DC, then I am obliged to take a cheap room. If I go to Washington DC, then I desire to stay close to the conference site. I intend to go to Washington DC.* This example can be represented by the following rules:

$$\begin{aligned}
\rho = 5 \quad & (Go2DC \wedge ChRm) \overset{\text{B}}{\hookrightarrow} \neg Close2ConfSite \\
\rho = 4 \quad & (Go2DC \wedge Close2ConfSite) \overset{\text{B}}{\hookrightarrow} \neg ChRm \\
\rho = 3 \quad & Go2DC \overset{\text{D}}{\hookrightarrow} Close2ConfSite \\
\rho = 2 \quad & Go2DC \overset{\text{O}}{\hookrightarrow} ChRm \\
\rho = 1 \quad & \top \overset{\text{I}}{\hookrightarrow} Go2DC
\end{aligned}$$

Lets examine a specific type of social agent, i.e. BIOD. Let the input of the agent be empty. Then, following the extension calculation mechanism, we first derive all beliefs and intentions, resulting in the following extension:

{Go2DC}

Because it is a social agent (i.e. the fourth rule has a higher priority than the fifth rule), the obligation rule is applied first. This results in the following intermediate extension:

{Go2DC, ChRm}

This extension is fed back into the B component where it triggers the first belief rule (i.e. the first

rule), because the second belief rules is not applicable as we already have ChRm. This produces the following final extension:

{Go2DC,ChRm,¬Close2ConfSite}

This extension denotes the situation in which the agent has decided to go to Washington DC and takes a cheap room not close to the conference site, which is indeed social behavior.
However, if we exchange the priority of the fourth and the fifth rules the agent becomes a selfish agent 'BIDO'. Then, the *D*-rule would be applied before any obligation rule is applied, resulting in the following final extension:

{Go2DC,¬ChRm, Close2ConfSite}

Sending the results back to the belief component does not make any difference here. This extension denotes the situation in which an agent has decided to go to Washington DC and takes an expensive room close to the conference site, which is indeed selfish behavior.

Concluding Remarks

We have briefly discussed the generic component based BOID architecture that is developed for cognitive agents. Each component in the BOID architecture represents a mental attitudes of the agent. The output of components may conflict. Some of the conflicts that may arise among BOID's components are presented. In the BOID architecture the conflicts are resolved by the order of output generation from different components. We have shown that the order of output generation determines the type of an agent. In general, the order of output generation can be used to identify different types of agents. We have shown that these conflict resolution mechanisms provide some well-known agent types and an interesting set of unknown agent types. In particular, we have shown that for a realistic agent beliefs are generated before obligations, intentions or desires; for a stable agent intentions are generated before desires; and for selfish agents desires are generated before intentions.

We believe that the way the BOID components are updated depends also on the type of the agent. The integration of updating various components have the highest priority in our research agenda. Another issue which in on our future research agenda is the incorporation of agent types derived from plan generation module and its interaction with goal generation modules. In the BOID architecture, the plan generation module influences the computation of extensions and therefore may play an important role in agent type classification. For example, when a generated extension cannot be transformed into a sequence of actions, another extension should be selected. The exact choice for a new extension should depends on the type of agent as well.

References

[1] M. E. Bratman. *Intention, plans, and practical reason.* Harvard University Press, Cambridge Mass, 1987.

[2] J. Broersen, M. Dastani, Z. Huang, J. Hulstijn and L. van der Torre. The BOID Architecture: Conflicts between beliefs, obligations, intentions, and desires. Proceedings of Fifth International Conference on Autonomous Agents (AA'01), "9–16",ACM Press (2001)

[3] R. A. Brooks. A robust layered control system for a mobile robot. IEEE J. Robotics Auromat., RA-2(7):14–23, Apr. 1986.

[4] C. Castelfranchi. Prescribed Mental Attitudes in Goal-Adoption and Norm-Adoption. In AI and Law, Special Issue on Agents and Norms, 7, 1999, 37-50.

[5] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[6] M. Gelfond and T. Cao Son. *Reasoning with Prioritized Defaults.* Proceedings of Logic Programming and Knowledge Representation 1997, 164-223, Port Jeerson, New York, October 1997.

[7] M. P. Georgeff and A. L. Lansky. *Reactive reasoning and planning.* In Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87), pages 677–682, 1987.

[8] A. Rao and M. Georgeff. *An abstract architecture for rational agents.* In Proceedings of the KR92, 1992.

[9] A. Rao and M. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*, 1995.

[10] R. Thomason. Desires and defaults. In *Proceedings of the KR'2000.* Morgan Kaufmann, 2000.

Declarative and Procedural Strategies in Problem Solving: Evidence from the Toads and Frogs Puzzle

Fabio Del Missier (delmisfa@univ.trieste.it)
Department of Psychology, via S. Anastasio 12
Trieste, I-34134 Italy

Danilo Fum (fum@univ.trieste.it)
Department of Psychology, via S. Anastasio 12
Trieste, I-34134 Italy

Abstract

The relationship between theoretically-grounded hints, strategy selection, and solution performance in the Toads and Frogs puzzle, a well-structured problem in which weak methods are difficult to apply, is investigated through an experiment and an ACT-R simulation. The main results show that providing a state specific hint is useful in speeding up the adoption of a hybrid solution strategy, comprising both the retrieval of previous moves and the proceduralization of a domain-specific heuristic that avoids any kind of forward search. The implications of the results for the problem solving theory are discussed.

Introduction

Research work on problem solving has attained to significant success in identifying the sources of difficulty for several kinds of well-structured problems (Newell & Simon, 1972). Working memory limitations (Miyake & Shah, 1999), in particular, play a prominent role in explaining why some problems are "so hard", and many factors have been identified that affect the working memory load. A partial list comprises the execution of legality tests on the operators (Kotovsky, Hayes & Simon, 1985), the number of options to be considered, and the availability of useful external memories (Cary & Carlson, 2001; Zhang & Norman, 1994).

Independently from working memory limitations, problem solvers seem reluctant to engage in a high degree of forward planning (Atwood, Masson & Polson, 1980; Ward & Allport, 1997; Simon & Reed, 1976). People usually recur to heuristic strategies, often relying on weak methods such as hill-climbing or means-ends analysis (Anderson, 1982; Anzai & Simon, 1979; Simon, 1975; Simon & Reed, 1976), and take into account only a limited number of states.

It is however interesting to wonder what strategy would be used in problems requiring a substantial degree of search when weak methods are not directly applicable. Here we try to answer this question by carrying out a first empirical exploration of problem solving behavior in a new kind of task.

First, we introduce the *Toads and Frogs* puzzle (henceforth T&F), a problem we found particularly suitable for the present research because of the peculiar structure of its problem space. We describe then a set of candidate strategies for solving it, and present an experiment designed to study the effectiveness of two types of hints delivered through the interface. Next we summarize the results of an ACT-R simulation aimed at identifying the strategies actually used by participants, and at tracing their development. Finally, we discuss the findings in the light of two main classes of problem solving strategies (memory-based vs. rule-based).

The Toads and Frogs Puzzle

The T&F puzzle (Berlekamp, Conway & Guy, 2001) is a well-structured problem that, to the best of our knowledge, has never been previously utilized in psychological research.

In the variant used here, three toads are placed on the three leftmost squares of a seven-square board while three frogs are placed on the three rightmost squares (Figure 1). The central square is initially empty. The goal of the game is to switch the animals' positions by having the toads occupy the three rightmost, and the frogs the three leftmost squares, respectively. A square can be occupied by only an animal at a time, and an animal can move only into the empty square. Toads can move only rightward and frogs only leftward. There are two possible types of move: a Slide to the next (empty) square and a Jump over an animal of a different type to a two-square apart empty position. A solution can be reached in exactly 16 moves, 9 Jumps and 7 Slides. Two symmetrical solution paths are possible, depending on the animal that is moved first (the solution sequence for the frog-move-first type of problem is presented in Figure 1).

Some of the moves in the solution path are forced (Jump only or Slide only) while in the remaining cases it is necessary to choose between two Slides, or between a Jump and a Slide, or between two Jumps (off the solution path only). Excluding the first move that allows two possible options, there exists a single right move for every

solution step, and it is not possible to retract a move when it has been done.

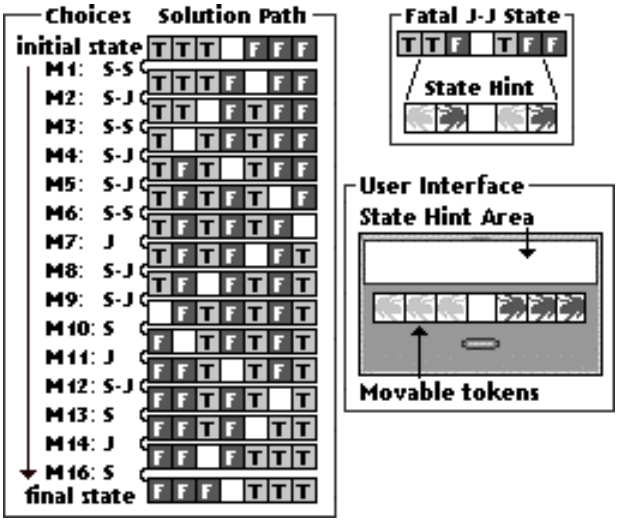


Figure 1: The Toads and Frogs problem

From a normative point of view, it is possible to identify two rules that foster the T&F puzzle solution. First, in deciding between two Slides, avoid the move that brings to a fatal Jump-Jump configuration, i.e., a state in which both legal moves are Jumps (an example is given in Figure 1). Every such configuration lies off the solution path and leads eventually to a dead end. Second, in choosing between a Jump and a Slide, select the Jump.

From a psychological viewpoint, the problem presents some interesting properties. It cannot be solved through hill-climbing, because this strategy does not help in deciding what to do when facing the critical Slide-Slide choices. A pure forward planning approach will not work either, because detecting dead ends would require an unattainable degree of look-ahead. The problem is also hard to solve by means-ends analysis, because of the difficulty to find out useful subgoals. The natural subgoal choice (i.e. putting the most advanced animal to its target location) cannot be used in the first solution steps due to the necessity to plan up to seven moves and to detect ten potential dead-ends. Finally, it is unlikely to find the solution by pure chance (p=0.0039).

Candidate Problem Solving Strategies

From the analysis of some concurrent verbal protocols collected in a pilot study and by combing the problem solving literature, we were able to define a set of candidate solution strategies for the T&F puzzle. We will briefly present the strategies and the hypotheses about the performance measure that can be derived from them.

AVOID BLOCKS (AB) This strategy can be reduced to two rules: (1) Avoid moves that bring to a "blocked" state. A state is considered blocked whenever the moved animal is preceded in its moving direction by an animal of the same type that (a) cannot be moved, and (b) is not placed in its final position. (2) In cases when none of the two legal moves brings to a blocked state, choose randomly between them. This strategy reflects the general heuristic of avoiding bad states, and it hypothesizes chance-level performance in the Slide-Slide choices.

JUMP AND RANDOM (JRND) In the Jump-Slide choices always select a Jump move. In the Slide-Slide choices choose randomly. The "always Jump" rule could be acquired by proceduralizing the hill-climbing heuristic. The strategy predicts a higher error rate in the Slide-Slide decisions in comparison with the Slide-Jump ones.

MOVE PATTERN (MP) Because the sequence of moves to reach the solution is highly patterned, implicitly learn to perform the pattern of moves. For instance, when the first move is the Slide of the most advanced frog [F1], the solution pattern is: [F1] T1 T2 F1 F2 F3 T1 T2 T3 F1 F2 F3 [T2 T3 F3] with the three final moves being on-the-target moves. This strategy (described, among others, by Reber & Kotovsky, 1997) predicts an approximately similar percentage of errors for each class of choices and no difference in the associated decision times.

JUMP AND RETRIEVE (JR) In the Jump-Slide choices, always select a Jump move. In the Slide-Slide choices, try to retrieve the last decision taken in the same situation, using the current state as a memory cue. If the last trial is remembered as a success, repeat the retrieved move, else select the alternative legal move. The strategy constitutes a partial version of the trial-and-error weak method. It predicts higher times (due to retrieval costs) and higher error rates (due to retrieval errors) in the Slide-Slide choices than in the Slide-Jump ones.

JUMP AND SPACE (JS) Always Jump in the Jump-Slide pick. In the Slide-Slide choices, select the move that brings to a state in which there is exactly one interposed square between each neighboring pair of animals like the moved one. This rule implements a domain-specific preference for states in which animals of the same type are regularly spaced. The strategy stems from the verbal protocols of participants stating their desire to reach an "alternating sequence" of animals. They claimed they wanted "to make some space" between the animals of the same type to allow the possibility for the other animals to "jump into". The strategy could be acquired by a perceptual noticing mechanism (Ruiz & Newell, 1989) and by the use of the perceived features as subgoals within a means-ends approach (Anzai & Simon, 1979). It requires: (a) to imagine a state stemming from the execution of a move; (b) to maintain the imagined state in working memory; (c) to perform two distance tests on the imagined state; (d) to select the right move depending on the test outcome. This process is time expensive and can be error-prone, therefore the strategy predicts higher times and error rates in the Slide-Slide decisions than in the Slide-Jump ones.

The Experiment

We performed an experiment to analyze the effect of two interface hints on the performance in the T&F puzzle, and on the acquisition of the solution strategies.

The first hint is motivated by the work of Kotovsky, Hayes & Simon (1985), who suggested that the execution of legality tests on the operators constitutes a major source of difficulty in the Tower of Hanoi isomorphs (the so-called rule-application hypothesis). According to the authors, the working memory load associated with these tests initially hinders the discovery of a solution strategy. Assuming that legality tests could be a source of difficulty also in the T&F problem, we devised an interface hint that completely removes any cost associated with their execution. With the "legality hint" enabled, the movable tokens pop out in the interface, being displayed on squares of a different color. In this way the legality tests are embedded into the interface. This manipulation should free working memory resources, and make them available for problem solving and for the acquisition of a solution strategy.

The second hint is related to the structure of the problem space and, in particular, to our hypothesis that the Slide-Slide choices are the biggest sources of difficulty in the T&F puzzle because they cannot be handled by the weak methods commonly used in problem solving. With the "state hint" enabled, an image pattern representing the common part of the fatal Jump-Jump states is presented in the State Hint Area of the User Interface Window (Figure 1) whenever participants face a Slide-Slide choice. The participants were instructed to avoid executing a move that will bring them in a state corresponding to the hint pattern. According to our hypothesis, the hint should be very effective in removing the main error source, and in helping participants to find a good decision policy in the Slide-Slide choices.

Finally, it is reasonable to expect a synergic interaction between the two kinds of hints, since the working memory unload provided by the first hint should enhance the effect of the state hint.

Method

Participants and Materials The participants were 72 undergraduates students, aged between 19 and 30. The sample was approximately balanced for gender. All the participants had a basic familiarity with computers, and were able to use the mouse. Two different versions of the T&F problem were used by having the first move (Slide a toad vs. Slide a frog) automatically generated by the computer. The first choice actually splits the problem space in two almost completely non-overlapping sub-spaces (there are only a few common states, placed off the solution paths).

Procedure Participants read an instruction document that explained the rules of the T&F, described the task, and showed how to use the interface. Depending on the experimental conditions, the document presented also the available hints. In order to decrease the likelihood of a random solution, we adopted as the learning criterion the attainment of the final state in two consecutive trials. The interface did not allow undoing a previously executed move. After getting stuck, or after a voluntary interruption of a trial, the solver should start again from the beginning. Participants were required to solve both versions of the problem in the order specified by the experimental design. They had allotted a maximum time of 20 min for the first problem and of 10 min for the second one. No limits were placed on the number of trials.

Apparatus A PowerMacintosh G3 was used for the experiment. A program implementing the T&F puzzle was written using MCL 4.3.1 and CLIM2. The program recorded each participant move and the associated time. The interface window was composed by two parts. The upper part was only used to display the state hint. The lower part showed the puzzle board and an "Interrupt" button. To move an animal, it was only required to click on the square containing it. If the move was allowed by the problem rules, the positions of the animal and of the blank square were immediately updated; in case of an illegal move, a warning sound was delivered.

Experimental Design Two between-subjects independent variables (State Hint and Legal Hint availability) were manipulated in a 2x2 factorial design. The 72 participants were randomly assigned to the four experimental groups. We adopted a transfer design, presenting the two different versions of the problem in a counterbalanced order. The hints were available only for the first problem, and the participants were made fully aware of this by the instructions and by the experimenter. The basic dependent variables for each problem were the achievement of the criterion, the total time, and the total number of trials needed to achieve it. More detailed dependent measures were the percentages of errors, and the mean move latency for each choice class.

Results

We will first present the results on the whole data to test the hypotheses about the effectiveness of the interface hints. Then we will report two series of results concerning the participants who reached the criterion in both the problems: the first to assess the transfer, the second to provide detailed performance analyses that will be compared with the predictions of the various strategies.

Hint Effectiveness

Criterion. Table 1 presents the frequency of problem solving outcomes for each hint group and problem.

Table 1: Frequency of problem solving outcomes.

GROUP	Total Failure	Criterion Problem 1 Only	Criterion Problem 2 Only	Criterion Problem 1 & 2
Control	5	0	6	7
Legal Hint	1	2	3	12
State Hint	1	2	0	15
Both Hints	3	4	2	9

In the first problem, a higher number of participants in each hint group reached the criterion in comparison with the control group (no hint). The Fisher Exact test showed significant differences between the control group and the Legal Hint group (p=0.0205), the State Hint group (p=0.0005) and the group with both hints (p=0.0461), respectively. No significant differences were found in the second problem. Contrasting the frequency of criterion attainment in both problems with the frequency of any other outcome, only the State Hint group resulted significantly better than the control group (Fisher Exact test, p=0.0076).

Times. A 2x2 ANOVA on the aggregate problem times yielded only the significant main effect of the State Hint (F(1,68)=8.37, MSE=50.98, p<0.01). The participants without the state hint had to devote more time to the problems (State Hint: M=15.08 min, No Hint: M=19.95 min).

Number of trials. A 2x2 ANOVA on the aggregate problem trials showed a significant interaction between State Hint and Legal Hint availability (F(1,68)=4.78, MSE=216.65, p<0.05). Only the main effect of the State Hint was significant (F(1,68)=10.00, MSE=216.65, p<0.01). The Tukey HSD test showed significant differences between the State Hint group and the control group (p<0.01) and between the State Hint group and the Legal Hint group (p<0.05). The State Hint group had the lowest mean (M=15.89), followed by the control group (M=27), the Legal Hint group (M=30.39), and the group with both hints (M=34.44).

Transfer

Times. We computed a 2x2 ANOVA (Problem x State Hint availability) on the problem solution times. The analysis showed a two-way interaction between Problem and State Hint (F(1,41)=9.20, MSE=9.04, p<0.01). The Unequal N Tukey HSD test underlined a significant difference between the participants with the state hint and those without it, but only in the first problem (p<0.05). In both the conditions, the time to criterion significantly decreased from the first to the second problem (State Hint: n=24, $M1$=7.26 min, $M2$=4.56 min, p<0.05; No Hint: n=19, $M1$=10.85 min, $M2$= 4.19 min, p<0.001).

Number of trials. A 2x2 ANOVA (Problem x State Hint availability) showed a two-way interaction between Problem and State Hint (F(1,41)=9.96, MSE=25.86, p<0.01). The Unequal N Tukey HSD test highlighted the significant difference between the participants with the state hint and those without it, again only in the first problem (p<0.01). Only in the condition without the state hint, the number of trials significantly decreased from the first to the second problem (State Hint: n=24, $M1$=8.42, $M2$=8.12; No Hint: n=19, $M1$=16.63, $M2$=9.37, p<0.001).

Detailed Performance The choices of each participant in the non-forced moves were categorized depending on their location: the first Slide-Slide choice (SS-1), the second Slide-Slide choice (SS-2), or a Jump-Slide decision point. Then, the percentage of errors for each choice point was computed, according to the participant's number of transitions for that state. Finally, given the low value of the percentages for each of the six Jump-Slide choices, an overall mean error percentage (JS-M) for each participant was computed.

Errors. A 2x2x2 ANOVA (State Hint availability x Problem x Error type) yielded a significant three-ways interaction (F(2,82)=3.30, MSE=251.3, p<0.05). There was no significant difference on the JS-M error percentage between the State Hint and No State Hint conditions and between the first and the second problem (Unequal N Tukey HSD test). The JS-M error percentages were significantly lower than the SS error percentages in each State Hint condition and problem (least significant difference: p<0.05; JS-M means between 2 and 0). The difference on the SS-1 error percentages between the State Hint and No State Hint conditions was significant only in the first problem (M-h=18, M-nh=43, p<0.01). A similar result was obtained for the SS-2 error percentages, but the difference only approached significance (M-h=26, M-nh=47, p=0.059). Some single sample t-tests, carried out to evaluate the null hypothesis that the error percentages of the SS choices were not different from random performance (50%), did not allow us to reject the hypothesis only in the condition without state hint of the first problem. In the other cases the percentages were significantly lower (least significant difference: p<0.05; means ranging from 18 to 35).

Move times. A 2x2x2 ANOVA (State Hint availability x Problem x Error type) on the mean move times, showed a significant three-ways interaction (F(3,123)=7.28, MSE=17.57, p<0.001). The analysis comprised also the two kinds of forced moves (Jump and Slide only) and used the aggregated SS decision times. There were significant differences between the first and the second problem, but only for the real choices. The Unequal N HSD test yielded the following results in the conditions with the state hint: p<0.001 for SS ($M1$=10.40 s, $M2$=4.94 s) and p<0.001 for SJ-M ($M1$=5.68 s, $M2$=3.30 s). In the conditions without state hint the results are analogous: p<0.001 for SS ($M1$=7.06 s, $M2$=4.16 s) and p<0.001 for SJ-M ($M1$=5.23 s, $M2$=2.93 s). For both problems, the SS times were significantly higher than the JS times (State

Hint: $p<0.001$ and $p<0.001$; No State Hint: $p<0.001$ and $p<0.05$). The JS times were significantly slower that the forced moves ($p<0.001$ for each condition and problem). The only significant difference involving the State Hint condition was that on the SS decisions in the first problem (*M-h*=10.40 s, *M-nh*=7.06 s, $p<0.05$).

Discussion

The experimental results provided strong support for the effectiveness of the state hint. This hint promoted the achievement of the criterion in the first problem, while its removal did not produce any performance decrease in the second one. A significant support for the effectiveness of the legal hint was not reached, but the limited power of the tests advises caution in the interpretation.

The results on the percentages of errors clearly showed that the main sources of difficulty for the problem were the Slide-Slide choices. The state hint worked by reducing the errors in these decision points in the first problem, thus allowing a faster development of a decision policy.

However, about 60% of the participants were able to reach the criterion in both the problems, thus demonstrating the possibility to acquire an advantageous solution strategy with a sufficient amount of practice. Furthermore, the performance of the successful solvers in the second problem was quite similar across the experimental conditions.

The error and time results were not compatible with the adoption of the MP strategy. The use of the AB or JRND strategies could not excluded in the conditions without the state hint, but only in the first problem. The JR and the JS strategies were the only two strategies potentially in accordance with the evidence on the second problem.

The Simulation

We decided to undertake an ACT-R simulation (Anderson & Lebiere, 1998) to formulate more detailed predictions from the partially supported strategies: Avoid Blocks (AB), Jump and Random (JRND), Jump and Retrieve (JR), and Jump and Space (JS), and to test them against the appropriate subset of data. We implemented the four strategies as separate ACT-R models.

While the JRND strategy leads to a direct implementation, the AB and JS require to mentally simulate the execution of the possible legal moves, and to evaluate the states deriving from them. We implemented the construction and storage of the imagined states via time-costly productions, and we assumed that the state evaluations were always performed errorless. Given that verbal protocol data indicated that detecting a block situation is quite an easy task, we assumed also that the AB model did not need to retrieve the simulated move. Conversely, the JS model, being engaged in more difficult distance tests, had to use the error prone memory retrieval. Thus, the JS model often dictated a move that did not comply with the strategy requirements.

The implementation of the JR strategy required, on the other hand, the memory storage of each Slide-Slide choice, and of the outcome of each trial. When facing a Slide-Slide problem state, the model tries to retrieve the last decision taken in the same situation using the current board configuration as a retrieval cue. It also tries to retrieve the outcome of the last performed trial. If the trial is remembered as a success, the retrieved move will be executed, otherwise the alternative legal move will be carried out. Thus, the JR model sometimes makes the wrong selection due to retrieval errors or to a temporal mismatch between the retrieved move and the trial.

Procedure and Results

We compared the AB and JRND predictions with the data of the first problem without the state hint. Then, we contrasted the JR and JS results with the data of the second problem, in the conditions without the state hint.

For each strategy, we executed a number of simulation runs (2000) sufficient to provide an efficient estimation. The dependent variables were the number of trials to reach the criterion, and the error percentages on the two Slide-Slide choices. The JS-M percentages were not taken into account, because all the strategies predicted few errors in the JS choices and the empirical values are always very close to zero. The experimental data and the simulation results are presented in Table 2 and Table 3 (the number near the strategy label stands for the problem being simulated).

Table 2: Experimental data.

PROBLEM	Trials to Criterion	SS-1 Error	SS-2 Error
FIRST	16.63	43	47
SECOND	9.37	30	35

Table 3: Simulation predictions. (Values within the 95% confidence interval of data are marked with *)

STRATEGY	Trials to Criterion	SS-1 Error	SS-2 Error
AB (1)	13.95*	43*	40*
JRND (1)	13.08	42*	39*
JR (2)	7.56*	34*	37*
JS (2)	13.70	42	39*

The results showed that the AB strategy obtained a slightly better fit than the JRND on the first problem results. The best fit for the performance on the second problem was obtained by the JR strategy. Thus, the most probable explanation is that participants shifted from the AB to the JR strategy as a consequence of their increased experience with the task.

It seems reasonable to assume that the state hint was able to foster the adoption of the JR strategy. A simple memorization of the state hint seems quite implausible, given the move latency data. In the second problem, if the participants were retrieving the state hint and using it to carry out the SS move selection, we should have observed a significant increase in the mean latency. On the contrary, the mean times for the SS moves resulted much lower in the second problem. Furthermore, the whole second problem performance was very similar for the groups with and without the state hint. So, it could be parsimoniously hypothesized that the indirect suggestion of the correct move made available through the state hint could have simply speeded up the natural development of a more general memory-based strategy.

Conclusions

The evidence provided can help us to answer the question that motivated our research on the T&F puzzle. People were able to acquire an effective solution strategy even when hill-climbing or means-ends analysis were not directly applicable. We gained support for a state-avoidance strategy in the initial problem solving attempts, and for a memory-based strategy in later trials. Finally, we demonstrated the effectiveness of a specific type of interface hint.

From a broader prospective, it is worth noting that some findings, obtained in very different settings (Howes & Payne, 2001), seem to bring converging evidence for a kind of memory-based problem solving when weak methods are not applicable or not efficient.

Another meaningful point is that our hybrid memory-based JR strategy was probably derived partly from the weak method of trial-and-error, and partly from the proceduralization of the state-avoidance heuristic. This raises an interesting theoretical issue concerning the ontology of multi-step problem solving strategies. The strategies commonly proposed in the literature seem to belong either to the algorithmic or to the memory-retrieval class. Our work seems to suggest that, in some multi-step situations, people are able to acquire hybrid strategies, relying on memory retrieval to handle problem solving steps where procedural methods don't work. So, in our view, it seems necessary to make a distinction between the strategies that require the intentional usage of memorized instances (Logan, 1988), the ones that keep track of the previous history in a procedural form (Lovett & Anderson, 1996), and the hybrid formulations. This also means that it will be generally necessary to make explicit all the potential candidate solution strategies, and to contrast them in their capacity to fit the data.

References

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 396-406.

Anderson, J.R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anzai, Y. & Simon, H.A. (1979). The theory of learning by doing. *Psychological Review*, 86, 124-140.

Atwood, M.E., Masson, M.E. & Polson, P.G. (1980). Further exploration with a process model for water jug problems. *Memory & Cognition, 8*, 182-192.

Berlekamp, E.R., Conway J.H. & Guy, R.K. (2001). *Winning Ways for Your Mathematical Plays*, v1, A K Peters.

Cary, M. & Carlson, R.A. (2001). Distributing working memory resources during problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 836-848.

Kotovsky, K., Hayes, J.R. & Simon, H.A. (1985). Why are some problems hard?: Evidence from the Tower of Hanoi. *Cognitive Psychology*, 17, 248-294.

Howes, A. & Payne, S.J. (2001). The strategic use of memory for frequency and recency in search control. *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 425-440). Hillsdale, NJ: Erlbaum.

Logan, G. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Lovett, M.C. & Anderson, J.R. (1996). History of success and current context in problem solving. *Cognitive Psychology*, *31*, 168-217.

Miyake, A. & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Newell, A., & Simon, H. A. (1972*). Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Reber, P.J & Kotovsky, K. (1997) Implicit learning in problem solving: The role of working memory capacity. *Journal of Experimental Psychology: General, 126, 178-203.*

Ruiz, D. & Newell, A. (1989). Tower-noticing triggers strategy-change in the Tower of Hanoi: a Soar model. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 522-529). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, *7*, 268-288.

Simon, H. A. & Reed, S.K. (1976). Modelling strategy shifts on a problem solving task. *Cognitive Psychology*, *8*, 86-97.

Zhang, J. & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*, 87-122.

Ward, G. & Allport, A. (1997). Planning and problem solving using the 5-disk Tower of London task. *Quarterly Journal of Experimental Psychology*, *50*, 49-78.

Teaching with Dialectic Arguments vs. Didactic Explanations

Ravi Desai (rpdst6@pitt.edu)
Intelligent Systems Program, University of Pittsburgh,
LRDC, 3939 O'Hara Street
Pittsburgh, PA 15260 USA
Kevin D. Ashley (ashley@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260 USA

Abstract

We compared two automated approaches to teaching *distinguishing*, a skill that involves assessing the relevant differences among cases in a context-sensitive way. We implemented the two approaches in two versions of an Intelligent Tutoring System designed to teach law students basic skills of case-based legal argument. The original version of CATO employs didactic explanation. The newer version, CATO-Dial, teaches the same skill with a simulated dialectic argument in a courtroom setting. We hypothesized that students would learn the skill better by engaging in the simulated argument than by receiving interactive explanation. We expected that dialectic argument would help students to construct the target knowledge on their own as they developed responses to arguments, and that this would lead to more robust learning. We showed that students in the dialectic argument simulation group performed significantly better on a section of the post-test aimed at assessing transfer of their skills of distinguishing. We discuss a number of cognitive and motivational benefits that may explain this effect

Introduction

The skill of distinguishing is important in dialectical domains such as law, applied ethics, policy analysis, and business, where arguments by analogy are routinely employed in professional education and practice. Distinguishing is a way to respond to an analogical argument. The argument claims that a target problem should be decided in the same way as a cited source case by virtue of their relevant similarities, that is, factual patterns the cases share that form the basis of legal reasons for deciding them in the same way. A *distinction* is a factual difference underlying a legal reason to decide the target problem differently from the cited case. There may be many differences, but only those that give rise to legal reasons for treating the cases differently are distinctions.

A distinction may be a factual strength of a side (i.e., a party, either plaintiff, the one who commences suit, or the defendant) in the target problem not shared in the source case, or a factual weakness in the source case not shared in the target. For the cases in our experiment, students play the role of the defendant's attorney; a distinction is defined as a pattern of facts that strengthens defendant's legal argument in the problem not found in the cited-case, or a fact pattern that weakens defendant's side in the cited-case not found in the problem.

In order to distinguish effectively, one must be sensitive to the argument context in which a case has been used. The context includes the role a relevant difference plays in an argument, its underlying legal significance both in absolute terms and relative to the other combinations of facts in the target problem and cited precedent.

Professors note that law students often demonstrate only a shallow understanding of the concept of a distinction. Students may be able to find different facts but fail to realize that only some differences are distinctions. Students may also ignore which side a difference favors, or they fail to view the significance of a difference in the context of the other facts in the problem and cited-case. Because of their shallow knowledge, students may make arguments citing differences that hurt, rather than help, their side's argument.

Ideally, students "pick up" the skill of distinguishing through the trial-and-error experience of making arguments. In a law school, students engage in arguments in the classroom. Sometimes, however, students are reluctant to expose themselves in class by making arguments, and, in any event, could benefit from additional instruction outside the classroom. In an effort to meet this need, Aleven and Ashley (1997) developed the CATO program (i.e., Case Argument TutOrial), an Intelligent Tutoring System designed to teach basic case-based argument skills, including how to distinguish cases (see also Aleven, 1997).

In designing an ITS to teach distinguishing, one may try two approaches. *Didactic explanation* involves presenting good and bad examples of distinguishing. The system explains

why the examples are instances of successful or unsuccessful argument. The bad examples illustrate the various kinds of shallow knowledge. This is how CATO teaches distinguishing.

Dialectic argument attempts to teach students distinguishing by engaging them in arguments, affording an opportunity to learn the skill through a process of trial-and-error. Technically, it is harder to design this kind of pedagogical interaction. Before undertaking to develop a large-scale dialectic argument system, we wanted to see whether it was likely to improve students' learning. We therefore developed a variation of CATO, called CATO-Dial, which employs dialectic argument to teach distinguishing.

We hypothesized that students would learn better to distinguish by engaging in dialectical argument than in didactic explanation. We speculated that students engaged in role-playing and arguing would encounter information in a more relevant context and would be motivated to process the information more thoroughly. They would develop deeper knowledge of the role a difference plays in the argument context, its interactions with other facts in the target problem and source case and the underlying reasons why it matters. Our experiment tested this hypothesis.

In aiding students to learn deeper knowledge, we also hoped that CATO-Dial would promote better transfer of knowledge. A major problem in knowledge transfer is that people tend to access prior knowledge that bears superficial rather than structural similarity to the problem at hand (Thompson et al., 2000).

Legal argument is not as determinate a form of problem-solving as, say, physics or geometry. Legal problems rarely have provably correct answers; there may be reasonable arguments on both sides of a dispute based on analogies to competing cases (Ashley, 1990). CATO-Dial attempts to engage students in argument dialogues that focus them on comparing cases.

Law school professors engage students in Socratic dialogues about cases in the casebooks. Some of the earliest ITSs used Socratic dialogue and an inquiry teaching method to teach subject matter such as geography (SCHOLAR (Carbonell, 1970)) or meteorology (WHY (Collins and Stevens, 1982)). A subsequent tutoring system (Wong et.al, 1997) incorporated the inquiry teaching method into an ITS shell and geography tutor. The OLIA ITS (Retalis et al., 1996) used a related argument dialogue strategy, playing devil's advocate. Research suggests that students tutored manually with Socratic dialogues learned targeted physics concepts (i.e., rules) better than those taught with more didactic dialogues (Rose et. al, 2001). In the latter, the human tutor provided more explanation before asking questions but asked fewer open-ended questions.

CATO vs. CATO-Dial

CATO is one of the few case-based Intelligent Tutoring Systems that teaches a process of case-based reasoning. (Aleven, 1997, pp. 197-8). It provides a set of specialized tools, accessible through an X-server connection to CATO running on a Unix workstation, and a web-accessible Casebook and Workbook. The Casebook presents excerpts from important legal case opinions in trade secret law. A small set of argumentation and discussion questions follows each case, much like an ordinary legal casebook. The Workbook helps students use CATO's tools to analyze and respond to the argumentation and discussion questions.

Experiments show that CATO is an effective teacher (Aleven & Ashley, 1997; Aleven, 1997). Students work with CATO's textual case summaries and abstract representations of cases in terms of *factors*. Each factor represents a stereotypical collection of facts, which tends normally to strengthen or weaken a conclusion that a side should win a particular kind of legal claim (Ashley, 1990). A Factor Hierarchy represents legal reasons why a factor makes a difference to the legal claim (Aleven, 1997).

CATO helps students analyze target problems and compare them to past cases. It teaches novices to identify factors in a target problem, test hypotheses about their significance against cases in its database, and make legal arguments about how to decide the target problems citing cases. Students encounter problems based on real litigated cases. Novice users identify conflicting factors in the problem, which give rise to conflicting reasons for decision. CATO teaches them how to make arguments to resolve such conflicts.

CATO's Argument Maker tool provides a tutorial on distinguishing. To enable CATO to employ dialectic argument, we developed CATO-Dial, a modification of the program that engages novice users in courtroom-style arguments about target problems. In using the CATO-Dial version of the tutorial on distinguishing, students encounter examples like the *Lynchburg Lemonade* case.[1] With the Case Analyzer tool, students have

[1] In the *Lynchburg Lemonade* case, Tony Mason, the plaintiff, developed a cocktail he dubbed "Lynchburg

identified conflicting factors in the *Lemonade* problem and have begun to consider the conflicting factor-based legal reasons about how to decide its outcome. For comparing cases, the Case Analyzer presents them in a tabular form as in Figure 1. The problem has six factors, four of which favor the plaintiff (p) and two of which favor the defendant (d). The *Boeing* case, won by plaintiff (p), shares two of these factors, the relevant similarities (marked with "="). The relevant differences (i.e., distinctions) are the four unshared factors marked with "*". These favor deciding the *Lemonade* case for the defendant (i.e., differently from *Boeing*). Note that F16 strengthens the defendant in the *Lemonade* case and is not found in *Boeing*, whereas F4, F12 and F14 strengthen the plaintiff's position in *Boeing* and are not found in the *Lemonade* case. F10, F15 and F18 are also unshared factors, but they are not distinctions because they favor deciding the *Lemonade* case for plaintiff (i.e., the same as in *Boeing*).

With CATO-Dial, students play the role of an advocate, Perry Mason, who has to argue a case in court. As shown in Figure 2, the student may put arguments in the mouth of Perry Mason by selecting argument moves and values from a menu. CATO-Dial responds on behalf of the Judge, who mediates the proceedings, Hamilton Burger, Perry's opposing counsel, and Della Street, Perry's savvy assistant, who offers helpful hints.

In the dialogue, Mr. Burger's responses (such as step 5 in Figure 2), generated by CATO-Dial, take advantage of any weaknesses in Mr. Mason's argument, based on the students' menu selections. The Judge's reaction emphasizes the student's mistake, and Della's hints, also generated by CATO-Dial, provide instruction on how to correct them. CATO-Dial can engage in dialogues like this for any pair of relevant cases in its database.

Lemonade". Since Tony took some measures to protect his recipe's secrecy, and since his was the only tavern producing this drink, we say factors F6, Security-Measures, and F15, Unique-Product, apply; both tend to favor the plaintiff (p). On the other hand, Tony disclosed his recipe in negotiations with a sales agent of the defendant, Jack Daniel's Distillery. Thus, F1, Disclosure-In-Negotiations, applies, a factor that tends to favor the defendant (d). The agent was aware, however, that the recipe was a "secret formula", so F21, Knew-Info-Confidential, also applies, tending to favor the plaintiff. Without compensating Tony, the defendant started marketing exactly the same kind of cocktail; F18, Identical-Products applies and favors plaintiff. Finally, the recipe could have been obtained by reverse engineering the cocktail; F16, Info-Reverse-Engineerable, applies and favors the defendant.

CATO-Dial's dialectic argument contrasts with CATO's didactic explanation. Originally, CATO's tutorial engaged students in an interactive exercise employing didactic explanation. In these explanations, CATO did not engage the student in as direct a manner. Instead, the student was presented with an argument that a hypothetical opponent could make. Below is an excerpt from CATO's original version of the dialogue of Figure 2. First, CATO makes the same argument as in Step 1 of Figure 2 and invites the student to select the factors that he would emphasize in distinguishing the *Boeing* case on behalf of the defendant. Second, the student again selects three differences (i.e., factors F15, 516, and F18) which though not shared between the two cases, actually strengthen the argument that plaintiff should win in the problem situation. CATO then responds as follows:

3. CATO: Suppose defendant made the following argument: The *Boeing* case is distinguishable, because in the *Lynchburg Lemonade* case, defendant's product was identical to plaintiff's [F18] and plaintiff was the only manufacturer making the product [F15]. This was not so in *Boeing*. Would this really help defendant? When you have answered this question for yourself, please click on "Continue".

4. STUDENTS: (Continue)

5. CATO: Plaintiff could respond to defendant's argument, as follows: Your Honor, as defendant's counsel correctly pointed out, in the current problem, defendant's product was identical to plaintiff's [F18] and plaintiff was the only manufacturer making the product [F15]. *Boeing*, which was won by plaintiff, does not have these strengths on plaintiff's part but is otherwise similar to the problem at hand. Where these additional strengths are present, a decision in favor of plaintiff is warranted even more strongly than it was in *Boeing*. In short, even though factors F18 and F15 are differences, defendant cannot use them to distinguish *Boeing*. Defendant needs to focus on other differences, namely, on those differences that make the *Lynchburg Lemonade* case look stronger for defendant than *Boeing*. Please click on "Continue". *See* (Aleven, 1997 at 141f).

The argument that CATO uses to show the inadequacy of the student's initial selection of distinctions is the same in both versions; the difference lies in the manner in which it is presented to the students and the way in which students are engaged in the task

Description of Experiment

We compared the impact of teaching distinguishing to senior undergraduates using the two versions of CATO. The students had all been accepted into law schools and were receiving

preparatory instruction through a summer institute. The students were randomly assigned to two groups. The experimental group used the dialectic argument version of CATO-Dial and initially numbered 22 students. The control group worked with the didactic explanation version of CATO and initially numbered 23 students. Each group worked in a series of eight two-hour sessions over a span of about one month from June 5 through July 11, 2000. For each session a student was paired with a different partner from the same group.

Prior to the series of instructional sessions, all students took a pre-test comprising three questions designed to assess argumentation skills. For Questions 1 and 2, students read a problem situation and three short cases. Students were asked to make and respond to arguments about the problem given the cases. Question 3 asked them to define the concepts of a relevant similarity and relevant difference.

Lynchburg Lemonade Case	*Boeing* (p)
= Fl Disclosure-In-Negotiations (d)	= Fl Disclosure-In-Negotiations (d)
= F6 Security-Measures (p)	* F4 Agreed-Not-To-Disclose (p)
F15 Unique-Product (p)	= F6 Security-Measures (p)
* F16 Info-Reverse-Engineerable (d)	F10 Secrets-Disclosed-Outsiders (d)
F18 Identical-Products (p)	* F12 Outsider-Disclosures-Restricted (p)
= F21 Knew-Info-Confidential (p)	* F14 Restricted-Materials-Used (p)
	= F21 Knew-Info-Confidential (p)
= shared factor * distinction	

Figure 1: Case Comparison In Terms of Factors

Court is in session…

1. Mr. Burger for Plaintiff (CATO): Your Honor, where plaintiff adopted security measures [F6] and defendant knew that plaintiff's information was confidential [F21], even though plaintiff disclosed its product information in negotiations with defendant [Fl], plaintiff should win a claim of trade secrets misappropriation, as in *The Boeing Company v. Sierracin Corporation*, 108 Wash.2d 38, 738 P.2d 665 (1987).

2. Judge (to Defendant's counsel): Your response, Mr. Mason?

3. <STUDENT (selecting from menu): Distinguish *Boeing* [F15] [F16] [F18] (Continue)>

Mr. Mason for Defendant (Student): *The Boeing Company v. Sierracin Corporation* is distinguishable, Your Honor, because in the *Lynchburg Lemonade* case, plaintiff's product information could be learned by reverse-engineering [F16]. This was not so in *Boeing*. *Boeing* is also distinguishable, because in the *Lynchburg Lemonade* case, unlike *Boeing,* defendant's product was identical to plaintiff's [F18] and plaintiff was the only manufacturer making the product [F15].

4. Judge (to Plaintiff's counsel): Mr. Burger?

5. Mr. Burger for Plaintiff (CATO): Your Honor, as Mr. Mason concedes, in the current problem, defendant's product was identical to plaintiff's [F18] and plaintiff was the only manufacturer making the product [F15]. *Boeing*, which was won by plaintiff, does not have these strengths on plaintiff's part but is otherwise similar to the problem at hand. Where these additional strengths are present, a decision in favor of plaintiff is warranted even more strongly than it was in *Boeing***.**

6. Judge (to Defendant's counsel): I am sorry, Mr. Mason, but I am afraid I have to agree with Mr. Burger.

7. STUDENT: Help! (Continue)>

8. Della Street (whispering into Perry's ear): Here is a hint. Even though factors F18 and F15 are differences, defendant cannot use them to distinguish *Boeing*. Defendant needs to focus on other differences, namely, on those differences that make the *Lynchburg Lemonade* case look stronger for defendant than *Boeing*.

9. <STUDENT (selecting from menu): Distinguish *Boeing* [F4][F12] [F14] [F16] (Continue)>

Mr. Mason for Defendant (Student): *Boeing* is distinguishable, because in *Boeing*, defendant used materials that were subject to confidentiality restrictions [F14], plaintiff's disclosures to outsiders were subject to confidentiality restrictions [F12], and defendant entered into a nondisclosure agreement with plaintiff [F4]. This was not so in the *Lynchburg Lemonade* case. Also, in the *Lynchburg Lemonade* case, plaintiff's product information could be learned by reverse-engineering [F16]. This was not so in *Boeing*.

10. Judge (to Plaintiff's counsel): Mr. Burger?

11. Mr. Burger for Plaintiff (CATO): Uhh. Well, as you know, Your Honor, *Boeing,* is not the only case we cite in our brief.

12. Della Street (whispering into Perry's ear): Brilliant job, Perry!

Figure 2: Sample Dialectic Argument in CATO-Dial

During the first six sessions, the experimental and control groups were treated identically. In Sessions 1 to 4, both groups of students familiarized themselves with the factor based representation system and the CATO environment. In Sessions 5 and 6 all students learned how to use the Argument Maker tools.

The control group and experimental group were treated differently in Sessions 7 and 8. These two sessions focused on teaching students how to distinguish cases. During both Sessions 7 and 8, students were expected to complete eight pair-wise case comparisons involving distinguishing. The eight pairs of cases were the same in both groups. The only difference was the manner in which CATO and CATO-Dial taught the lesson. The experimental group worked with CATO-Dial's simulated courtroom dialogues like that in Figure 2. The control group worked with the original CATO didactic explanations.

After Session 8, all students took a post-test comprising six questions. The first three questions were worded identically to the pre-test questions, but Questions 1 and 2 involved a different problem and cases. The other three questions tested the following transfer skills:

- Question 4 put students in a new role – instead of making arguments they critiqued an argument.
- Question 5 tested students' recall of a particular problem situation they had encountered in the instruction. This problem had been used extensively in the teaching sessions as a basis of the argumentation lessons. Students were asked to make and respond to an argument about the problem, which they had to recall from memory, by analogizing it to and distinguishing it from a new case presented with the question.
- Question 6 tested students' ability to apply skills they had learned to an unfamiliar new domain: the copyright law doctrine of Fair Use. As such, this question is a telling measure of their ability to distinguish cases.

The director of the University of Pittsburgh School of Law legal writing program graded all but one of the pre-test and post-test questions. The grader was provided a one-page summary of grading criteria and instructed to assign a gestalt grade (between 1 and 10) to each question. He was blind as to the identity of the test writers, but did know which were pre-tests and which were post-tests. The exception was Question 5, the recall question, for which we developed an objective grading scheme. Students were awarded a maximum of ten possible points on the basis of how many of the factors in the problem they referred to in their argument.

Analysis

Post-test data were available for only 22 of the 45 students, 15 in the experimental group and 7 in the control group. The rest of the subjects either dropped out before the end of the experiment or did not complete enough work in Sessions 7 and 8, the only sessions involving differential treatment of the two groups.

Pre-test scores were analyzed for the 22 students, who provided both pre-test and post-test data. For each student, responses to the three pre-test questions were summed, and the mean response of students in the experimental group was compared to that of students in the control group, using a two-tailed t-test. Results showed no statistically significant difference between the two groups. Since the students were paired with different partners across sessions, we used the individual student rather than the pair as the unit of analysis for both pre-test and post-test analyses.

Post-test scores were also analyzed for the 22 students who provided both pre-test and post-test data. A two-tailed t-test indicated no significant difference in the mean post-test scores of the experimental and control groups with respect to the first five questions. For question 6, however, the mean post-test score of the experimental group was significantly higher than that of the control group ($t(7.1) = 3.87$, $p < .05$, effect size of 1.57).[2] On question 5, the mean post-test score of the experimental group was nearly statistically significantly higher than that of the control group ($t(20) = 1.39$, $p = 0.052$, effect size of 0.62). While the difference was not significant the experimental group scored substantially higher on question 4 with an effect size of 0.88.

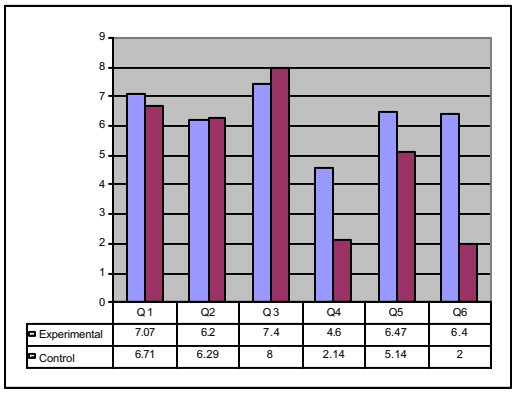


Figure 3: Comparison of Mean Post-Test Scores

Discussion and Conclusion

Our hypothesis was that students would learn the skill of distinguishing better by engaging in simulated dialectical argument than in interactive didactic explanation. The data confirmed our hypothesis in an interesting way. Whereas dialectic argument was not more effective than didactic explanation in teaching argumentation skills, it was more effective in helping students transfer the skills they learned to new tasks and significantly better in helping students apply the skills to an unfamiliar legal domain. Dialectical argument may have induced students to construct a schema for making and responding to arguments, resulting in deeper knowledge and thus better performance on transfer skills. Experimental students also appeared to have a better understanding of the importance that context plays in the task of finding distinctions. The grader also evaluated the

[2] Degrees of freedom for this test were reduced from 20 because Levene's test for equality of variances indicated that the variances of scores in the experimental and control groups were not equal.

answers in terms of four grading criteria, each involving a simple binary positive-or-negative scale. Two differences emerged. Students in the experimental group were more often rated positively on the criteria "Avoids making opponent's argument" and "Avoids errors regarding which side strengths favor". These results support the conclusion that the experimental manipulation helped students to learn better when a difference is a distinction.

It is intriguing that the rather superficial transformation from CATO to CATO-Dial in the presentation of the lesson on distinguishing had such a dramatic benefit. After all, both programs presented the same basic information. The critical difference, we believe, is that CATO-Dial's dialectical argument simulation provided that information in a more useful way. The dialectical argument offers several potential benefits, any or all of which may explain the observed difference in transfer scores.

Students may have found the increased level of involvement and the competitive element in the courtroom simulation motivating and thus conducive to paying attention and learning.

Furthermore students may have found it easier to understand the program's responses in CATO-Dial than in CATO. It is awkward for CATO to explain the quality of a student's response by example. The dialectical argument simulation, by contrast, provides a more natural context for illustrating the effect in an ongoing dialogue regarding a student's choices. Students in the experimental group did report finding the dialogues somewhat (though not significantly) more helpful than did those in the control group. When asked, "When CATO did provide instructional feedback, how helpful was it?", the CATO-Dial students rated it as more helpful than did the CATO students (Ms = 6.76 and 5.56 out of 10, respectively).

Role-playing in a courtroom argument, with its cognitive and emotional expectations, may also be important. Courtroom simulation explicitly prompts the student. An interactive style of human tutoring, in which tutors prompted students, supported learning even when tutors did not provide explanations and feedback (Chi et al., 2001). Having been prompted to participate, students may have been more likely to argue the merits to themselves, a task cognitively similar to self-explanation (Chi et al., 1989). Dialectical argument may also induce a student to feel worse about making a mistake than does didactic explanation. If so, students are more likely to pay attention and to care about learning in the former context. Role-playing may also induce students to compare the cases more carefully, helping the transition from shallow to deeper knowledge. In a recent investigation, business school students who compared cases in a study phase were three times more likely to transfer the cases' implicit principle to a new application than were those who simply read the cases for the purposes of advice-giving (Thompson et al. 2000).

The results suggest that the CATO-Dial approach is potentially quite valuable. Since laws change, law professors value students' ability to transfer skills. Our subsequent work will focus on converting as much of the CATO curriculum as possible to a dialectical format.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 9720359. We thank Professor Kevin Deasy, University of Pittsburgh School of Law, for his many contributions to this work.

References

Aleven, V. (1997) *Teaching Case-Based Argumentation Through a Model and Examples*, Ph.D. Diss., U. Pgh, LRDC.

Aleven, V. & Ashley, K.D. (1997) "Teaching Case-Based Argumentation Through a Model and Examples". In *Proc. 8th World Conf. AI in Education Society*. pp. 87-94. IOS Press: Amsterdam. Kobe, Japan. August.

Ashley, K.D., (1990). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. The MIT Press: Cambridge.

Carbonell, J.R. (1970). "AI in CAI: An Artificial Intelligence approach to Computer Aided Instruction". *IEEE Transactions on Man Machine Systems* 11(4) 190-202.

Chi, Michelene T.H., S. Silver, H. Jeong, T. Yamauchi, & R. Hausmann (2001) "Learning From Human Tutoring" in *Cognitive Science*, Vol. 25, pp. 471-533.

Chi, Bassock, Lewis, Reimann & Glaser (1989), "Self Explanations: How students study and use examples in learning to solve problems", *Cognitive Science, 5,145-182*

Collins, A. & Stevens, A. L. (1982). "Goals and Strategies of inquiry Teachers". In *Advances In Instructional Psychology*, R Glaser (ed.) pp. 65-119. Hillsdale, NJ: Erlbaum.

Retalis, S., H. Pain & M. Haggith. (1996) "Arguing with the Devil; Teaching in Controversial Domains". In *Intelligent Tutoring Systems, 3d Int'l Conf., ITS '96*, 659-667. Berlin: Springer.

Rose C. P., J. D. Moore, K. VanLehn, D. Allbritton. (2001) "A Comparative Evaluation of Socratic versus Didactic Tutoring", 2001 LRDC Tech Report LRDC-BEE-1.

Thompson, L., Gentner, D. and Loewenstein, J. (2000) "Avoiding Missed Opportunities in Managerial Life." in *Org. Behavior and Human Decision Proc.*, 82, No. 1. May. pp. 60-75.

Modeling Human Error in a Real-World Teamwork Environment

Stephen Deutsch (sdeutsch@bbn.com)
BBN Technologies, 10 Moulton Street
Cambridge, MA 02138 USA

Richard Pew (pew@bbn.com)
BBN Technologies, 10 Moulton Street
Cambridge, MA 02138 USA

Abstract

An initial model of human error in a real-world teamwork environment has been developed. The captain and first officer of a commercial aircraft and the air traffic controllers with whom they interact are modeled as the crew executes an approach and landing followed by taxi operations that take them to their assigned gate. Scenario details and human model development were based on the results of a series of full-task simulation experiments using commercial pilots as subjects. The focus of the experiment was on errors committed by the aircrews during taxi operations. The models developed exhibit the robust behaviors typically exhibited by aircrews and identify psychologically grounded windows for error within that robust behavior.

Human Error Modeling Applied to Taxi Operations

NASA Ames Research Center conducted two full-mission studies of airport taxi operations under low visibility and night conditions. The subject of the studies was the Taxiway Navigation and Situation Awareness (T-NASA) system, aircraft flight deck technology designed to improve commercial airport taxi operations in poor weather while maintaining a high degree of safety (Hooey, Foyle, & Andre, 2000). The T-NASA system includes a head-up display, a head-down electronic moving map display, and directional audio alerts. The studies included a series of baseline trials run without the T-NASA system and a series of trials using various configurations of the T-NASA system. The T-NASA system effectively eliminated very nearly all error, hence the focus of the human error modeling effort was on the baseline trials.

The NASA Ames Advanced Concept Flight Simulator (ACFS) used in the studies provided a generic glass cockpit simulator with a 180-degree field of view and a high fidelity rendering of Chicago O'Hare Airport replicating the airport layout including runways, taxiways, signage, painted markings, lights, concourses, and structures (Hooey & Foyle, 2001). In the first study, 16 two-pilot commercial crews each completed six land and taxi-to-gate trials based on current operations using Jeppesen charts for navigation. Half of the trials were under low visibility conditions with runway visual range (RVR) of 700 feet, and half under night visual meteorological conditions (VMC). In the second study, 18 commercial two-pilot crews each completed three land and taxi-to-gate trials based on current operating conditions under 1000 foot RVR conditions. In evaluating these studies, Hooey and Foyle (2001) defined navigation errors as taxiing on a portion of the airport surface on which the aircraft had not been cleared and deviating from their cleared centerline by at least 50 feet. Their analysis revealed 26 navigation errors in 150 current-operation trials—errors were committed on 17.3% of the trials.

Modeling Robust Nominal Performance as a Prelude to Modeling Error

As we set out to identify the sources of error (c.f., Deutsch & Pew, 2001) and then to model error in taxi procedures, we started by refining earlier Distributed Operator Model Architecture (D-OMAR) models (Deutsch, 1998; Deutsch & Adams, 1995) that captured the robustness in aircrew procedures. The models represent the multiple task behaviors of each player as the product of a mix of goals and procedures that operate concurrently to proactively address the player's agenda. Expectations integrate anticipated events while anticipated or unanticipated interruptions must be accommodated. Ongoing tasks determine their own execution times and run to completion unless another procedure defined as a competing procedure with greater priority intervenes. A mix of automatic and thoughtful behaviors are modeled without resorting to a central executive responsible for explicitly scheduling all future actions. A thoughtful cognitive act of decision-making is defined as just that, another procedure that determines the action to follow.

The NASA Information to Modelers package included a Nominal Task Sequence (NASA, 2001a) for the T-NASA baseline conditions. This was used as the basis for the development of the approach-and-landing and taxi procedures that the models of the captain, first officers, and air traffic controllers employed. Approach-and-landing is one of the busiest phases of flight, making high demands on the aircrew. In spite of the high demands of getting the aircraft safely on the ground, it is also the time at which the first steps in the subsequent taxi operations are initiated. The crew is in the process of approaching a given runway and already know the concourse and gate toward which they will be heading. Moreover, as specified in the Nominal Task Sequence, at about eleven miles out they discuss with the air traffic controller and among themselves which runway exit

they will take. As we will argue below, the crewmembers each now have in mind one and perhaps several taxi routes they might take to the gate. Once the runway exit information is in hand, the focus of attention returns to landing the aircraft and rollout.

The information provided in the Nominal Task Sequence was also used as the basis for the modeling of the subsequent landing and rollout sequence. As the rollout sequence is completed and the aircraft approaches the designated runway exit, the taxi sequence is initiated. The first officer provides information to the captain on their position relative to the preferred exit based on notes taken when the preferred exit was agreed on. He/she then informs the controller that the aircraft is clearing the runway, both crewmembers then switch their radio frequency, and the first officer contacts the ground controller. At this point, the ground controller provides the crew with the taxi route to the gate and the first officer writes down the taxi route.

It was at this point that we encountered the first instance of a requirement for a coping strategy. Many of the high-speed exits at O'Hare have a very short run to the first intersection and taxiway routings can be unusually lengthy. We encountered this first when modeling a landing on runway 9R using high-speed exit M6 with an immediate left turn onto taxiway M. The first officer was head-down writing out the taxi directives and was late in providing information to the captain on the upcoming immediate turn. At this point, the captain was also listening to the taxi routing and could go with what he/she heard or slow the aircraft and obtain confirmation on the upcoming turn from the first officer. The coping strategy that we modeled had the captain go ahead with the turn as heard and notify the first-officer of the turn as it was executed.

The process for each subsequent turn in the taxi sequence followed the same pattern. As a turn was completed, the first officer would consult his/her routing notes and the airport diagram, and then prompt the captain on the taxiway and direction for the upcoming turn. As expected, the modeled nominal process proved very robust. By simply changing the routing that the ground controller provided, the captain and first officer were able to execute any desired taxi routing. With these robust aircrew processes in place, the challenge was to model taxi sequences that produced errors consistent with those in the baseline T-NASA experiments.

As the captain and first officer meet their responsibilities during taxi operations, the inherent nature of the tasks that they perform provide them each with a different sense of their immediate location and their location with respect to their assigned taxi routing. They each achieve and maintain different levels of local and global situation awareness (Wickens & Prevett, 1995). If they are working well as a team, they will strive to fill each other's gaps in awareness. In building the aircrew models, we felt that it was essential to reflect this level of teamwork.

The captain was modeled as predominantly head-up during taxi operations. He/she announced each turn as it was executed to keep the first officer informed of their immediate location during such periods as the first officer might have been head-down. Meanwhile, the first officer, working with the airport diagram and written notes on the runway exit and taxiway routing provided the captain with a more global view of their taxiway routing than would have otherwise been available. The teamwork skills of the modeled aircrews were effective in repairing gaps in one another's situation awareness. One effect of providing this level of detail in good crew performance was of course to make the taxiway procedures just that much more robust and error that much less likely.

Making the Wrong Turn at an Intersection

The particular process that produced the errors of interest was the preparation for and execution of the next turn in the taxi sequence as governed by the captain. As modeled, the captain, if left to his/her own resources, must rely on his/her memory of the taxi sequence as conveyed by the ground controller as the aircraft cleared the landing runway. However, the captain gets significant support in this task from the first officer. The first officer takes notes on the taxi sequence as it is received from the ground controller and will, under nominal conditions, prompt the captain with the name of the next taxiway and the direction of the turn required.

Under nominal conditions, the first officer prompts the captain on the upcoming turn and one can reasonable expect that the captain will correctly act in accordance with the prompt. Acting counter to the prompt is an error possibility that we did not pursue. Hence, to find a source for making a turn error at an intersection, we had to construct reasonable scenarios in which the first officer was otherwise occupied and unable to provide the prompt in a timely manner and of course identify an underlying reason for a mistake on the part of the captain. The events that prevented the first officer from providing the prompts are discussed below in the sections providing details on the error scenarios. Here, we examine possible sources for the errors committed by the captain in executing the incorrect turns.

Intention-to-Act

A classical view of the taxiway process might be that, in approaching a turn, the captain has a planning problem whose resolution is then followed by plan execution. Do we in fact need to make a turn at the upcoming intersection and if so, which way? There might be a schema in place for executing the next turn with slots to be filled in for the name of the next taxiway and the direction to turn. In this view of the process, error might come about by incorrectly filling the slot for the next taxiway name, but more probably, the slot for the direction of the turn to make.

We would like to argue in favor of an alternate view in which there are typically several *intentions-to-act* concurrently in process. The intentions may be established at different points in time. One or more of them may lead to a correct turn or to making an error at the intersection. A *winner-take-all* process leads to the execution of one of the intentions-to-act and the correctness of the outcome is the product of the winning intention. At the point of execution, the remaining intentions cease to contend. We label the process intention-to-act to suggest that the process is not the

product of a conscious decision process—it is not a deliberative planning process followed by a plan execution process. There is the immediacy of an automatic, atomic process rather than a sequential process of planning and acting. Each of the intentions-to-act is instantiated with established slot values, rather than with unfilled slot values to be filled by a deliberative process.

Most of the time there is more than one intention-to-act. In the nominal case where the first officer has provided the correct prompt for the turn, the turn is, most likely, simply executed in response to the prompt and most likely in accordance with a pre-existing intention. In lieu of the prompt from the first officer, the captain will act on a pre-existing intention that might lead to the execution of his/her intention to turn or alternatively to pause and query the first officer on the next turn. (We have not pursued the case of the captain's slowing or stopping the aircraft and querying the first officer.) That is, most of the time in the taxi environment, it is reasonable to expect that the captain has an intention-to-act in place and ready to be acted on.

Rather than having a single planning process with slots to be filled from various sources that is followed by a plan execution step, there are multiple intentions-to-act with selection through a non-conscious winner-take-all process. Each of the intentions-to-act has a complete set of immediately filled slots. In the following section, we provide the reasoning supporting this viewpoint.

Intentions-to-Act as Automaticity

At this point, we want to build the case for the idea that in performing relatively simple tasks like correctly executing the next taxiway turn, there may be several competing intentions-to-act. Most may arise as automatic processes that require little or no conscious deliberative thought. They may emerge from different ongoing processes competing in a winner-take-all process to determine the action taken. Occasionally, the winner will determine an action that is in error. During the course of this study, we have attempted to identify some of the sources for these intentions and to provide reasoned explanations on why the errors emerge.

For most of us, there are a broad range of everyday activities that we perform quickly and effortlessly—they appear to be automatic and involve little thought or conscious awareness (Logan, 1988a; James, 1890). Logan (1988a) characterizes this automaticity, the execution of these activities, as fast, effortless, autonomous, stereotypic, and unavailable to conscious awareness. That is, we experience them as fast, effortless, stereotypic, and unavailable to conscious awareness. They are autonomous in the sense that the acquisition of these skills comes about independent of any deliberate intention to learn them.

Logan (1988a) developed the "Instance Theory of Automaticity," a theory for how automatization is constructed. The theory was developed in part through a series of experiments in learning alphabet arithmetic—learning to solve problems of the type "A+2=?" where the answer is "C." Initially, most people solve these problems by explicitly counting out the required steps through the alphabet—they employ an algorithm that they step through when the problem is presented. Through experience they "learn or remember" the answers.

Logan suggests that *each* learned instance is remembered. When presented with a new problem, there is a concurrent attempt to access a remembered instance of a previous solution *and* an explicit algorithm-based problem-solving computation. The memory access is a comparatively fast process, the algorithm-based process comparatively slow. If the memory access is successful in retrieving a solution, there will be a rapid response to the posed problem. If the memory access is not successful, the response will be slower. Through experience, more and more solutions are acquired and at some point, the deliberative process is simply not a contender in the winner-take-all process. For any given problem, there may be several remembered solutions. Due to the remembering of each solution instance, there may potentially be several correct retrievals. It is the one that is first retrieval that determines the time required to solve the problem.

Logan (1988b) further argues that the memory traces that support automaticity may well support declarative as well as procedural knowledge. Logan (1988b) suggests that we "look more broadly for automatic processes. They need not be restricted to procedural knowledge or perceptual-motor skill but may permeate the most intellectual activities in the application environment." Bargh and Chartrand (1999) further suggest that limits on conscious, intentional control requires that non-conscious processes support much of moment-to-moment psychological life. Here we are suggesting that the captain's procedures for addressing the next turn in the taxiway sequence may sometimes be automatic and that while these will often lead to correct behaviors, they may sometimes lead to errors such as those seen in the baseline T-NASA experiments.

Intentions-to-Act as a Source of Error

Our review of the NASA-provided data on the T-NASA experiments pointed to two important factors that we felt deserved particular attention in our modeling effort. NASA (2001c) identified the importance of the location of the destination gate and its relation to the taxi route. Five errors occurred in 48 instances of required turns *away from* the shortest route to the concourse gate while only seven errors occurred in 534 instances of turns *toward* the concourse gate. At any given intersection, the aircrews had a bias to turn toward their destination concourse gate. When the correct turn was one away from the concourse gate, there was a greater tendency toward making an error.

The second observation was the straightforward one that time pressure can lead to error. There was a greater chance of error when a second turn in the taxi sequence closely followed the previous turn. The time pressure of a second turn closely following a first turn was an important factor in each of the errors that we generated in the modeling effort.

To date, four sources of contending intentions-to-act have been identified and modeled. The first is episodic memory—a source for habit-based actions. Similar situations have been encountered in the past and we have a ready source of responses that have worked well. These are

responses that in the past have proven successful and are generally able to carry us through most of the activities of the day. When they fail this is what Reason (1990) refers to as "strong-but-wrong." In our particular case, the aircrews had a history of previous landings at Chicago O'Hare.

A second source of intention-to-act is context-based expectation, driven by partial knowledge. Explicit partial information provided in the current situation prompts a particular intention. Within the taxi-framework, the captain knows the location of the concourse gate and, based on this knowledge, may reasonably have an expectation that the next turn will take them on the shortest route to the gate. These particular situation-specific information points are sufficient to set up an intention for the next turn.

The third source of intention-to-act is the remembrance of the taxi sequence as provided by the ground controller when the aircraft exited the landing runway. As the aircraft approaches a turn, several minutes may have passed since the ground controller provided the taxi directive. The remembrance may or may not be correct, but it can be the source of an intention-to-act.

The fourth source of intention-to-act in the taxi-framework, and the best-grounded source of intention, is the explicit prompt by the first officer based on written notes on the taxi directives provided by the ground controller. In the nominal case, the first officer's prompt will match the captain's intention and will lead to error free performance.

We modeled the contention between these intentions as a winner-take-all process mediated by priority and explored the impact of varying the priorities of the contending intentions. Within the winner-take-all framework, at the winning intention's transition from intention to action, the remaining intentions cease to contend—within the framework of the model, the procedures that would have implemented those intentions fail. The occurrence and timing of the events that drive the intentions determine how they play out, producing successful behaviors or mistakes that lead to an incorrect turn on the taxiway. In particular, to provide a window for error to occur, it was necessary to set up realistic event chains that prevented the first officer from providing the prompt on the next turn to the captain.

As we have suggested, the team-based nature of the taxi procedures makes them very robust and the challenge has been to create situations in which mistakes will lead to error. This effort focused on two error sequences, each requiring two turns in rapid sequence. For case one, there were two instances of the same error as crews took high-speed exit M7 from runway 9R. At the first intersection after the high-speed exit, each captain turned left toward the concourse gate rather right away from the gate as directed by the ground controller. In the second case, there were two scenarios that shared a similar turn sequence: after turning onto taxiway F in the first instance and M2 in the second instance, there was a quick right turn onto taxiway B. In each of the scenarios, one of the captains turned left rather than right. The errors were noteworthy, because in committing the error the captains each turned *away* from their intended concourse gate rather than toward the gate as directed.

Error Driven by Partial Knowledge Our hypothesis is that the incorrect turn following the high-speed exit (see Figure 1) was driven by the captain's expectation that the shortest route to the gate was the route to be taken. (The small arrows that denote the errors in Figure 1 indicate the incorrect left turns taken just after the high-speed runway exit. They are in red when viewed in color—in grayscale, they may be difficult to make out.) The intention-to-act arose at the point of the early discussion of the runway exit with the approach controller and the first officer. At this point, the captain knew the runway exit and the concourse gate, and might reasonably have expected to turn left from the high-speed exit at taxiway M taking him/her toward the gate. It became one intention contending to be executed at the first turn after exiting the active runway. From Reason's (1990) perspective, this is an automatic retrieval process based on similarity-matching and frequency-gambling that opens a window for error.

As the scenario played out in the nominal case, the first officer completed the task of taking notes on the taxiway

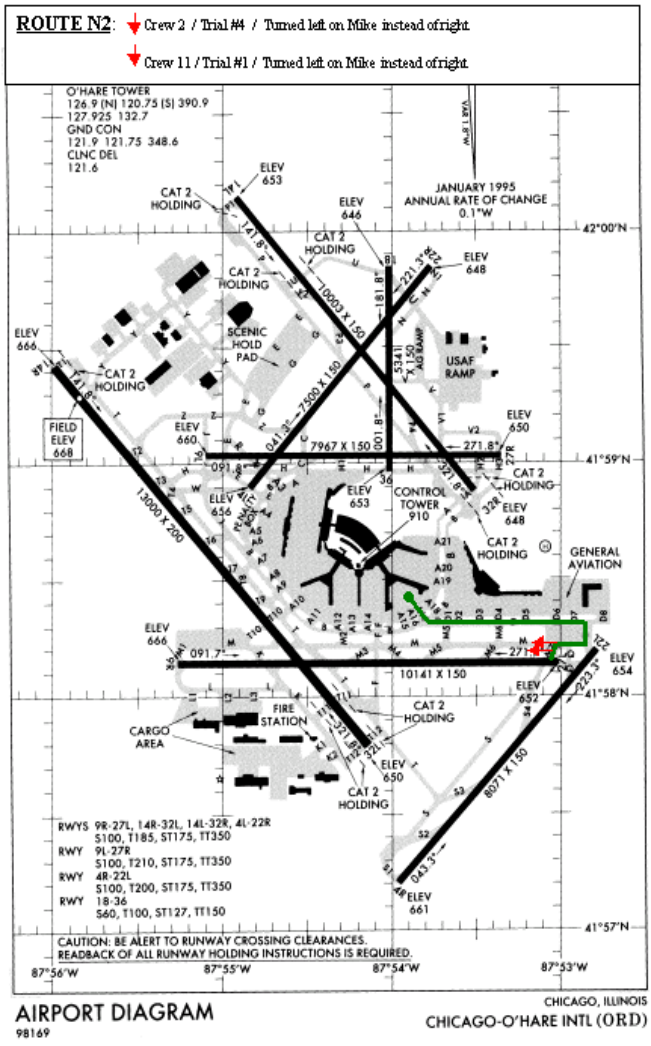


Figure 1: Errors driven by partial knowledge (NASA 2001b).

sequence and then prompted the captain on the first turn following the runway exit. The first officer's prompt triggered a new, contending intention-to-act on the captain's part. The new intention may or may not have been consistent with preexisting contentions. In the nominal case, it dominated and the captain turned right correctly. Given the correct prompt by the first officer, we deemed it highly unlikely that the captain would incorrectly execute the turn.

To open a window for an error to occur, it was necessary to construct a situation that reasonably occupied the first officer, preventing him/her from providing the captain with the explicit prompt on the upcoming turn. The very short run to the first turn after the high-speed exit was the essential factor. The first officer was already busy taking notes on the taxiway routing. Indeed, in some scenarios the taxiway routing was so lengthy that in the nominal case the first officer was still taking notes as the first turn was executed. In this scenario, this was not the case, hence a "mistake" was needed to additionally task the first officer. The failure to preset the radio frequency for the transfer to the ground controller provided the delay. The few seconds necessary to set the new radio frequency provided enough delay to prevent the first officer from prompting the captain before the turn. This was a mistake on the part of the aircrew in the sense that it is always incumbent upon them to complete an action at the earliest available time, rather than risk a situation such as this in which there are contending tasks in process.

Let us recap the captain's intentions-to-act as the aircraft approached the first turn onto taxiway M after the high-speed exit on taxiway M7. The first officer has been otherwise occupied and has not provided the captain with the explicit prompt on the upcoming turn. Based on the coping strategy described earlier, the captain might have a correct intention-to-act based on having attended to the ground controller's taxi directive and an incorrect intention based on the expectation of receiving a shortest route to the concourse gate. Much of the time the coping strategy might be expected to win the winner-take-all competition and lead to a correct turn—some of the time the expectation-based intention-to-act might be acted upon, leading to a taxiway error. Hence, a reasonable, grounded source for an error consistent with the T-NASA experiments has been identified and modeled.

Error Driven by Habit The second scenario examined the surprising cases in which an aircrew incorrectly turned away from the shortest course to the gate (see Figure 2). (In Figure 2, the small arrow denoting the error indicates the incorrect left turn taken just after the short north-bound segments near the center of the airport diagrams. It is in red when viewed in color—in grayscale, it may be difficult to make out.) The basic intention to take the shortest route to the gate would have led to the correct behavior, yet it was not the one acted upon. There were two instances of this error at similar intersections. In the first case (Figure 2), the aircraft was proceeding north on taxiway F and had been instructed to turn right onto taxiway B, but the captain turned left instead. In the second case, the aircraft was proceeding north on taxiway M2 and had been instructed to turn right onto taxiway B, but the captain turned left. We speculated that a crew whose company gates were on the opposite side of the airport from those required by the scenario might incorrectly turn toward their company gates, exhibiting an error based on long established habit. Requiring an aircrew to proceed to a gate opposite in direction from their company gates might be considered an artifact of the particular scenario, but in a commercial air travel environment that has seen many company failures and mergers, it is not uncommon for aircrews to find themselves working for new companies with new gate locations.

The turn at which the errors occurred closely followed a previous turn, creating a time-pressured situation. Once again, we manipulated the situation such that the first officer was not able to provide a timely prompt to the captain on the upcoming turn. Conflicting taxiway traffic was present on the first officer's side of the aircraft during the approach to the first turn. The first officer informed the captain of the

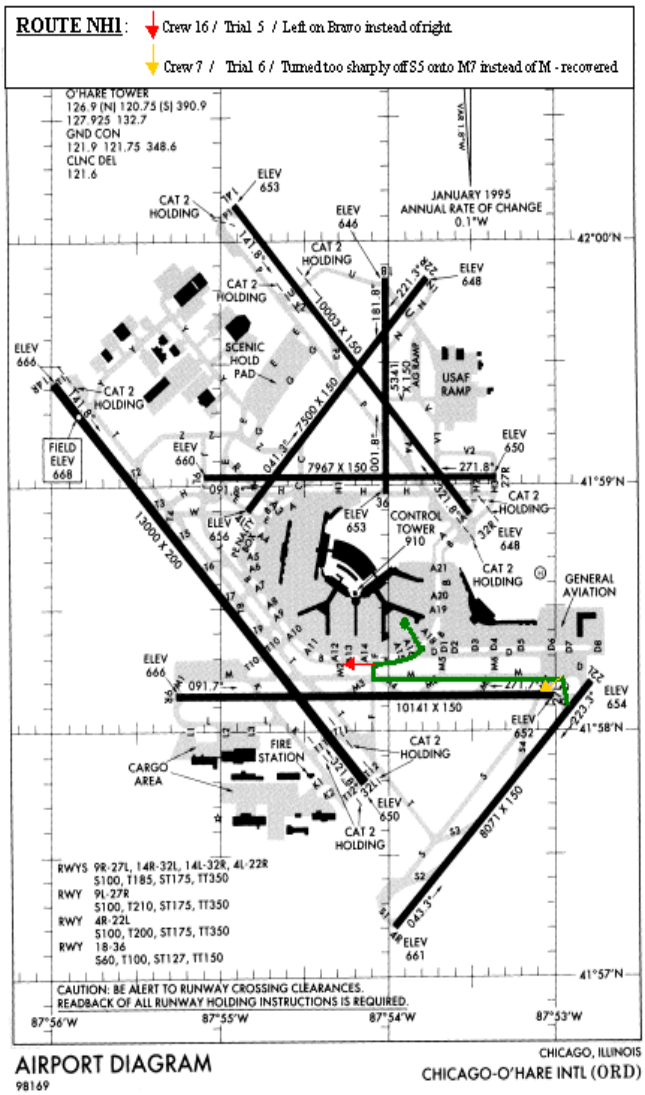


Figure 2: Error driven by habit (NASA, 2001b).

presence of the traffic and continued to monitor the other aircraft. Consequently, the first officer was delayed in going head-down to review his/her notes on the upcoming taxiway turn and checking the airport diagram. Following the delay, the first officer's prompt on the upcoming turn was immediately interrupted by a message from the ground controller directing the other aircraft to hold short of the upcoming intersection, allowing the first aircraft to proceed with the turn. Very slight changes in timing of the interruption would have opened the window for a timely and successful prompt.

In the absence of the prompt, there were still multiple intentions-to-act. As modeled, there were intentions-to-act based on the remembrance of the ground controller's taxi directive and on habit based in episodic memory. When the captain's habit-based intention-to-act won the winner-take-all competition and was acted upon, the error was committed. An informal post hoc analysis of the human subject trial error provided support for the speculation that the model represented (B. Hooey, personal communication).

Heuristically Guided Search of the Error Space

The incidence of error in the current-equipment T-NASA experiments was strikingly high when compared to the typical behaviors of professional aircrews. In general, the low frequency of mistakes and the even lower frequency of mistakes combining to produce errors renders a simple stochastic exploration of the behaviors space impractical. The robustness of aircrew team procedures that employ checking and cross-checking of critical actions means that most mistakes will be caught, further compounding the search task. Estimating error frequency for error types can also be a problem. The frequency of some errors (e.g., discrimination of taxiway signage) might be reasonably estimated; the frequency of others (e.g., the onset of a particular intention-to-act) is more difficult.

Timing is also critical. Very small variations in timing can open or close the window in which an error might occur. Timing was particularly critical in the scenario in which the habit-based error occurred. The combination of the demand on the part of the first officer for head-up time to monitor the approaching traffic and the precise moment of the ground controller's interruption of the first officer's prompt for the upcoming turn was necessary to open the window to error. It might well have been possible to generate many hundreds of runs, slightly varying several of the timings, and never have produced a single habit-based error.

To address this problem, we have employed a heuristically guided search of the space in which forced sequences of mistakes are generated, looking for those that lead to error. The errors produced to date are initial examples of the product of such a process. We have identified several novel potential sources of mistakes and worked to create situations in which they might reasonably be expected to occur. We have taken advantage of the time pressure inherent in the closely spaced turn sequences to manipulate the timing of events to construct sequences of mistakes that do in fact lead to error. For the present, this heuristically guided exploration of the error space has been manipulated by hand. In the future, we would like to move toward a more automated exploration of the error space.

Acknowledgments

The research reported on here was funded by the NASA Aviation Safety Program Human Error Modeling Element. The D-OMAR human modeling research effort is supported by the Sustainment Logistics Branch of the Air Force Research Laboratory.

References

Bargh, J. A, & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462-479.

Deutsch, S. E. (1998). Interdisciplinary foundations for multiple-task human performance modeling in OMAR. *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society,* Madison, WI.

Deutsch, S. E., & Adams, M. J. (1995). The operator model architecture and its psychological framework. *Proceedings of the 6th IFAC Symposium on Man-Machine Systems.* Cambridge, MA.

Deutsch, S. E., & Pew, R. W. (2001). *Modeling Human Error in D-OMAR.* (Tech. Rep. BBN-8328). Cambridge, MA: BBN Technologies.

Hooey, B. L., & Foyle, D. C. (2001). A post-hoc analysis of navigation errors during surface operations: Identification of contributing factors and mitigating strategies. In *Proceedings of the 11th Symposium on Aviation Psychology.* Columbus, OH: Ohio State University.

Hooey, B. L., Foyle, D. C., & Andre, A. D. (2000). Integration of cockpit displays for surface operations: The final stage of a human-centered design approach. In *Proceedings of the AIAA/SAE World Aviation Congress*. Warrendale, PA: SAE International.

James, W. (1890). *Principles of Psychology.* New York: Holt.

Logan, G. D. (1988a). Toward an instance theory of automatization. *Psychological Review, 95,* 492-527.

Logan, G. D. (1988b). Automaticity, resources, and memory: Theoretical controversies and practical implications. *Human Factors, 30,* 583-598.

NASA. (2001a). *Nominal task sequence.* NASA Working Paper. Moffett Field, CA: NASA ARC.

NASA. (2001b). *Routes and errors: Locations and types.* NASA Working Paper. Moffett Field, CA: NASA ARC.

NASA. (2001c). *Enhanced descriptions of off-route navigation errors.* NASA Working Paper. Moffett Field, CA: NASA ARC.

Reason, J. (1990). *Human Error.* Cambridge: Cambridge University Press.

Wickens, C. D., & Prevett, T. T. (1995). Exploring the dimensions of egocentricity in aircraft navigation displays. *Journal of Experimental Psychology: Applied, 1,* 110-135.

The Quality of Test Context and Contra-evidence as a Moderating Factor in the Belief Revision Process

Kristien Dieussaert (kristien.dieussaert@psy.kuleuven.ac.be)
University of Leuven, Department of Psychology
Tiensestraat 102, Leuven, Belgium

Walter Schaeken (walter.schaeken@psy.kuleuven.ac.be)
University of Leuven, Department of Psychology
Tiensestraat 102, Leuven, Belgium

Gery d'ydewalle (gery.dydewalle@psy.kuleuven.ac.be)
University of Leuven, Department of Psychology
Tiensestraat 102, Leuven, Belgium

Abstract

In this study, we describe the influences of qualitative changes to the reasoning problem on the reasoning process. The first manipulation is the quality of the test context: A rule is learned in a certain context and contradicted in another. The belief in the rule is then measured in the learning context, the contradictory context, and a new context. The second manipulation is the quality of contradiction: The contradictory rule can neutralize or inverse the learned rule. Both qualitative changes influence the belief revision process.

Introduction

Research in Artificial Intelligence is often conducted to develop systems that think/act rationally (or like humans, depending on the approach). Minsky (1975) was one of the first to point out the problem with deductive systems, and from then on, several researchers developed non-monotonic reasoning systems (for an overview, see Brewka, Dix, & Konolidge, 1997).

While these systems can be very interesting from the viewpoint of an engineer, Rips (1994) mentions two main reasons why these nonmonotonic logics are less than ideal for cognitive purposes: They do not lend themselves to simple implementations (higher order logics are incomplete) and they do not seem to reflect the deliberations that actually underlie human reasoning with defaults. He (1994, p. 299) stated: "For purposes of philosophy (and perhaps AI), we need normative rules about how to change our minds in the face of conflicting evidence; for purposes of psychology, we also need descriptive information on how people deal with the same sort of conflict."

Elio and Pelletier (1997) wrote a pioneering article on this topic. In their experiments, they first presented participants with a conditional premise and a categorical premise that affirmed the antecedent [or denied the consequent]. Then, they added a third piece of information, which conflicted with the conditional and the categorical premise. This piece of information was a categorical premise denying the consequent [or affirming the antecedent]. The three pieces of information together are in contradiction with the valid Modus Ponens [Modus Tollens] argument: If A then B, A, thus B [If A then B, not-B thus not-A].

Participants were asked to resolve the contradiction by rejecting one of the two first pieces of information. Elio and Pelletier (1997) observed that participants chose to disbelieve the conditional premise rather than the categorical one when resolving the contradiction.

Their results were refined by Dieussaert, Schaeken, De Neys, and d'Ydewalle (2000) and by Politzer and Carles (2001). They found that the initial belief in the conditional premise influenced the belief revision choice that participants prefer to make. When participants had a strong belief in the conditional premise, the preferred to reject (doubt) the categorical premise, when the conflicting information is added. When participants had a weak belief in the conditional premise, the results of Elio and Pelletier (1997) were confirmed: In this case, participants prefer to reject (doubt) the conditional premise.

This shows that it is important in belief revision research to be aware of the belief state participants hold before conflicting information is presented. Therefore, we conducted a measure of the initial belief state in the following experiments, before adding conflicting information.

The following experiments were inspired by research in the field of conditioning. Among others, Bouton (e.g., 1988, 1994) showed that the extinction of behavior does not necessarily means the rejection of a learned rule. One of the phenomena that confirm this hypothesis is 'renewal': When a behavior is learned in context A, and extinguished in context B, the behavior might show up again with a new test in context A. Bouton explains this phenomenon as follows: Individuals learn a dominant rule and exceptions to this rule in certain contexts (see also: Holyoak, Koh, &

Nisbett, 1989). As a consequence, a stimulus that has lost his value as a reinforcer, becomes an ambiguous stimulus from which the specific value is determined by the context. In other words, Bouton points to the importance of the presence of a certain context as an indicator of the belief in the (conditional) rule when an individual is confronted with conflicting information to that rule. In the forthcoming experiments, we will also work with context embedded situations, to gain more insight in the relative value of the contra-evidence.

Another uniqueness of the present experiments is that the participants' belief state is measured by the behavior they pose, and not by what they say their belief is. We think it is as important to know how individuals act on their beliefs, as it is to know how they describe their beliefs. Moreover, De Neys, Dieussaert, Schaeken, and d'Ydewalle (2000) found a strong correspondence between what participants say about their beliefs and how they actually act upon these beliefs, in experiments comparable with the ones presented below.

Briefly, on a technical level, the following experiments are designed as follows: first, the initial belief in a rule is tested. Then, contradictory evidence is presented in a different context, and finally, the belief in the rule is again evaluated by examining participants' manifested behaviours. On the content level, the experiments describe the influence of qualitative changes to the reasoning problem on the belief revision process.

Experiment 1: Pilot study

In this part, we describe a study that tests whether formerly used instantiations of the quality of contra-evidence were well chosen.

Before that, we explain the two hypotheses that are at issue in this manuscript and we depict the important role that the pilot study has regarding the second hypothesis. Subsequently, we give a brief overview of the content of the experiments.

Two hypotheses are evaluated in this manuscript. The main hypothesis is that when a conditional rule is acquired in a first context (A) and contradicted in a second context (B), the belief in the conditional is not affected by the contradictory information when tested in the first context (A) or in a new context (C). Indeed, if the rule acquired in the first context really is a dominant rule, and if the conflicting information learned in a second context is perceived as exceptional, one should also apply the dominant rule in a new context.

A second hypothesis is that the quality of the contradictory information plays a role in the belief revision process. Bouton (1988) focused on the *extinction* of learned rules. With extinction, a learned rule is often only neutralized (e.g. food – no food). Our hypothesis is that when the conflicting information has a stronger impact (compare with: food – shock), this may affect the belief revision process in a different way. More precisely, we hypothesize that when the contradictory information merely neutralizes the acquired conditional rule, its effect on the belief revision process is smaller than when it reverses the acquired conditional rule. Or, in other words, we suppose that when the acquired rule is reversed, this might affect the final belief state tested in context A and C anyway, despite the fact of a buffering context element (B) when the contradictory information is provided.

It is made very clear in the fore mentioned examples that the absence of an appetitive stimulus (neutralizing) is not the same as the presence of an aversive stimulus (reversing). An adequate test of the second hypothesis is only possible if both stimuli are well chosen.

In a former series of experiments (Dieussaert, Schaeken, & d'Ydewalle, 2002), the manipulation of the quality of the contra-evidence did not result in significant differences, contrary to our second hypothesis. Before drawing any theoretical consequences from this finding, we examine whether those instantiations were appropriately chosen. The pilot study (Experiment 1) was set up to test this hypothesis.

To convey a good understanding of this pilot study and its consequences possible, we briefly describe the content of the former experiments (Dieussaert et al., 2002). The motivation behind the experiment is related to that behind the conditioning experiments of Pineño, Ortega, and Matute (2000). Participants are in a war area, leading a rescue mission. They are told they should rescue as many refugees as possible from a building and they can do this by loading the refugees on a truck. Importantly, they should only fill the truck with refugees if the road is free of mines. They can learn whether the road is safe because coloured lights indicate it. Therefore, they should learn as fast as possible the meaning of these lights.

The participants learn the meaning of the lights in a first location (context), and then move to a second location where they learn that the meaning of some lights is reversed or neutralized. For example, a green light in the first location might indicate that the road is safe, while in the second location it indicates that the road is mined (reversal) or that no information about the road is available (neutralizing). Finally, they move to a third location. At this location, identical or different from one of the former contexts, the participants' belief regarding the meaning of the lights is examined.

In this pilot study (Experiment 1), we manipulate the feedback that participants received in the test phase. Three forms of feedback were distinguished: clearly

positive feedback, clearly negative feedback, and the feedback 'no information available'. The hypothesis is that participants interpret the latter feedback as negative.

Method

Participants. The 36 participants are candidate students at the University of Leuven, Department of Psychology and they participated as a partial fulfillment of a course requirement. Each student was randomly attributed to one of two groups (12 per group).

Design One variable was manipulated between subjects: Feedback. The three levels of this variable were manipulated in the test phase.

The dependent variable is the number of times participants press on the space bar during a fixed time interval when the cue light appears. The meaning of a space bar press is that a person is put in the truck. Each press is thus equivalent to the saving of one person. This dependent variable is measured in the three phases of the experiment: the confirmation phase (9 blocks), the contradiction phase (9 blocks) and the test phase (4 blocks). This behavioral measure is taken as a measure of belief state. The idea behind it is that the more the participant believes that the road is free of mines, the more persons (s)he will put in the truck.

Each block consists of four trials. A trial lasts four seconds. During these four seconds, a colored light is shown (green, blue, red or white). The sequence of the lights is randomized within each block. Participants learn the meaning of these lights (see below) by pressing the spacebar. One of these lights is the cue light, one is the neutral light and the other two do not have a fixed meaning. Each light's meaning is counterbalanced between participants.

The first eight blocks are considered as learning blocks; they are not included in the analysis. All participants acquired a constant pressing level for the cue light within these blocks. The cue light indicates a safe road. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, safe, or closed. The context is a location, which we refer to as context A.

The confirmation phase consists of nine blocks. The meaning of the lights is equal as in the learning blocks. The context does not change either.

The contradiction phase consists of nine blocks. Here, the cue light indicates a mined road. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, safe or closed. The context is a location, which we refer to as context B.

The test phase consists of four blocks. The test phase always takes place in the same context as the confirmation phase (A), but the feedback varies between groups. For the ABA/+/ group, the message 'n persons saved' is displayed after the cue light lit up in the test phase, while for the ABA/-/ group and the ABA group, the messages are 'n persons died' and 'no information available', respectively. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, safe, or closed.

Procedure The experiment was carried out individually, on computer. All participants received the same instructions. They were told to imagine that they were in charge of a rescue mission. Refugees were hiding in a building that could be attacked by the enemy. Furthermore, these people can only be rescued safely, by placing them in a truck, if the road is free of mines. People are loaded in the truck by pressing the spacebar. To know whether the road is safe, the participants have to learn the meaning of the lights that appear on each trial. Some of the four lights have a fixed meaning (road safe, road mined, road closed); other lights are distracters and have no fixed meaning. The feedback they receive is 'n persons saved', 'n persons died', or 'road closed'.

Moreover, sometimes no information is available about the situation of the road, so that the participants do not know whether the persons they placed in the truck were saved. It was stressed that they should only press the spacebar continuously in case they were very sure that the people would be saved in that trial.

Finally, they were told that different rescue missions could take place and that a beep sound signaled the start of a new mission. The importance of keeping in mind the location (context) where the rescue mission took place was stressed. They were told that the lights could have different meanings in different locations.

Results

Table 1 shows the mean number of space bar presses, (i.e., persons put in the truck) within a fixed time interval. Neither in the confirmation phase, nor in the contradiction phase, did we observe any differences between the groups.

Table 1: Mean number of presses per phase per group.

	CONFIRMATION PHASE	CONTRADICTION PHASE	TEST PHASE
ABA/-/	34.9	7.1	14.9
ABA	45.5	2.5	27.5
ABA/+/	42.6	4.9	39.1

An ANOVA shows a main effect of Feedback ($F(2,33) = 7.80$, $p<.005$). The ABA group scores higher in the test phase than the ABA/-/ group ($F(1,33) = 4.20$, $p<.05$), but does not differ from the ABA/+/ group. The latter group differs strongly from the ABA/-/ group ($F(1,33) = 15.58$, $p<.0005$).

A more detailed planned comparison analysis on the four blocks of the test phase moderates the picture. The three groups do not differ from each other in the first block. The ABA group differs significantly from the ABA/-/ group in the second block (36.7 vs.9.4; $F(1,33) = 14.24$), $p<.001$), but differs significantly from the ABA/+/ group in the third and fourth block of the test phase (17.8 vs. 42; $F(1,33) = 10.54$, $p<.005$ and 18.7 vs.36.3; $F(1,33) = 5.51$, $p<.05$, respectively). For these last two blocks, no difference with the ABA/-/ group could be observed.

Discussion

The feedback message 'no information available' does not function well as a neutralizer of a conditional rule. The message is not interpreted as the absence of an appetitive stimulus, but rather as the presence of an aversive stimulus. Indeed, after a few trials the effect of this feedback message does not differ from other negative feedback, such as the message 'n persons died'.

Experiment 2

Now that we have shown that the message 'no information available' is interpreted as negative rather than as neutral feedback, we have good reasons to believe that it was not appropriate to use this message as an instantiation of the neutralizing level as was done in former experiments (see: Dieussaert et al., 2002).

In the present experiment (Experiment 2), we test the two hypotheses at issue, and focus on the more appropriate qualitative manipulations of the contra-evidence. Our main hypothesis is that a conditional rule acquired in a first context functions as a dominant rule. Conflicting information learned in a second context is perceived as exceptional. Therefore, one should also apply the dominant rule in a new context.

Our second hypothesis is that the quality of the contradictory information plays a role in the belief revision process. More precisely, we suppose that when the acquired rule is reversed by contradictory evidence, this might nonetheless affect the final belief state in both context A and context C, despite the existence of the buffering context element (B).

Method

Participants Seventy candidate students at the University of Leuven, Department of Psychology, participated in this experiment. They did not participate in the former experiment and participated as a partial fulfillment of a course requirement. Each student was randomly assigned to one of six groups (11 in the ABB- and ABC-Weak group; 12 in the other groups).

Design The independent variables, Test Context and Contradiction Level, were manipulated between subjects. We distinguish three instances of the variable Test Context: The context in the test phase can be the same as the context in the confirmation phase (ABA), it can be the same as the context in the contradiction phase (ABB) or it can differ from both (ABC).

The second independent variable, Contradiction Level, consists of two instances: The contradiction can be Strong, when the meaning of the rule is reversed, or can be Weak, when the meaning of the rule is neutralized.

The dependent variable is the same as in Experiment 1: The number of times participants press on the space bar during a fixed time interval when the cue light appears. Again, this variable is measured in the three phases of the experiment: the confirmation phase (9 blocks), the contradiction phase (9 blocks) and the test phase (4 blocks). This behavioral measure is taken as a measure of belief state.

Each block consists of four trials. A trial lasts four seconds. During these four seconds, a colored light is shown (green, blue, red, or white). The sequence of the lights is randomized within each block. Participants learn the meaning of these lights (see below) by pressing the spacebar. One of these lights is the cue light, one is the neutral light and the other two do not have a fixed meaning. Each lights' meaning is counterbalanced over the participants.

The first eight blocks are considered as learning blocks; they are not included in the analysis. All participants acquired a constant pressing level for the cue light within these blocks. The cue light indicates a safe road. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, save, or closed. The context is a location, which we refer to as context A.

The confirmation phase consists of nine blocks. The meaning of the lights is equal as in the learning blocks. The context does not change either.

The contradiction phase consists of nine blocks. Here, the cue light indicates a mined road in the Strong contradiction level. In the Weak contradiction level, the message 'road closed' is displayed with a cue light. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, save or closed. The context is a location, which we refer to as context B.

The test phase consists of four blocks. In this phase, simultaneously with the cue light, the message 'no information is available' is displayed. The neutral light indicates that the road is closed. The meaning of the other lights differs randomly over blocks; they can indicate that the road is mined, save, or closed. The context varies over groups: It can be the same as in the confirmation phase (A), the same as in the contradiction phase (B) or different from those two (C).

Procedure See Pilot study (Experiment 1)

Results

No differences between the groups could be observed in the confirmation phase. This was expected since at that moment no manipulation was introduced yet. In the contradiction phase, the Strong groups scored lower than the Weak groups, as we hypothesized (2.4 vs.7.6; $F(1,64) = 7.6$, $p<.01$). Within the same level of contradiction, no differences were observed. Table 2 shows the means for each group.

Table 2: Mean number of presses per phase per group.

		CONFIR-MATION PHASE	CONTRA DICTION PHASE	TEST PHASE
ABA	Strong	45.5	2.5	27.5
ABB	Strong	41.2	1.8	2.1
ABC	Strong	41.0	2.8	2.5
ABA	Weak	44.0	10.9	40.4
ABB	Weak	44.0	8.3	7.3
ABC	Weak	46.2	3.6	2.9

An ANOVA shows a main effect of Test Context ($F(2,64)=73.21$; $p<.00001$). The score of participants from the ABA group (33.9) is higher than the score of participants from the ABC group (2.69; $F(1,64) = 118.36$, $p<.00001$) and the ABB group (4.70; $F(1,64) = 98.65$, $p<.00001$).

The manipulation of Contradiction Level resulted in a main difference between both levels (10.69 vs. 16.87 for the Strong and Weak contradictions respectively; $F(1,64) = 6.71$, $p<.05$) in the expected direction. Neither the ABB group, nor the ABC group has a different score in the Strong and Weak group. The main effect results from the difference in the ABA Strong and Weak group ($F(1,64) = 10.05$, $p<.005$).

The ABA Weak score in the test phase does not differ from it's score in the confirmation phase, while the ABA Strong score does ($F(1,64) = 19.79$, $p<.00005$).

Discussion

The effect of Contradiction Level is now adequately obtained due to an appropriate choice of the instantiation of the Weak level. Only the groups where learning and test context are the same, are affected by the Contradiction Level manipulation: When the contradiction merely neutralizes (i.e., does not reverse) a previously learned rule, no belief revision takes place. When the contradiction is strong, the belief in the rule decreases compared to the learning phase. This observation is surprising given that the contradiction takes place in another context, and could therefore be easily neglected.

On a theoretical level, it would be tempting to conclude that the rule acquired in context A can be considered as a dominant rule. One could state that the first and the second hypothesis are thus confirmed: The belief is not influenced by the contradictory information in context B, when this conflicting information neutralizes the learned information, but it is influenced when the conflicting information reverses the learned information.

However, this conclusion would be false because for each Contradiction level, the belief in the rule remains low when it is tested in a new context (C). If the 'dominant rule' hypothesis were correct, the belief in the new context should be as high as in the confirmation phase, at least for the Weak level group.

General Discussion

We manipulated two qualitative factors, Test Context and Contradiction Level, to test their influence on the belief revision process. Regarding Test Context, we hypothesize that when a conditional rule is acquired in a first context (A) and contradicted in a second context (B), the belief in the conditional is not affected by the contradictory information when subsequently tested in the first context (A) or in a new context (C).

This hypothesis was inspired by Bouton (1988, 1994), who stated that an acquired rule is protected against extinction when the conflicting information is presented in another context than the learning context, because individuals interpret the first rule as the dominant rule and other one as the exception.

Regarding Contradiction level, the hypothesis was that the quality of the contradictory information plays a role in the belief revision process. We supposed that when the acquired rule is reversed, this might affect the final belief state tested in context A and C anyway, despite the buffering context (B) of the contradictory information.

The influence of the variable Test Context is clear: Given the same learning (and confirmation) and test context, less belief revision takes place after contradictory information is presented in a different context, than when learning and test contexts differ.

The influence of the variable Contradiction Level is also clear. Given an appropriate choice of the absence of an appetitive stimulus (road closed) and of the presence of an aversive stimulus (road mined), the manipulation resulted in a significant difference between the Weak and the Strong level. This effect was mainly due to the effect in the ABA group. For this group, no belief revision took place when the contradiction neutralized the learned rule, but belief revision was observed when the contradiction inversed the learned rule.

Although this effect on the ABA group confirms both hypotheses, one should not be too optimistic about the theoretical consequences because no such effect was

found for the ABC group, which plays an important role as a control group.

Therefore, the results may put serious question marks on the theoretical translation into 'dominant' and 'exception' rules. The results of these experiments may indicate that different rules are learned within each context.

However, from an economical viewpoint on the learning process, this idea does not seem very fruitful. Learning and applying conditional rules would become a very heavy task if every new context implied a new rule.

An interesting distinction, that was not made thus far, could be between conditional rules that are learned (and mostly applied) in one restricted context and conditional rules that are learned (and applied) in a wide range of different contexts.

Our experiments investigated the former situation. It might be that in the latter situation rules that are applicable in more contexts may become dominant rules and rules that are only applicable in one or a few contexts may become exceptional rules. Surely, this idea needs supplementary testing.

In sum, the results of these experiments indicate that the belief revision process is influenced by various qualitative characteristics of the problems. More precisely, this study shows the effect of contradictory context and the level of contradiction. Although the theoretical consequences of these results are not clear for the time being, some handouts for further testing are provided.

Sharing the considerations of Rips (1994) regarding human reasoning, we conclude with an amendment to his statement: For purposes of psychology, we need more descriptive information *and theoretical explanations* on how people deal with conflicting information.

Acknowledgments

This study was carried out with support from the IUAP/PAI P4 and from the Fund for Scientific Research Flanders Project G.0239.02 .

We wish to thank the reviewers and Uri Hasson for their helpful comments.

References

Bouton, M. E. (1988). Context and ambiguity in the extinction of emotional learning: implication for exposure therapy. *Behavior Research and Therapy, 26*, 137-149.

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes, 20*, 219-231.

Brewka, G., Dix, J., & Konolidge, K. (1997). *Nonmonotonic reasoning: an overview*. Stanford: CSLI Lecture Notes 73.

De Neys, W., Dieussaert, K., Schaeken, W. & d'Ydewalle, G. (2000). *Assessing belief revision in a game context*. Poster presentation, 4th International Conference on Thinking, Durham.

Dieussaert, K., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2000). Initial belief state as predictor of belief revision. *Current Psychology of Cognition*, 19, 277-288.

Dieussaert, K., Schaeken, W., & d'Ydewalle, G. (2002). *A study of the belief revision process: The value of context and contra-evidence* (Tech. Rep. N°291). Leuven, Belgium: University of Leuven, Laboratory of Experimental Psychology.

Elio, R., & Pelletier, F. (1997). Belief change as propositional update. *Cognitive Science, 21 (4)* , 419-460.

Holyoak, K. J., Koh, K., & Nisbett, R.E. (1989). A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review, 96*, 315-340.

Minsky, M. L. (1975). A framework for representing knowledge. In Winston, P.H. (ed.), *The psychology of computer vision* (pp. 211-277). NY: McGraw-Hill.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking and Reasoning, 7(3)*, 217-234.

Pineño, O., Ortega, N., and Matute, H. (2000). The relative activation of associations modulates interference between elementally trained cues. *Learning and Motivation, 31 (2),* 128-152.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: Routledge.

The Role of Analogy in Teaching Middle-School Mathematics

Lindsey K. Engle (Lengle@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Keith J. Holyoak (Holyoak@lifesci.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

James W. Stigler (Stigler@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Abstract

Analogies produced in twenty-five US eighth-grade mathematics classroom lessons were analyzed according to their frequency and structure. Frequency findings suggest that analogies are a common part of mathematics classroom learning, and a component analysis revealed regular structural patterns in the way these analogies are produced. Teachers tended to organize the analogies by producing the target, source, and mapping steps before students become active participants. Students were most likely to then make inferences, adapt them to the target context, and solve target problems. Student participation was either independent or co-constructed with a teacher or other students. Findings address an important correlate with experimental research on analogical reasoning.

The United States' educational system is presently struggling to find teaching programs that facilitate mathematical understanding that goes beyond algorithmic knowledge. Standardized test results recently released in California indicate that despite improvements, the educational system remains far below state goals in mathematics (STAR, 2001). One major component of this difficulty is a lack of knowledge about how to teach abstract concepts so that students are able to transfer this learning across contexts.

Systematic use of analogy may be integral to a teaching program that meets that goal. Analogy is a comparative structure that highlights abstract structural relations (Gentner, 1983), and facilitates schema acquisition and transfer across problems (Gick & Holyoak, 1983). Learning mathematics requires development of generalized concept representations that can be applied across contexts (Bransford, Brown & Cocking, 1999).

Cognitive scientists have argued for decades that analogy plays a central role in human cognition, learning and problem solving (e.g., Holyoak, Gentner & Kokinov, 2001; Kolodner, 1997; Holyoak & Thagard, 1995; Gentner & Toupin, 1986; Piaget, 1950). However, there have been paradoxical findings in analogy research. While analogy has been demonstrated to be used in several everyday contexts (e.g. Dunbar 1995, 2001), most laboratory studies show low rates of spontaneous noticing and use of analogies for problem solving (e.g. Gick & Holyoak, 1980, 1983). It is necessary to understand this discrepancy between observed patterns of analogical reasoning in the laboratory and in everyday contexts, termed the analogical paradox (Dunbar, 2001), in order to design meaningful interventions to promote educational usage of analogy.

We suggest that in order to clarify the paradoxical findings concerning analogy use, detailed analysis of everyday analogy usage is essential, because important aspects of analogy use can only be understood through online analysis of the pragmatics governing analogy production. The current study uses discourse analytic techniques to explore analogy production in the context of teaching mathematics in eighth-grade mathematics classrooms.

Methods

Sample and Coding

Twenty-five videotaped eighth-grade mathematics lessons were analyzed to examine analogy activities. The lessons were randomly selected from a larger random probability sample collected as part of the Third International Mathematics and Science Study (TIMSS) directed by Jim Stigler (see Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1998). All selected classrooms were videotaped on one occasion. The classrooms were selected from US public, private, and parochial? schools in both urban and rural areas, and videotaping was conducted throughout the school year. The lesson content was not constrained, but most lessons drew from number theory, geometry, or algebra domains. Teaching styles similarly were not constrained and thus reflected a range of techniques and perspectives.

Lessons were analyzed using V-Prism, a computer software package designed to allow simultaneous viewing of a digitized video and its typed transcript on a computer screen. In the program, the video's transcript is time-linked so that the lines of text move temporally with the video.

The transcript may be marked with codes to designate when an episode begins and ends.

Six levels of analysis were developed such that every lesson was coded in six passes. Each pass was conducted by at least two coders and intercoder reliability was assessed. Codes and frequency data are presented below.

1. Identifying analogies

The definition of analogy used in this study was based on Gentner's structure mapping model (1983). Analogy was operationalized as a comparative structure between familiar objects, termed the *source* (or *base*) of the analogy, and relatively unfamiliar objects, termed the *target* of the analogy. *Objects* were defined as entities that function as wholes at a given level of analysis. The source and target objects are aligned according to their *predicate*, or relational, structure such that inferences are drawn from the source predicate structure to explain the target predicate structure. *Mapping* is defined as the process of aligning and drawing inferences between the source and target objects. Inferences are then drawn from the source structure and used to derive novel knowledge about the parallel target structure. Several constraints govern which mappings are made in each analogy, since all possible mappings are not typically completed (see Holyoak & Thagard, 1989).

Coders used a conservative measure of analogy, marking only units where a source, a target, and clear structural mapping between the source and target could be identified in discourse or explicit gestures. Two examples illustrate typical analogies. Reliability between the two coders was calculated on 105 protocols in 4 lessons (18% percent of the total sample). Agreement was 86%. Differences were resolved in discussion and consensus between the coders.

Figure 1 provides an illustration of the type of analogy typically identified in the data. This is a transcript of an analogy presented by a teacher to a whole class. The teacher is standing at the board, where the formula for circumference and a drawing of a circle is projected behind him.

Construction of Source object: *Highlighting the predicate structure:* *External layer of (peel, orange)*	Teacher: Now here's how I alway's looked at it. We're gonna say this- this circle right here is an orange Its an orange Alright? Its an orange Now lets say we're gonna take - stick a needle in the orange n' suck out everything inside except for the peeling of the orange. ((demonstrates gesturally)).
Mapping: orange peel to circumference	Okay we're gos- we're gonna pretend like that's our circumference right there.
Adapting the inference (this object will be an external layer) to the context of a geometric circle	((teacher uses a pointer to run along the outside edge of the circle on the overhead projector))

Figure 1: Example Analogy

A total of 103 analogies were identified as verbally produced by classroom participants. Lesson included a mean of 4.1 analogies (*SD* = 2.6). The range for individual lessons was between 1 and 11 analogies. Thus in every lesson examined, at least one explicit analogy unit was coded.

2. Participant Organization.

Each analogy unit was examined to record the roles played by teachers and students. Based on empirical and computational models of analogy, coders specifically determined whether a teacher, a student, a group of students, or no one generated the four steps of analogical reasoning demonstrated above in figure 1 (source, target, mapping, inference and adaptation, and problem solving).

Findings revealed that most analogies were initiated by teachers, and those produced by students tended to be highly superficial and only occurred in two classrooms (3% of the total number of analogies coded). Teachers produced 84% of sources, 77% of targets, and 89% of mappings. Students had a comparatively more active role in developing inferences and adapting them to target contexts, produced 27% alone or collaboratively with the teacher. Finally, students were most active in completing computations to provide final answers to target problems. Although only 42% of analogies had an explicit answer, students supplied 38% of those answers.

3. Source Structure

In order to examine the types of sources and targets produced typically in these mathematics analogies, the structure of the source and the target were independently coded for each analogy unit. Five classification categories were used. There were: *1)* decontextualized math problem, *2)* contextualized math problem (i.e., a word problem set in a non-math context), *3)* schema (general rule, no surface features), and *4)* outside math phenomena. If more than one source was used to explain the same target, or the target was stated in two ways, the source was coded " multiple." See Figure 2 for sample analogies of each structure.

Intercoder reliability was calculated separately for coding source and target. 22 protocols were used from 3 lessons (approximately 15% of the analogies coded). Reliability was calculated to control for chance using Cohen's Kappa, yielding k=72 and k=81 for source and target reliability (acceptable to good levels).

Both the sources and targets displayed significant differences between the frequency with which each structure

category was used, χ^2 (4) = 26.5; 68.4; p <.001 in both cases. While both sources and targets were most likely to be decontextualized math problems (40% and 44%), all four categories were represented. The next largest proportion of targets were math schemas (33%), suggesting that teachers were using analogies not only to prompt solutions to single problems, but also to aid in developing more general schemas. The most substantial distinction occurred between objects outside the domain of mathematics used as sources and targets. 15% of sources were outside the domain of mathematics, while only 1 out of 103 targets was a non-math phenomenon. This is not entirely unexpected since the math classroom is oriented towards mathematics learning, but it is conceivable that mathematics classrooms would discuss how to apply math structures to understand real-world problems. Finally, there were more multiple sources used (16%) than multiple versions of the target (5%).

Table 1: Structural composition of analogy sources and targets

STRUCTURE	FREQUENCY	
	Source	**Target**
Not math	15	1
Contextualized	19	18
Decontextualized	41	45
Schema	12	34
Multiple	16	5
Total	103	103

The overall level of similarity between the surface features of the source and the target was coded on a four point scale, as suggested by research indicating that this relationship governs the reasoning used to solve an analogy (e.g. Ross, 1987; Gick & Holyoak, 1983).

	Source	**Target**	**Surface Similarity**
1	Outside-math phenomena: *It s like balancing a scale, matter doesn t disappear, so to keep it balanced, whatever we do to undo one side we have to do to the other.*	Decontextualized math problem: *It is divided by negative sixty, so we multiply by negative sixty on both sides.*	Far distance
2	Contextualized math problem: *Lets say you ve got money. If you lost 88 cents and then you lost 5 cents more, would you add or subtract to find out the total amount you lost?*	Math schema *When you have a negative number minus another number, do you add or subtract?*	Schema involved
3	Decontextualized math problem: *Ok, don t put all that other stuff. What if it was just 16/20. How would you reduce it?*	Decontextualized math problem: *Now lets change the integers to monomials with variables. So then how would we reduce* $15xy^2z^4/25x^3y$*?*	Low similarity
4	Decontextualized math problem: *How would you multiply these?* (x + 2y)(x + y)	Decontextualized math problem: *In that case, how would you multiply* (5x+ y)(x + 3y)	High similarity

Figure 2: Analogy structure and surface similarity

Table 2 shows the frequencies of the surface similarity between analogy sources and targets. There were no significant frequency difference among the four categories, χ^2 (3) = 3.058 p=.3, indicating that all four types of analogy constructions were regularly used.

Table 2: Surface similarity between source and target

SURFACE SIMILARITY	FREQUENCY
Schema involved	28
Far distance// outside-math source/target involved	19
Low surface similarity	31
High surface similarity	25
Total	103

4. Function

The function of each analogy unit was coded to examine the context and purpose for analogy use. Primarily, this code examined whether the analogy was directed toward explaining a concept, a procedure, or a combination of the two. Additionally, a separate code marked whether the analogy was implemented following evidence of a student having trouble with a problem or concept.

	Mathematical Function	Paraphrased Examples
Code 1	Being a math student.	*Remove the parentheses, very carefully. Kind of like if you were a bomb squad called in to diffuse a bomb. If you mess up the first step, the whole problem will blow up in your face at the end!*
Code 2	Teach concepts only.	*When you're adding fractions, think about your denominators as units, like centimeters or feet. When you add length, the units must be the same in order to add them.*
Code 3	Teach concepts and procedures	*Ok do you remember how we found the perimeter and area of polygons last week? This time it is the same concept but we are going to use the formulas to solve for circumference and area of a circle.*
Code 4	Teach procedures only.	*What were the steps we used to solve the last example? OK lets do the same thing in this problem. First lets factor the numerator and denominator and then we'll see what we can cancel..*

Figure 3: Analogy Function

The mathematical function of each identified analogy revealed that some functions were more frequent than others χ^2 (3) = 20.1, p<.001. The raw frequencies are shown in Table 3. Teachers were most likely to use analogies to teach procedures only. The frequency distribution may be related to the tendency of US teachers to teach more procedures than concepts during a single lesson (Stigler et. al, 1998).

Table 3: Frequency of mathematical analogy functions.

FUNCTION	FREQUENCY
Being a math student	12
Concepts only	19
Concepts and Procedures	30
Procedures only	42
Total	103

We were interested in whether the function or purpose of the analogy was related to the type of analogy produced. Two steps were taken to investigate this question. First, a Pearson chi square was performed to compare analogy function with the structure of source generated. This test revealed that there is a significant relationship between the source generated and the structure of each analogy, χ^2 (12) = 35.8, p<.001.

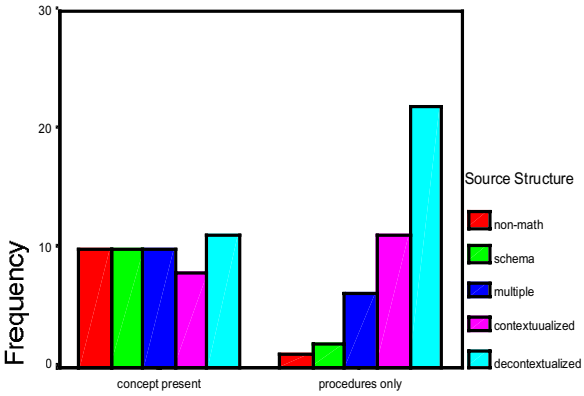


Figure 4: Source structure according to math teaching function.

Second, in order to determine whether the structure of the whole analogy is related to the function, a Pearson chi square was performed to compare the surface similarity between the source and target of each analogy with its function. This test revealed that there was a significant relationship between these variables, χ^2 (3) = 37.5 p<.0001. Far distance analogies were almost exclusively used to explain concepts or concepts paired with procedures. In contrast, relational mappings with high surface similarity between sources and targets were almost exclusively used to teach procedures only. Schemas were more likely to be used in analogies demonstrating concepts, but were also regularly used to teach procedures. Low surface similarity analogies, and mappings based on a decontextualized mathematics object, were used more frequently to teach procedures than concepts.

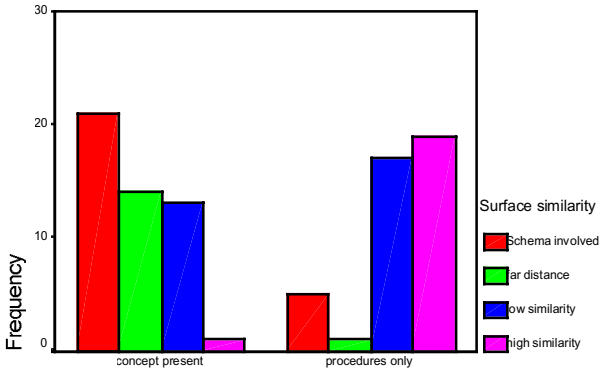


Figure 5: Surface similarity according to math teaching function.

Discussion

The present findings suggest that analogies are regularly produced in US mathematics classrooms, and are generated in reliable structural patterns. These patterns vary as a function of the immediate context and purpose of the analogy. The specific nature of these relationships suggest that analogy generation is highly constrained and organized by the goal of the analogy. When the goal of the analogy was to teach conceptual information, teachers were more likely to use distance analogies, schemas, and lower surface similarity than when they were teaching math procedures alone.

Patterns of analogy usage also revealed interesting associations between teachers' analogy practices and previous research. Teachers frequently presented multiple sources, techniques that have been demonstrated to facilitate schema acquisition and transfer (e.g., Gick & Holyoak, 1983). Thus it is possible that teachers may be using intuitive theories of analogy to guide their analogy production towards an effective teaching tool.

In addition, teachers provided substantial structure for each analogy, perhaps because they are aware that noticing analogies is a difficult task, as has been frequently demonstrated experimentally. While the teachers assistance makes the analogy more likely to be completed successfully than if students were responsible for generating more components, however, this design provides little information to the teacher about whether students are actually performing analogical reasoning when analogies are produced in this way. Although teachers frequently asked students to participate in producing analogies, they may not be aware that the way most students participate does not require that they perform higher order comparative reasoning.

Further, educational research suggests that enabling students to generate predictions and inferences about unknown problem types can be highly beneficial for their transfer performance (e.g. Carpenter, Fennema, Fuson & Hiebert, Human, Murray, Olivier & Werne, 1999). These findings therefore suggest that teachers might be reducing the learning benefits of analogy by highly structuring the comparisons for students. This is an empirical question that is currently under investigation.

Acknowledgments

Development of the TIMSS database was funded by the National Center for Education Statistics (NCES), the US Department of Education, and the National Science Foundation (NSF) grants to James Stigler. This research was additionally supported by an National Institute of Mental Health (NIMH) Developmental Cognitive Science Training Grant to Lindsey Engle.

References

Bassok, M. (2001). Semantic alignments in mathematical word problems. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.) *The analogical mind: Perspectives from cognitive science* (pp. 401-433). Cambridge, MA: MIT Press.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20,* 493-523.

Bransford, J. D., Brown, A. L., & Cocking, R.R. (1999). *How people learn: Brain, mind, experience, and school*. National Research Council, Commission on Behavioral & Social Sciences & Education. Committee on Developments in the Science of Learning. Washington, DC: National Academy Press.

Carpenter, T., Fennema, E.,; Fuson, K, Hiebert, J. Human, P., Murray, H., Olivier, A., & Wearne, D. (1999). Learning basic number concepts and skills as problem solving. In E. Fennema & T. Romberg (Eds.) ,*Mathematics classrooms that promote understanding* (pp. 45-61). Mahwah, NJ: Erlbaum.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg &J. E. Davidson (Eds),. *The nature of insight* (pp. 365-395) Cambridge, MA: MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461-493). Washington, D.C.: American Psychological Association.

Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 313-334). Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.

Gentner, D., Holyoak,, K. J.& Kokinov, B. (2001). *The analogical mind: Perspectives from cognitive science.* Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist, 52,* 32-24.

Gick, M.L.,& Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 15,* 306-355.

Gick, M.L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Holyoak & Thagard (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13,* 332-340.

Holyoak, K., Novick, L., Melz, E.R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In *2*. K. J. Holyoak & J. A.

Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol 2: Analogical* connections (pp. 113-180). Norwood, NJ: Ablex.

Kazemi, E., & Stipek, D. (2001). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *Elementary School Journal.* 102, 59-80.

Novick, L. (1988) Analogical transfer, problem similarity, and expertise. *Journal of Experimental psychology: Learning, Memory, and Cognition, 14*, 510-520.

Novick, L., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 398-415.

Reed, S. K., & Demptster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of experimental psychology: Learning, Memory and Cognition, 11,* 106-125.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 124-139.

Ross, B. (1987). This is like that: the use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629-639.

California Standards Testing and Reporting Program (STAR) (2001). Stanford 9 and California Achievement Test results. http://star.cde.ca.gov/star2001/

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1998). *Methods and findings of the TIMSS videotape classroom study*. Washington, DC: US Government Printing Office.

Category Size and Category-Based Induction

Aidan Feeney & David R. Gardiner
Department of Psychology
University of Durham, Stockton Campus
University Boulevard
Stockton-on-Tees, TS17 6BH
United Kingdom
{aidan.feeney, d.r.gardiner}@durham.ac.uk

Abstract

In this paper we investigate the role of category size in category-based induction. In a series of three experiments we asked participants about the strength of inductive inferences from arbitrary subordinate categories to their superordinates. We show that people use both subordinate and superordinate category size as a cue in category-based induction (Experiments 1 & 2). However, the results of Experiment 3 show that the effect of subordinate category size is smaller when the categories are said to be similar than when said to be dissimilar. On the basis of this result we suggest that people use category size as an indication of how much uncertainty remains concerning the superordinate rather than as a means of assessing how representative the category is as a sample of the superordinate. We conclude with a discussion of possible strategies for inductive reasoning.

One of the functions of categories is to promote inductive inference. Knowing that one set of instances possesses a certain feature allows us to consider whether other sets are also likely to possess the same feature. For example, knowledge that all of the chairs in the lecture room we are currently in are made of plastic will assist us in making a prediction about the chairs in the lecture theatre next door, the cafeteria at the end of the corridor and the provost's office. The experiments to be reported in this paper were all concerned with the role played by information about category-size in such inductive inferences. They ask whether participants are more likely to project a property if it is possessed by instances of a larger category than of a smaller category and whether people are more confident about conclusions concerning large or small groups. Furthermore, if people do turn out to be sensitive to size cues in this manner, what kind of reasoning underlies their use of such cues? As we will see below, there are a variety of ways in which category size might influence people's judgements of inductive strength.

Other researchers have been interested in induction based on categories and there are several models of categorical induction in the literature all designed to capture between 12 and 15 phenomena (for an excellent review, see Heit, 2000). One factor that is common to all of these models is inter-category similarity (Osherson et al, 1990) or featural overlap between categories (Sloman, 1993). To illustrate how similarity might affect the strength of an inductive inference consider the arguments below where the statement above the line is a premise and the statement below the line is a conclusion. With arguments of this type participants are asked to assume the premise to be true and to evaluate the degree to which it supports the conclusion.

Robins have an ulnar artery
Thrushes have an ulnar artery — Argument 1

Robins have an ulnar artery
Flamingos have an ulnar artery — Argument 2

As the categories in Argument 1 are more similar than those in Argument 2, people will judge the former to be stronger than the latter. Where the conclusion category is superordinate to the premise category, as in Argument 3 below, the degree to which the premise category is typical of the superordinate category informs people's judgements of inductive strength (see Rips, 1975).

Robins have an ulnar artery
Birds have an ulnar artery — Argument 3

A second factor which, in at least one model of category-based induction, impacts upon judgements of argument strength is 'coverage' (Osherson et al, 1990). Coverage is the degree to which the premise categories are similar to instances of the conclusion category (or, in cases where the conclusion category is not superordinate to the premise categories, to instances of the nearest superordinate category containing both premise and conclusion categories). So, for example, Argument 4 below would normally result in greater ratings of inductive strength than would Argument 5.

German Shepherds produce phagocytes
Poodles produce phagocytes
All dogs produce phagocytes — Argument 4

German shepherds produce phagocytes
Dobermans produce phagocytes
All dogs produce phagocytes — Argument 5

As the premise categories in Argument 4 are similar to a greater range of instances of the conclusion category than are the premise categories in Argument 5, the former is judged to be stronger than the latter. In general, the more diverse are the premise categories, the stronger is the argument (although for exceptions see Sloman, 1993).

There are several things to note about much of the existing work on category-based induction. First, although rarely formally contrasted with normative models of induction (for an exception see Heit, 1998), many of the effects in the literature have an intuitively strong normative basis. For example, both effects of similarity and coverage might be expected under the assumption that participants are sensitive to the representativeness of the samples about which they have some information. Samples that are either similar to, or typical of, the population to which the property will be projected, are, intuitively at least, more representative of that population. Similarly, diverse sets of premises intuitively seem to be more representative of the premises than are non-diverse premises.

A second characteristic of previous work on category-based induction is that researchers have been interested in investigating the effects of category knowledge on inductive judgements concerning natural kinds. As it is not normally possible to know the size of many naturally occurring categories (for example, how many members are there of the category 'bird'?) research has tended to concentrate on the role played by inter-category relationships. This may be contrasted with work on, for example, statistical judgement where both the a priori probabilities of the hypotheses as well as the probability of the evidence given each hypothesis has been manipulated. By presenting participants with problems concerning arbitrary categories in which category size was manipulated, the work to be described here attempted to address the role that category size plays in category-based induction.

Category Size and Category-Based Induction

Consider the following scenario:

672 people work in a 10 story office block. Of these, 313 work on floor 2 and 35 work on floor 7.

Given this scenario, which of these arguments is the strongest?

All 313 people who work on floor 2 have an identity number beginning with the letter Z

All 672 people who work in the office block have an identity number beginning with the letter Z — Argument 6

All 35 people who work on floor 7 have an identity number beginning with the letter Z

All 672 people who work in the office block have an identity number beginning with the letter Z — Argument 7

There are at least two reasons for preferring Argument 6 to Argument 7. The first line of reasoning is that the sample size in Argument 6 is larger than that in Argument 7. As larger samples are held to be more representative of the populations from which they are drawn than are small samples, Argument 6 is stronger than Argument 7 (for a recent discussion of the psychological literatures on sensitivity to sample size see Sedlmeier & Gigerenzer, 1997, and Keren & Lewis, 2000). However, since Nisbett et al's (1983) work on statistical heuristics in induction, it has been known that the variability of the feature being projected interacts with sample size to determine inductive strength. For example, Nisbett et al found that only a very small sample was required for participants to project features for which there is little within category variability (e.g. colour in a specific species of bird) whereas a much larger sample was required for the projection of more variable features. In the scenario above, the information that people work on different floors may suggest variability in staff identity numbers. That is, if category structure is made salient by a scenario, sample size may not be considered relevant in determining the strength of the inference.

The type of reasoning described above relies on indirect inference. That is, an inference about characteristic of a population is made on the basis of evidence about the prevalence of that characteristic amongst members of a sample. A less sophisticated, but more direct, way of making the inference is to think about the sample as a proportion of the population. Thus, if a large proportion of the population is known to possess the characteristic, then there is less uncertainty about the remaining members of the population and hence, a greater probability that the characteristic is universally possessed.

If we find that participants are sensitive to category size when asked to evaluate category-based inductive inferences, then the question arises as to what form of statistical reasoning underlies that sensitivity. The first two experiments to be reported here were designed to investigate premise and conclusion categories as cues to inductive reasoning whilst the final experiment was designed to compare contrasting accounts of any category size effect.

Experiment 1

Method

A total of 40 participants from the undergraduate population of the University of Durham (Stockton campus) took part in this experiment. Of these, 11 were male and 29 were female. The average age of participants was 22 years.

Experiment 1 had an entirely within participants design. The dependent variable was the number of problems for which participants chose as strongest the argument

concerning a large premise category. Each participant received a set of instructions and eight reasoning problems. Each problem described a superordinate category and two subordinate categories. The absolute size of each category was described such that the *Large* subordinate category was 25-40%, and the *Small* subordinate category 5-8%, of the size of the superordinate category. Participants received problems such as the following:

Extensive research has shown that there are several strains of the dreaded, and always fatal, Xanthrax virus. 1,000 people are known to have died from the virus. One form of the virus is Strain 6 from which 300 people have died. Another form is Strain 3 from which 60 people have died.

and were then asked to indicate which of two arguments was the stronger. These arguments consisted of a premise, concerning one or other of the subordinate categories, and a conclusion concerning the superordinate category:

Xanthrax Strain 6 produces a blotchy rash in sufferers
All 1,000 Xanthrax fatalities displayed a blotchy rash

Xanthrax Strain 3 produces a blotchy rash in sufferers
All 1,000 Xanthrax fatalities displayed a blotchy rash

The order in which the arguments appeared was controlled whilst the eight problems appeared in one of eight randomly determined orders. The other seven problems concerned books in a library, articles from several issues of a journal, the age of trees in a forest, houses sold by an estate agent, workers in an office block, characteristics of historical artefacts and works of art.

Results and Discussion

As expected, participants displayed a marked preference for arguments involving the large subordinate category. Out of a maximum of eight, the mean number of such arguments selected as being stronger was 6.13 (S.D. = 1.99). The difference between the number of large subordinate category arguments that were selected as stronger and the number that would be predicted by chance was statistically significant across all problem contents ($t(40) = 6.76$, $p < .001$). This preference for large premise categories was also statistically significant in all eight problem contents ($\chi 2(1) > 8$ in 7 out of the eight cases). Response frequencies, broken down by content, are displayed in Table 1.

The results of Experiment 1 confirm our intuition that participants are more likely to project a property to a superordinate category from a large rather than a small subordinate. In Experiment 2 we kept subordinate category size constant and, instead, manipulated the size of the superordinate category. Our strong intuition was that participants would be happiest projecting a feature to a small, rather than a large, category. An analogous effect exists in the literature (Osherson et al, 1990) where participants have been demonstrated to prefer projections to lower level, and hence smaller, categories.

Table 1: Large and Small argument selection from Experiment 1.

Problem Content	Subordinate Category	
	Small	Large
Disease	13	27
Library	8	32
Housing	11	29
Forest	10	30
Journal	10	30
Office Block	10	30
Artefacts	7	33
Gallery	6	34

A second aim of this experiment was to demonstrate category size effects in a between participants design. The literature on base rate neglect (see Koehler, 1996) contains demonstrations that participants are more likely to take the base rate into account when base rate is manipulated within participants. Hence, in Experiment 2 we wished to investigate whether participants would take category size into account in a between participant design.

Experiment 2

Method

The experiment had a 2 x 3 mixed design. Population size was manipulated between participants whilst each participant received three different problems asking them to rate the strength of an inductive argument.

A total of 116 participants from the undergraduate population at the University of Durham (Stockton Campus) took part in this experiment. Of these, 58 were male and 58 were female. The average age of participants was 21 years.

Participants received a booklet containing a set of instructions followed by three reasoning tasks. These tasks asked participants to evaluate the strength of arguments projecting a feature possessed by a subordinate category to all members of its superordinate. The problems concerned sub-types of a disease, individual production lines in a factory, and different variants of a plastic. Participants were

requested to rate the strength of the arguments on a 1-10 scale (very weak – very strong).

In cases where the superordinate category was small, the subordinate category accounted for between 45 and 55% of the superordinate. When the superordinate category was large, subordinate categories accounted for between 5 and 8% of the larger category. Importantly, only the size of the superordinate category was altered in this experiment. Approximately equal numbers of participants attempted the problems in each of the six possible orders.

Results and Discussion

The means and standard deviations from this experiment are presented in Table 2. A 2x3 Anova analysis revealed a significant effect of population size on the strength ratings assigned to arguments ($F(1, 114) = 5.32$, $MSE = 9.98$, $p<.03$). As expected, the mean ratings for large population arguments (mean = 2.81, S.D. = 1.68) were significantly smaller than for the small population arguments (mean = 3.59, S.D. = 1.96). Neither of the other effects tested by the analysis were significant.

Table 2: Means and standard deviations from Experiment 2.

Condition	Content		
	Disease	Factory	Plastic
Large	3.84 (2.41)	3.14 (2.23)	3.79 (2.71)
Small	2.74 (2.07)	2.84 (2.07)	2.84 (1.96)
	3.29 (2.31)	2.99 (2.14)	3.32 (2.40)

As expected, participants rated arguments projecting features to small conclusion categories more highly than they did arguments concerning large conclusion categories. However, one striking aspect of these results is that mean ratings of argument strength were very low. It would appear that, at least in a between participants design, making subordinate categories salient in the scenario causes participants to doubt the conclusion regardless of conclusion category size. One possible reason for this is that participants may expect there to be differences between subordinate categories. This expectation may contribute to their unwillingness to project a feature to the superordinate on the basis of evidence concerning only one subordinate. In Experiment 3 we investigated the effects of explicit information about similarities and dissimilarities between subordinate categories on ratings of inductive strength.

Experiment 3

Experiments 1 and 2 have demonstrated that participants are sensitive to both subordinate and superordinate category when evaluating the strength of inductive arguments. In Experiment 3 we attempted to contrast two possible accounts of people's use of category size as a cue for induction.

The first account we considered was that people's judgements are affected by subordinate category size because they use a sample size heuristic (see Nisbett et al, 1983). That is, people realise that larger samples are more representative of the population from which they are drawn than are smaller samples. Alternatively, people may reason that a large subordinate leaves a smaller proportion of the superordinate unaccounted for than a small subordinate and hence, makes for a stronger argument.

To test these alternative accounts we gave one group of participants reason to believe that all subordinate categories in the domain were similar whilst telling another group that they were dissimilar. Our reasoning was that 'similar' problems should result in fewer attributions of variability to the superordinate category. If participants were cautious in using a sample size heuristic in Experiment 1 due to worries about the representativeness of the sample, then for 'similar' problems participants in this experiment should be less cautious and a greater effect of category size should be observed. On the other hand, 'dissimilar' problems should produce even more caution and a smaller effect of category size.

An alternative hypothesis is that the effect of category size in Experiment 1 may have been due to an assumption that the subordinate category was unrepresentative of the superordinate. That is, participants may have been unwilling to project a property to the superordinate because the subordinate category for which information was available may not have 'covered' the superordinate category. Accordingly, participants may have endorsed inferences from large premise categories more strongly because large premise categories leave fewer cases unaccounted for in the conclusion category. Thus, telling participants that the subordinate categories in the scenario are similar may cause them to rely less on category size as a cue and to assign higher ratings of inductive strength regardless of category size. Explicitly telling them that members of the superordinate are dissimilar may lead them to rely even more heavily on sample size.

Method

Eight male and 52 female undergraduate students (average age 27 years) from the University of Durham's Stockton Campus took part in this experiment which had a 2 x 2 entirely within participants design. The factors manipulated were the similarity said to hold between subordinate categories in each of the problems (similar vs. dissimilar) and the size of the subordinate category (large vs. small).

Each participant received a nine-page booklet comprising a set of instructions and eight inductive reasoning problems.

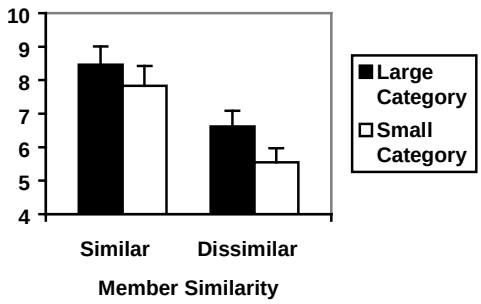
In each of these problems participants were told that instances of a subordinate category possessed a feature and were asked to evaluate the strength of an argument projecting this feature to all members of the superordinate category. The problems concerned workers in an office block, symptoms of a disease, features of handmade chairs, materials used by an engine manufacturer, properties of a type of plastic, the type of material used in indigenous art, production lines in a factory, and stock at a fish farm. Each participant received one version of each problem and two problems in each condition of the experiment. Each problem content appeared equally often in each condition of the experiment.

The problems were designed so that the large subordinate category was 45-55% of the size of the superordinate category, whilst the small subordinate category contained 5-8% of the superordinate population. To achieve the similarity manipulation participants were explicitly told that some similarity/dissimilarity existed between all members of the superordinate category. For example, in the office block problem participants were told that the workers either worked for the same, or different, divisions of a company.

Results and Discussion

As participants completed two problems per condition of the design, we calculated a mean score per condition for each participant. We carried out a 2 x 2 entirely within participants ANOVA on this data, the means from which may be seen in Figure 1. Our similarity manipulation had a highly significant effect upon the ratings of argument strength ($F(1, 59) = 24.70$, $MSE = 10.38$, $p<.001$). Mean ratings of argument strength were higher for categories containing similar members (8.15) than for categories containing dissimilar members (6.08). This finding replicates previous work (e.g. Nisbett et al, 1983) suggesting that within category variability significantly affects people's willingness to project a property from a sample to a population.

Figure 1: Interaction between Category Size and Similarity from Experiment 3



Our category size manipulation also had a significant effect on ratings of argument strength ($F(1,59) = 6.03$, $MSE = 7.19$, $p<.02$). Arguments involving the large subordinate category were rated as stronger (7.54) than those involving the small subordinate (6.69). Although the interaction between these factors did not approach significance ($F(1, 59) = .34$, $MSE = 8.21$, $p > .5$), planned comparisons revealed a significant effect of category size when category members were said to be dissimilar but not when they were said to be similar.

When participants are explicitly told that the members of the superordinate category are similar, premise category size ceases to have a significant effect on judgements of inductive strength. Instead, it would appear that category size is more important under conditions where 'indirect' inductive inference from a subordinate to a superordinate is likely to be unsafe. This suggests that category size functions as an indicator of the number of cases for which uncertainty remains.

General Discussion

Our finding, that category size acts as a cue in category-based inductive inference, is entirely novel (if not entirely unexpected). Likewise, the finding that the effect of category size decreases when participants are explicitly told that subordinate categories are similar is also novel We will discuss possible interpretations of these findings and their implications for the question of strategy use for induction.

The results of Experiment 3 might be regarded as being consistent with Nisbett et al's findings as in their study participants were found to be insensitive to sample size when the property that they were required to project was unlikely to vary. Similarly, our participants were also relatively insensitive to sample size when told that there were similarities between subordinate categories. It might even be argued that participants' greater sensitivity to sample size in the dissimilar condition is also consistent with Nisbett et al's results and is evidence that participants realised that in conditions of variability, a large sample is safer.

We are unsure about this reading of the results because we find it implausible that participants told that there were dissimilarities between subordinate categories should consider a bigger subordinate category more likely to be representative of the population than a smaller subordinate. We find it much more plausible that participants attributed greater strength to arguments that accounted for the most population members. It is plausible, however, that in the similar condition, participants projected features to the population on the basis of characteristics of the sample. Even in the small category size condition, premise categories always had more than 20 members. This figure

was chosen as 20 was the largest sample size given to participants in Nisbett et al's experiment. Given the existence of a subordinate category structure, participants may have considered the increase in inductive strength from small to larger premise categories to be insubstantial.

We contend that our results suggest that people are flexible in the strategies that they adopt for inductive inference. In conditions of low variability, people will project properties to a population on the basis of the characteristics of a sample. However, where high variability exists, people may be more likely to base their judgements of inductive strength on the number of cases outstanding.

The 'direct' strategy is a very interesting one, partly because it has been somewhat neglected in the literature (although see Evans & Dusoir, 1977). The conditions for its application are population size being both finite and approximately known. In addition, we suspect that a 'direct' strategy will be used in conditions where the population is small. This is because indirect induction only becomes necessary where the population is large or infinite and it is difficult, or impossible, to check all members. With small populations a direct strategy based on checking as many members as possible is more tractable.

Evidence that group size can affect how people perform induction comes from Wang (1996) who showed that the demonstration of classic framing effects depends on the size of the group being reasoned about. His explanation for this finding is that we possess social-group domain-specific reasoning abilities. As Wang only gave scenarios concerning social groups to his participants, we are unwilling to subscribe to the notion of social-group domain-specific abilities. However, we agree that the human species is likely to have made inferences about relatively small populations throughout most of its history. In addition, formal notions of induction evolved relatively recently (see Gigerenzer et al, 1989), most probably in response to a need for safe inferences about population whose size was unknowable or infinite. Although the 'direct' strategy is primitive when compared to more 'indirect' forms of induction, such a strategy may work very well with small populations.

References

Evans, J.St.B.T. & Dusoir, A.E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica*, 41, 129 - 137.

Gigerenzer et al (1989). *The empire of chance*. Cambridge: Cambridge University Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. & Chater, N. (Eds.) *Rational models of cognition*. Oxford, UK: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, 7, 569-592.

Keren, G. & Lewis, C. (2000). Even Bernoulli might have been wrong. *Journal of Behavioral Decision Making*, 13, 125-132.

Koehler, J.J. (1996). The base-rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences*, 19, 1-53.

Nisbett, R.E. et al (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90, 339-363.

Osherson, D.N., *et al* (1990). Category-based induction. *Psychological Review*, 97, 185-200.

Rips, L. (1975). Inductive judgements about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14, 665-681.

Sedlmeier, P. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, 10, 33-51.

Sloman, S.A. (1993). Feature based induction. *Cognitive Psychology*, 25, 231-280

Wang, X.T. (1996). Domain specific rationality in human choices: Violations of utility axioms and social contexts. Cognition, 60, 31 - 63.

Acknowledgements

This research was funded by grant R000222426 from the Economic and Social Research Council of the United Kingdom

Why Example Fading Works: A Qualitative Analysis Using Cascade

Eric S. Fleischman (esfleisc@colby.edu)
Randolph M. Jones[*] (rjones@colby.edu)
Department of Computer Science, Colby College, 5830 Mayflower Hill Drive
Waterville, ME 04901–8858 USA

Abstract

"Faded" examples are example problems that provide a solution, but first require students to generate a portion of the solution themselves. Empirical studies have shown that such examples can be more effective teaching aids than completely worked examples that require no work from the student. Cascade is a model of problem-solving skill acquisition that was originally developed to explain other empirical regularities associated with human problem solving and learning, most notably the self-explanation effect. Past research demonstrated that Cascade might also explain the mechanisms underlying the effectiveness of example fading. This paper analyzes new protocol data, and finds that it is consistent with predictions derived from Cascade.

Overview

Renkl, Atkinson, and Maier (2000) empirically demonstrated the qualitative result that, when learning problem-solving skills, students studying a series of "faded" examples show improved post-test performance over students studying only completely worked examples. Jones and Fleischman (2001) argue that this result can be explained by Cascade (VanLehn, Jones, & Chi, 1991), a computational model of problem-solving skill acquisition. Cascade was originally developed to understand the mechanisms of the self-explanation effect (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Pirolli & Anderson, 1985). Jones and Fleischman demonstrated that the mechanisms underlying self-explanation might also explain the effectiveness of studying faded examples. Although they showed that Cascade is consistent with the fading result, the explanation involved assumptions that had not yet been tested empirically. Therefore Jones and Fleischman (2001) finished with a small set of predictions and suggestions for new experiments to confirm or dispute Cascade's account. Since that time, Renkl, Atkinson, and their colleagues have run additional experiments, collecting detailed transcripts of subjects studying two types of faded sequences of problems. Although the experiments are not yet complete, we have been able to perform a qualitative analysis of the protocol data for eight of the subjects. Additionally, we have fine-tuned Cascade's knowledge base (but not its underlying mechanisms) to more faithfully model the current data. This paper reports the result of using Cascade to develop a qualitative analysis of the eight subjects. The primary result is that the findings remain consistent with Cascade's account of example fading, as well as the predictions made by Jones and Fleischman (2001).

Background

Years of research have demonstrated effective techniques for teaching students problem-solving skills in a variety of task domains. In particular, a number of studies show that students benefit from being given a series of completely worked example problems, followed by a series of unworked practice problems (e.g., Chi et al., 1989; Pirolli & Anderson, 1985; Renkl, 1997, VanLehn, 1996). Other studies show that the effectiveness of such a curriculum depends in part on the willingness of the students to explain the worked examples to themselves in detail, rather than simply giving the examples a superficial read (Chi et al., 1989; Fergusson-Hessler & de Jong, 1990; Pirolli & Bielaczyc, 1989). VanLehn and Jones (1993a, 1993b; VanLehn et al., 1991) developed Cascade in order to determine the cognitive mechanisms behind this *self-explanation effect*. In essence, Cascade suggests that thorough study of worked examples help students consciously expose and patch gaps in their task knowledge. In addition, self-explanation provides contextual memories that can guide future problem solving by analogy to familiar examples.

Subsequent experiments by Renkl et al. (2000) suggest that student learning can improve even further by *fading* a curriculum from fully worked examples to partially worked examples. The partially worked examples provide a complete solution to the problem (as with fully worked examples), but first require students to derive one or more steps on their own. This in turn requires the students to understand the rest of the example in at least enough detail to be able to attempt a solution.

Jones and Fleischman (2001) argue that the reason faded examples improve learning is that they retain much of the guidance provided by the context of a solved example, but they force the students to work on particular parts of the problem, in turn possibly forcing them to expose and patch knowledge gaps. This is in contrast to studying completely worked examples, where it is basically up to the students to decide whether they are going to put any effort into understanding the examples (because the students are not required to produce any answers in that case). This argument came directly from the assumption that Cascade is

[*] The second author is also affiliated with Soar Technology, Inc.

an accurate model of human problem solving and learning (at this level of abstraction).

Jones and Fleischman ran Cascade on a mock "faded" curriculum in order to demonstrate the plausibility of their hypothesis. This exercise confirmed that the proposed explanation is a sufficient account of the general fading results, but the explanation rests on a number of assumptions that had not yet been confirmed by empirical data. The first assumption was that the classical physics problem domain (implemented in Cascade and studied by Chi et al., 1989) is sufficiently similar to computing simple probabilities (studied by Renkl et al., 2000). The second assumption was that Cascade's underlying processes accurately match what subjects do when learning from faded examples. Data had simply not yet been collected to argue this point either way. Thus, Jones and Fleischman (2001) presented three specific predictions to be confirmed or denied by subsequent empirical research:

1. "Faded examples cause effective learning by forcing the student to encounter and overcome an impasse."
2. There is likely "…at least some benefit to example fading from the learning of search control knowledge."
3. "The primary benefit of a faded example is that it forces the student to process parts of the example that they might otherwise ignore."

They also suggested that these predictions be tested with new experiments that include the collection of protocol data.

The current work tackles both of these issues. To begin with, Renkl and Atkinson (and their colleagues) have initiated an additional study to collect more detailed subject data, including transcribed talk-aloud protocols generated by the subjects while studying and solving problems. Although their experiment and analysis is not yet complete, they provided us with eight initial transcripts, enabling us to generate a partial coding that tests the predictions listed above.

We have also generated a new task knowledge base for computing probabilities, so Cascade can solve precisely they same problems given to the subjects in the new experiments. This allows us to remove a model assumption, and verify the Cascade results with a more accurate match to the data. The next section describes the methods we used to generate the new knowledge base and encode the protocols. The following sections present the results of those activities in more detail.

Methods

Given the study material presented to the experimental subjects, we first performed a thorough task analysis. This involved identifying the probability equations required for solving the set of study problems. This set serves as the target knowledge base that we would expect a "perfect learner" to have acquired after complete study of the curriculum. The task analysis allowed us to replace Cascade's physics task knowledge with task knowledge about computing probabilities. It is important to note that we only changed Cascade's task knowledge. We did not change any of the underlying problem-solving or learning mechanisms built into Cascade. Once we defined the target knowledge base, we represented each problem as a set of given and sought quantities, using Cascade's representation language within Prolog.

After doing the task analysis, we ran Cascade on each problem in order to perform a content analysis. The content analysis records the required equations for each solution. This allows us to predict interactions between performance on separate problems by a single subject. For example, if a subject fails to use an identified equation in one problem (as suggested by an error combined with protocol evidence), but then correctly solves a subsequent problem that requires the same equation, we can safely hypothesize that some sort of learning took place even if there is no direct evidence of a learning episode in the protocol transcript. This helps us track learning across a series of problems. The content analysis also helps constrain the encoding of the subject protocols. In the face of ambiguous utterances that lead to a correct solution, we can generally infer which equations the subject must have used correctly.

Our final task was to encode the subject protocols for behavior episodes relevant to the predictions reported above. Future work (when all subject protocols are available) will contain thorough quantitative analyses of various protocol encodings. For the current effort, however, we performed a qualitative analysis, looking for general trends in the data. The goal was to investigate whether there were any relationships between example fading and learning, the use of analogies for search control, and the generation of self-explanations. We constrained the protocol encoding by performing a goal decomposition to match each subject protocol. Running Cascade on the same problems generates similar goal decompositions, which we can then use to inform the coding process.

Task and Content Analysis

As mentioned above, the task analysis determined all of the equations, or knowledge chunks, required to solve the set of probability problems from the empirical study. We did not include more basic arithmetic reasoning (such as addition, multiplication, and the ability to isolate variables) in the analysis. This is because, for the domains Cascade has been used to study so far, it simplifies the model to assume that subjects have well rehearsed knowledge of these tasks. For the problems in question, we identified twelve distinct, required equations. Some of the equations compute simple probabilities by dividing the cardinality of various sets of objects. The task knowledge includes cases for choosing objects at random with and without replacement. The target knowledge base also includes various equations for combining probabilities. These include the special multiplication rule

$$P(e1 \text{ and } e2) = P(e1) \times P(e2)$$

the addition rule

So I have a total of 12 bottles and there are 4 that are turned into vinegar. So a total of 4 vinegar and 8 drinkable. Probability of vinegar is 1/3 and drinkable is 2/3. Now if we take one then we're left with…There is a 1 in 3 chance that that will be vinegar…	S: value(p(event1a)) S: solve(p(event) = size(selectionpool) / size(totalpool)) S: value(size(selectionpool)) F: value(size(selectionpool)) = 4 S: value(size(totalpool)) F: value(size(totalpool)) = 12 F: solve(p(event) = size(selectionpool) / size(totalpool)): 1/3 F: value(p(event1a)) = 1/3

Table 1. A protocol excerpt and its corresponding analysis in the form of a Cascade trace.

$$P(e1 \text{ or } e2) = P(e1) + P(e2) - P(e1 \text{ and } e2)$$

and the subtraction rule

$$P(\text{not } e) = 1 - P(e).$$

It is worth comparing some features of this domain with classical mechanics, the domain used in previous studies with Cascade. The physics and probabilities domains share the feature of involving mostly symbolic problem-solving skills, which we feel is the defining characteristic that allows Cascade to model both well. However, there are also some potentially significant differences between the two domains.

To begin with, the target knowledge base for computing probabilities is much smaller than the physics knowledge base. In contrast to the current set of 12 equations, Cascade required explicit representation of 62 separate chunks of knowledge for classical physics. Another significant difference is that subjects often relied on common-sense reasoning to explain and learn physics skills. Thus, the Cascade model for physics included a number of general common-sense rules that could be used to guide the learning of correct (and sometimes incorrect) physics knowledge. In contrast, there seems to be much less opportunity to generate common-sense explanations for the rules of probabilities (although there are certainly some). In the protocols we have studied so far, subjects generally made little use of common-sense principles when they got stuck.

To perform the content analysis, we encoded the problems provided by Renkl and Atkinson into Cascade's representation. This essentially involved translating the problems' given and sought quantities into a Prolog-style predicate representation. Once complete, we ran Cascade to ensure it could solve each problem with the complete knowledge base. Each problem run generated an execution trace that provides explicit detail about which equations are necessary to solve each problem. With these tools in hand, we proceeded to analyze each subject protocol to track the usage of individual equations, self-explanation behavior, the use of analogy to guide problem solving, and learning.

Protocol Analysis

The basic approach to the protocol analysis was to assume that Cascade provides an accurate model of each subject's behavior and then to look for inconsistencies. We patterned this approach after Jones and VanLehn's (1992) evaluation of Cascade's ability to model the fine-grained behavior of individual subjects studying and solving physics problems. For each subject-problem pair, we generated the hypothetical solution trace that Cascade would have to generate in order to produce the utterances observed in the subject. We allowed ourselves to tune the trace only by assuming that Cascade has missing or incorrect knowledge about computing probabilities. Any other discrepancies between Cascade and the protocol data are marked against Cascade's ability to explain the subject's performance. Table 1 presents an example protocol excerpt and the corresponding Cascade trace that matches the internal behavior suggested by the subject's utterances.

We constructed Cascade-like traces for a number of problems, and used those results to guide the rest of our protocol analysis. Recall that the predictions presented by Jones and Fleischman (2001) focused on self-explanation, forced impasses, knowledge acquisition from impasses, and knowledge tuning. Thus, our protocol analysis focuses on these three issues.

Self-Explanation

For Cascade's account of fading to be correct, it must mean that subjects tend to generate more self-explanations (or problem-solving activity) for faded examples than they do for completely worked examples. This prediction is borne out in the protocol data we examined. Subjects rarely engaged in self-explaining on fully worked examples. Subjects were much more engaged in the faded examples, presumably because the faded examples demanded them to generate some kind of answer. The protocols show evidence that sometimes even this was not enough to ensure self explanation. Sometimes, subjects would simply "click through" the faded portions of the examples, and skip on to the solutions. For example, after revealing a solution step, one subject simply said "Boy…summmsumm…I don't know this right now," and proceeded to the next step. However, there were certainly many more instances of self-explanation for faded examples than there were for completely worked examples.

Impasses

Subjects encountered impasses during fully worked examples even more rarely than they bothered to self-explain the examples. This is because a subject cannot experience an impasse without first engaging in some self-explanation. However, subjects experienced many impasses when working on faded examples. This is because, if the

subject made a mistake, they received relatively immediate feedback by then being shown the correct solution step. If the subject bothered to read the revealed solution, they would have to acknowledge a discrepancy and go back to refigure things. The following excerpt shows an example of a subject first reading a completely worked example (Problem 5) with no impasse, and then working a faded example (Problem 6) that forces an impasse. Both problems require precisely the same set of equations to generate a correct solution.

> Problem 5:
> S: Okay, let's see here. Probability is 1/10 time 1/5, okay, I see how they did it, allright. Probability of stitching and/or color defects is 1/10 plus 1/5 minus the total probability that's 1/50, and that equals (reads aloud) okay, next.
>
> Problem 6:
> S: Okay, allright. Now. This is the difference, that's going to be 1 minus the 2/50 plus the 23/50, that's going to be, 1 minus okay, 2/50 or, 2/50 equals .04, and plus 23/50 equals .46, now, .46 and .04, give me .50, 1 minus .50 equals .50, so, I'm doing this right, it should be .50, no, okay, allright, let's see, okay, I guess I…, okay, so that's 1 minus that, okay, I see what I did.

In this excerpt, the subject essentially just reads Problem 5 and claims to understand it (which is, paradoxically, a hallmark of subjects that are *not* doing enough self explanation). The subject generates an answer to Problem 6 that they think is correct, but when they reveal the correct solution step they discover they are wrong. This forces an impasse. In this particular excerpt, there is no convincing evidence that the subject actually resolved the impasse and learned the correct solution sequence, but the impasse at least gave them the opportunity. The following sections discuss analysis of actual learning episodes in the protocols.

Knowledge Acquisition

We were surprised to find no obvious episodes of knowledge acquisition in the protocols. That is, we found no evidence that subject were missing entire chunks of knowledge that they were then able to discover in response to an impasse. This was particularly surprising because Jones and Fleischman (2001) assumed a key role for knowledge acquisition episodes during their initial Cascade study with physics problems. It appears that this is a place where differences in the task domains are significant. As mentioned previously, knowledge acquisition episodes were an extremely important part of Cascade's account of the self-explanation effect for the physics domain. However, in all of the protocols for subject computing probabilities, it appears that they already *know* all of the equations they need; they just have not yet learned the right times to *use* them. This is admittedly a subtle distinction that cannot always be verified in the protocol data, so we plan to give it a much closer look in future studies. However, since our original proposal gave such a large role to knowledge acquisition, we feel we are being conservative by suggesting that there are no knowledge acquisition episodes at all in the current protocol data. There are clearly other types of learning episodes in the data, which we describe below, and we feel that those remain consistent with Cascade's predictions about fading.

Knowledge Tuning

One of the predictions about Cascade's account of fading was that fading enables students to tune knowledge they have already acquired, by allowing them to use it in a useful problem-solving context. In Cascade, all knowledge tuning occurs via a process of analogical search control. Thus, we expect to see subjects learn after they have successfully drawn an analogy between two problems. We observed many such episodes in the current protocol data. The following excerpt provides one of the clearest examples:

> Problem 2:
> S: …The chance of it being drinkable is 8 to 11 so the probability of her drinking, probability that the first bottle is vinegar but the second is drinkable, 2 red balls and 2 white balls is 4, probability is 1/2 so if we multiply 1/3 times 8/11 that will be 8/33…

Problem 2 involves a collection of bottles containing wine and vinegar. However, in the middle of the excerpt, the subject makes an explicit analogical reference to Problem 1, which deals with selecting a particular configuration from a collection of red and white balls. In a subsequent problem that uses precisely the same solution technique, the subject easily solves the problem correctly, without any evidence of an impasse or overt analogical reference.

It seems clear that this particular episode involves knowledge tuning via analogy. There are other episodes of knowledge tuning that are not overtly analogical. The current Cascade model dictates that all knowledge tuning occurs by analogy, but that happens at a low enough cognitive level that it is difficult to prove or disprove. Certainly there are many cognitive theories that posit some sort of similarity-based memory for skills and facts.

There is one aspect of knowledge tuning that Cascade does not model well. Some subject protocols show basically the same pattern as the excerpt above, but the tuning occurs more gradually across 3 or 4 problems. Cascade's knowledge tuning mechanism is more of an all-or-nothing proposition. As soon as Cascade solves one problem by analogy, it can immediately retrieve the same knowledge in similarly structured future problems. This seems to be a clear weakness in the Cascade model. However, it does not invalidate Cascade basic account of fading. The subject protocols show that the use of analogy occurs more frequently during fading-driven impasses than during the study of completely worked examples.

In prior Cascade studies in the physics domain, there were strong interactions between the knowledge tuning and knowledge acquisition mechanisms (VanLehn & Jones, 1993b). We expect that similar interactions could help

explain the effectiveness of fading examples. However, since we have so far seen no evidence of knowledge acquisition in the current study, it has not been important to analyze potential interactions.

Conclusions

We conclude by reiterating the predictions that Jones and Fleischman (2001) proposed to gather evidence for Cascade's account of the benefit of faded examples:

1. "Faded examples cause effective learning by forcing the student to encounter and overcome an impasse."
2. There is likely "…at least some benefit to example fading from the learning of search control knowledge."
3. "The primary benefit of a faded example is that it forces the student to process parts of the example that they might otherwise ignore."

We feel that our initial analysis of protocol data from Renkl, Atkinson, and colleagues confirms each of these predictions to some extent. The basic effect is strong: students often do not expend much effort on understanding completely worked examples, but fading the examples gives the students a strong impetus to do so. This encouragement to work out portions of the examples leads to more opportunities to identify incorrect (or incorrectly applied) knowledge, which in turn provides opportunities to correct or tune that knowledge. We were surprised to find that knowledge acquisition did not appear to play a significant role in the probabilities task domain. However, future analysis will more closely search for such episodes. In addition, although it makes the model less interesting in some ways, the preponderance of analogical knowledge tuning is entirely consistent with the Cascade model. Since Jones and Fleischman's (2001) original study of fading with Cascade did not focus strongly on knowledge tuning (because it played less of a role in the physics domain), an important future task is to run a thorough set of experiments with Cascade to confirm that knowledge tuning can account for all of the observed improvements in problem-solving skill.

It is also certainly possible that future analysis will uncover data that is inconsistent with Cascade's predictions. With this possibility in mind, our next course of action is to gather even more data and perform a more thorough quantitative analysis. We also expect that we will find some ways in which Cascade should be improved. For example, we already know that the knowledge tuning mechanism should be adjusted to account for more gradual forms of knowledge tuning observed in the subjects. Any further mismatches of the model to the data should also serve to improve our understanding of how humans learn problem-solving skills and, as a consequence, inform how we ought to teach them.

Acknowledgments

This project could not have proceeded without the gracious collaboration of Alexander Renkl and Bob Atkinson, together with their research groups. This work was also made possible by the support of Colby College, particularly the Colby College Senior Scholars program, and Soar Technology, Inc.

References

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Ferguson-Hessler, M. G. M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor solvers. *Cognition and Instruction, 7*, 41–54.

Jones, R. M., & Fleischman, E. S. (2001). Cascade explains and informs the utility of fading examples to problems. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp.459-464). Mahwah, NJ: Lawrence Erlbaum.

Jones, R. M., & VanLehn, K. (1992). A fine-grained model of skill acquisition: Fitting Cascade to individual subjects. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 873–878). Hillsdale, NJ: Lawrence Erlbaum.

Pirolli, P., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39*, 240–272.

Pirolli, P., & Bielaczyc, K. (1989). Empirical analyses of self-explanation and transfer in learning to program. In G. M. Olson & E. E. Smith (Eds.), *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 450–457). Hillsdale, NJ: Lawrence Erlbaum.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1-29.

Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 393–398). Mahwah, NJ: Lawrence Erlbaum.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513–539.

VanLehn, K., & Jones, R. M. (1993a). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning*. Boston: Kluwer Academic.

VanLehn, K., & Jones, R. M. (1993b). Integration of explanation-based learning of correctness and analogical search control. In S. Minton (Ed.), *Machine learning methods for planning*. Los Altos, CA: Morgan Kaufmann.

VanLehn, K., & Jones, R. M. (1993c). Better learners use analogical problem solving sparingly. *Machine Learning:*

Proceedings of the Tenth International Conference (pp. 338–345). San Mateo, CA: Morgan Kaufmann.

VanLehn, K., & Jones, R. M. (1993d). What mediates the self-explanation effect? Knowledge gaps, schemas or analogies? *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 1034-1039). Hillsdale, NJ: Lawrence Erlbaum.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1991). A model of the self-explanation effect. *Journal of the Learning Sciences*, *2*, 1–59.

Evolution of Gender in Indo-European Languages

Harry E. Foundalis (hfoundal@cs.indiana.edu)
Computer Science and Cognitive Science
Indiana University, Center for Research on Concepts and Cognition
Bloomington, IN 47405 USA

Abstract

In a recent paper, Lera Boroditsky and Lauren A. Schmidt (2000) examined the degree to which the linguistic category of grammatical gender of nouns influences people's perception of the cognitive category of biological gender, or sex. Their conclusion was that English speakers' intuitions about the gender of certain nouns (animals) correlate with the gender assigned to those nouns in languages such as German and Spanish. More important, they found that people's ideas about the putative biological gender (sex) of objects are strongly influenced by the grammatical gender of those objects in their native language. In this study I sought to reproduce Boroditsky and Schmidt's results in order to show that the interpretation they supplied is unwarranted, and that the authors conflate the concepts of biological gender (sex) and "formal gender", which is employed by most Indo-European languages (as opposed to "natural gender", in English). I compare the intuitions of 20 American monolinguals with the statistics of formal gender as it appears in 14 Indo-European languages. Moreover, I discuss the possible origin and evolution of gender in such languages, and suggest an explanation for the relation between grammatical and biological gender.

Introduction

The idea that our native language may shape our thought, in part or in whole, is usually associated with the work of Whorf and Sapir, in what is known as "the Sapir-Whorf hypothesis" (Whorf, 1956). This is an intriguing hypothesis because it implies that different cultures — speaking different languages — may perceive the world in different ways. For example, whereas one culture may differentiate objects on the basis of shape, another culture may differentiate them on the basis of material (Imai and Gentner, 1997), and this may be reflected in the corresponding languages. To what extent, then, does language (and culture) force a person's cognition to perceive the world in one way rather than another?

A possible manifestation of this idea was examined by Boroditsky and Schmidt (henceforth B&S), in studying the way grammatical and biological gender interfere with each other in the minds of native speakers of languages such as Spanish and German. B&S support the idea that a speaker whose language assigns the genders masculine and feminine to nouns — whether they refer to people, animals, things, or ideas — is bound to subliminally think of an object as having a corresponding biological gender, male or female. (To avoid circumlocutions, I use the word "sex" to refer to biological gender, reserving "gender" for the grammatical category.)

B&S's proposal rests on the assumption that there is an inherent equating of the concepts of gender and sex in such a speaker's mind. So, for example, a young learner of an Indo-European language employing "formal gender" could associate a specific category of nouns discernible only through the behavior of neighboring words (say, the feminine nouns) with a perceptual property of entities of the world (say, the femaleness of individuals), even before encountering the words for "feminine" and "masculine". Although the latter point to a certain relation between gender and sex (which undoubtedly exists), we will see that such an assumption is untenable. First, however, we should briefly review the category of gender as it appears in various languages, in order to understand what it is, and what relation we may expect between the concepts of gender and sex.

Although many people are familiar with gender as it appears in Indo-European languages, the notion of gender as understood by linguists is much more general. As a "definition", I will follow Charles F. Hockett's description: "Genders are classes of nouns reflected in the behavior of associated words" (Hockett, 1958:231). A characterization like this is general enough to encompass all noun categories that linguists call "genders", whether they are labeled "masculine", "feminine", "neuter", "common", or even "class IV".

A language may have two or more classes of nouns that qualify as genders, or it may have none, in which case we say that the language lacks a gender system. Such is the case with several of the major families of Asian languages (e.g., Mandarin Chinese). Tamil, a member of the Dravidian family in south India, divides nouns into "rational" (i.e., people, gods) and "non-rational" (animals, and everything else), and further subdivides rational gender into "masculine" and "feminine" (Corbett, 1991:8–10). Thus, Tamil employs a "natural gender system", which means that given the semantics of a noun we can predict its gender, and vice-versa. English, a Germanic language, has a natural

gender system like Tamil, reflected only in personal, possessive, and reflexive pronouns. There are a few exceptions to semantic association: "she" may be used for a ship or country, "he"/"she" for an animal (of unknown sex), and "it" for downgrading humans (Mathiot and Roberts, 1979). Other languages show a less well-defined assignment based on semantics: Zande, a language spoken mainly in the Democratic Republic of the Congo, assigns nouns generally to four genders: masculine, feminine, animal, and neuter (Corbett, 1991:14). There are, however, about 80 exceptions, including such concepts as heavenly and metal objects, and edible plants, which are placed in the animal gender. Dyirbal, an Australian language, also has four genders, denoted by "class I, II, III, and IV". It has been shown (Dixon, 1972:308–12) that male humans and non-human animates belong to class I; female humans, water, fire, and fighting to class II; non-flesh food to class III; and everything else to class IV. Thus, the rules are semantic but non-obvious. However, children learning the language do not appear to learn the gender of nouns individually.

Turning now to typical Indo-European languages, we see an even smaller dependence on semantics. Nouns denoting people — assigned to masculine or feminine gender according to sex — are a minority. The "exceptions" (non-sexed objects assigned to either of those two genders) are the majority, thus making the semantic association a rather useless predictor for the gender of a noun. This fact, as we shall see, is very important for a correct assessment of B&S's work.

B&S's Experiment 1

In their first experiment, B&S investigated whether "the grammatical genders of nouns do in part reflect the properties of their referents" (Boroditsky and Schmidt, 2000:2). If true, they predicted "a correspondence in the assignment of genders across languages, and also a correspondence between Spanish and German genders and English speakers' naive intuitions". Although their testing of the prediction of correspondence across languages was rather inadequate (regarding the number of languages; I improve this test in the present study), they did a more thorough test of the naïve intuitions of 15 English speakers, none of whom were familiar with either Spanish or German (though we do not know if they were monolinguals). The subjects were asked to exclusively classify each of 50 animal names and 85 names of artifacts as either masculine or feminine (B&S do not give a list of those words).

Their comparison of gender agreement between Spanish and German yielded a correlation coefficient of $r = 0.21$, $p < 0.05$. This, they termed an "appreciable agreement". Although I would think a value of $r = 0.21$ (hence, $r^2 = 0.04$) indicates a rather appreciable *dis*agreement, B&S pointed out that the two languages "agreed more on the genders of animals (r=.39, p<.01), [than] on the genders of artifacts (r=.10, p<.35)".

To test B&S's hypothesis on the agreement of gender across languages, I examined 84 common nouns in 14 Indo-European languages. The nouns were chosen so that they represented more-or-less common referents: 20 artifacts, 22 natural objects, 20 abstract ideas, and 22 animals. The 14 languages were chosen so that a fairly representative set of the Indo-European family tree was obtained (three Germanic: Dutch, German, Icelandic; four Romance: French, Italian, Spanish, Portuguese; three Slavic: Polish, Russian, Serbo-Croatian; one Celtic: Irish; and also Albanian, Greek, and Kurdish.) Native speakers verified my choices of nouns (originally collected from dictionaries) for all languages but Albanian, Dutch, and Icelandic. The full assignment of genders is given in Appendix A.

The results of my study show that, predictably, the closer languages are in the family tree, the more they agree on gender. Languages as close linguistically as Portuguese and Spanish show a coefficient of determination[1] $r^2 = 0.75$. However, the coefficient between Spanish and German is $r^2 = 0.09$, $p < 0.01$ (so, $r = 0.30$; compare with B&S's $r = 0.21$), and the one between Spanish and Russian is $r^2 = 0.03$, exhibiting a complete uncorrelatedness (see Table 1). Overall, languages that belong to different subfamilies (e.g., a pair formed by a Romance and a Germanic language) show appreciable disagreement. For languages in the same subfamily, the part on which they agree — as given by the coefficient r^2 — is explicable not by reference to any inherent common intuition of people on the *sex* of things like a book and a tree, but by reference to the fact that Indo-European languages evolved from a common ancestor language, which employed gender, probably one with a strong semantic basis. As languages diverged, so did gender assignments, precisely *because* there is no objective and universal basis on which to decide the gender/sex of "flower", or the idea of "war", or even the words for "cat" and "butterfly". (See Appendix A: each of these words is nearly evenly assigned — close to 50% — between the masculine and feminine genders.) Table 1 shows the coefficients of determination (r^2) between the 14 languages.

B&S's second prediction is that English native speakers' naïve intuitions about the gender of nouns

[1] Since I observed no negative correlation, I prefer to use r^2, the *coefficient of determination*, rather than r, the correlation coefficient, because the former has a natural interpretation, which the latter lacks: r^2 shows the proportion of variation in one population that is explained by the variation in the other population. To be precise, I should employ the non-parametric r_s^2: *Spearman rank coefficient of determination*, since the populations are highly non-normal. However, in our case differences between r^2 and r_s^2 appear only in the second decimal place, so I will keep referring to r^2 in order to facilitate the comparison with B&S's results.

Table 1: Coefficients of determination (r^2) for the 14 languages, plus English monolinguals ('En', last row).

	French												
Fr		Italian											
It	.32		Portuguese										
Pt	.37	.32		Spanish									
Sp	.44	.24	.75		Dutch								
Du	.04	.00	.00	.01		German							
Ge	.07	.03	.06	.09	.24		Icelandic						
Ic	.14	.12	.17	.21	.07	.19		Irish					
Ir	.01	.00	.03	.04	.01	.02	.01		Polish				
Pl	.05	.02	.06	.13	.06	.14	.11	.01		Russian			
Ru	.00	.01	.01	.03	.02	.06	.04	.00	.29		Serbo-Croatian		
Se	.04	.09	.02	.03	.06	.05	.06	.01	.18	.27		Albanian	
Al	.11	.11	.18	.15	.01	.13	.16	.02	.01	.01	.00		Greek
Gr	.14	.10	.09	.11	.00	.14	.15	.00	.03	.02	.14	.19	
Ku	.07	.02	.08	.10	.09	.11	.09	.00	.04	.01	.01	.15	.04
En	.00	.01	.01	.01	.07	.05	.03	.01	.03	.04	.03	.03	.01

Kurdish **avg**

.11 **.03**

should show a correspondence with the assignment of gender in other Indo-European languages. To test this prediction I asked 20 monolingual native American English speakers (10 males and 10 females) to assign a gender, either masculine or feminine, to each of the 84 nouns listed in Appendix A. Subjects showed a remarkable consistency among themselves (average standard deviation $s = 0.18$), especially for words that have a natural association with maleness and potency (e.g., hammer, boulder, thunder, attack, war, lion), or with femaleness and beauty (e.g., flower, happiness, love, butterfly). The average assignments of genders by English monolinguals form a 15th population, which was compared against each of the 14 studied languages to determine the degree of correlation. The last row in Table 1 shows the values of r^2 for each case. We see that the opinion of native English speakers on gender shows a very weak correlation with each of the 14 languages, except possibly Kurdish (which can be attributed to statistical error). No negative correlation was observed. The average of all r^2 is $\langle r^2 \rangle = 0.03$. The *p*-values (indicating linear relationships) are statistically insignificant ($p > 0.05$) for all languages but Dutch, German, and Kurdish. However, it should be noted that the *p*-values are bound to converge to zero given a large enough sample size. What is important is not *whether* a linear relationship exists, suggested by the *p*-values, but *the magnitude* of correlation, given by r^2.

To explain why the correlation between English speakers' intuitions and gender assignment in the 14 studied languages is so weak, we must understand the cognitive processes of gender acquisition in such languages. Young learners of Indo-European languages with formal gender might notice the close correlation between gender and sex *when the noun being referred to is a person* (or even a pet of a known sex). However, learners could not miss noticing the clear *unrelatedness* of gender and sex when the object being referred to is not an animal, and thus lacks sex. In the young learner's world, the nouns for which gender and sex correlate nicely are a small minority compared to those for which the two notions cannot be correlated (because sex is *not* one of the perceived properties of the object referred to by the noun). The situation is depicted in Figure 1.

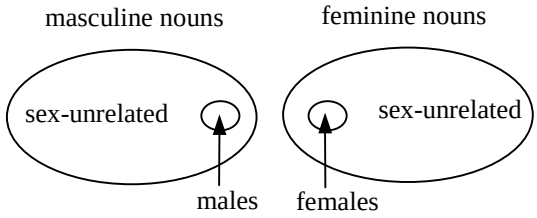


Figure 1: Gender vs. sex in "formal gender" languages

The sizes of the areas in Figure 1 are schematic, but relevant. Assuming the learner's cognitive mechanisms are tuned toward noticing the statistics and learning the regularities of this world, we conclude that the learner of such a language should not find the linguistic category of gender a particularly good predictor of the cognitive percept of sex. We should note that, at an early (pre-school) age, the learner is oblivious to the fact that the name of an observed category of nouns is "masculine", a word closely associated with maleness, while another category is called "feminine"; the learner simply notices the categories. Later, during formal education, the suspected (weak) relation between the notions of gender and sex may be reinforced, but it happens at a time when the learner has already acquired the linguistic category of formal gender, and has already noted that, as Figure 1 suggests, gender is not a good predictor of sex, and the two notions are only loosely related.

On the contrary, learners of languages that employ "natural gender", such as English, notice the close correlation between gender and sex. For such languages, the situation is depicted in Figure 2.

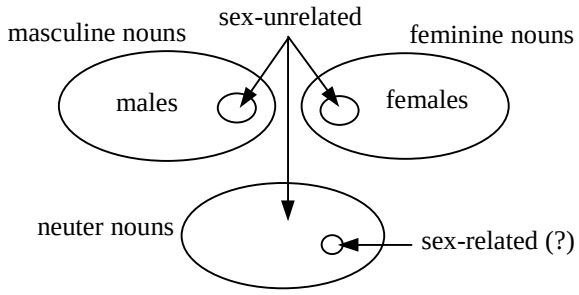


Figure 2: Gender vs. sex in "natural gender" languages

In this case, the intersection of masculine, feminine, and neuter nouns having the "correct" correlation with the percepts "male", "female", and "other" is large. The close correlation between gender and sex thus turns the percept of sex into a *good predictor* of the grammatical category of gender, and vice-versa. This fact may lead speakers of languages employing natural gender into *conflating* the two ideas, and possibly, as the B&S paper implicitly suggests, thinking that native speakers of languages with formal gender may perform a similar conflation. We should note in passing that one of the meanings of the word "gender" in English is "the state of being male, female, or neuter; sex" (Oxford English Dictionary, 1993). Thus, in English, the question "what is your gender?" is a meaningful one to ask a person. In Greek on the other hand, a typical Indo-European language employing formal gender, the same question ("pio eeneh to genos sou?") is absurd, because it implies the questioned entity is a *noun* —akin to asking in English: "what is your declension?"

B&S's Experiment 2

In their second experiment, B&S attempted to test whether "people's ideas about the genders of objects are strongly influenced by the grammatical genders assigned to these objects in their native language" (Boroditsky and Schmidt, 2000: Abstract). B&S based their hypothesis on an earlier study (Konishi, 1993) where German and Spanish speakers judged nouns that were masculine in their languages to be higher in potency than feminine ones, and the tested nouns belonged to opposite genders in the two languages. Subjects assigned subjectively a potency value for each noun, on a 7-point scale. B&S presented 24 pairs consisting of a noun (e.g., "spoon") and a proper name (e.g., "Erica") to 16 German and 25 Spanish native speakers during a learning phase. All nouns were given in English. The subjects' memory of *the sex of the proper name* that had been associated with a noun was examined during the testing phase. As expected, subjects were better able to remember the correct sex (82% correct) when the sex (e.g., "female") matched with the gender (e.g., "feminine"), than when this was not the case (74% correct). Since the nouns were chosen to have opposite genders in the two languages, subjects showed opposite memory biases. B&S concluded that "people's ideas about the genders of objects are strongly influenced by the grammatical genders assigned to those objects in their native language."

As with experiment 1, what is important is not the observation that there is an interference in memory retention between gender and sex, but the *explanation* for this phenomenon. B&S tacitly assume people make a direct connection between the concepts "masculine" and "male", and between "feminine" and "female". This direct connection may be "traversed" in the Spanish speaker's mind when presented with the word "moon" (in Spanish: " luna", of feminine gender), so that they match the "femaleness" of the moon with the femaleness implied by a name like "Karla". A German speaker performing the same task (being presented with "moon" – "Karla") would experience inhibition between moon's "maleness" (in German: " Mond", masculine), and Karla's femaleness, resulting in slightly worse memory performance.

Plausible as this explanation might appear, it makes more sense in the mind of a native speaker of a natural gender language (such as English), where "male" – "masculine" and "female" – "feminine" nearly coincide conceptually. For a native speaker of a formal gender language this explanation seems to be simplistically projecting the natural-gender speaker's view of the world onto everyone else. An alternative explanation is that the interference is caused by a much more indirect relation between noun and proper name than what B&S hypothesize. For example, the word "moon" in a Spanish speaker's mind evokes involuntarily, instantly, and subliminally, the Spanish word "luna". This word is of feminine gender, and is related to the feminine ending "-a", the pronoun " ella", and so on. The proper name "Karla" is also of feminine gender in the Spanish speaker's mind (75% of all female names tested by B&S ended in "-a", the marker of Spanish morphology for feminine nouns), and thus instantly and subliminally related to the same grammatical items ("ella", "feminine", etc). We should note that I make no reference to "moon's sex" in this conceptual plan. In other words, there is a lot of overlap in linguistic connections between "moon"–"luna" and "Karla" in the Spanish speaker's mind.[2] No experimental setting can sever these linguistic connections, and allow us to test exclusively the connections "feminine" – "female" and "masculine" – "male". I do not claim that such direct connections do not exist in the mind of a formal gender language's speaker. Such connections *do* exist. They may be learned in school, where the words for "masculine" and "feminine" are used as labels for categories of nouns the native speaker has already acquired at a very early stage; or they may be based on the small number of sex-related nouns. What I do claim is that experiments such as the one described by B&S (and Konishi) do not necessarily detect the direct influence of a supposed "sex of nouns" on cognition in speakers of languages with formal gender, but instead the very intricate and indirect connections between gender and sex in such languages, which are of both a perceptual as well as a linguistic nature.

[2] This argument is weaker for German speakers, but then we are not given the difference in performance between German and Spanish subjects in B&S's second experiment.

Evolution of Gender

What could the origin of grammatical gender be? B&S hint at possible common intuitions of people across languages, and attempt to quantify this assumption by examining the intuitions of speakers of English. I performed a similar comparison of such intuitions against Indo-European languages, and found that such intuitions do not show any particular correlation with the studied languages (Table 1). Moreover, it would be meaningful to talk about such a correlation *if languages agreed among themselves*. Otherwise, if we find a correlation between the intuitions of monolingual speakers of English and, say, Kurdish, we do not have any reason to assume there is anything other than chance involved. Looking back at the data in Table 1, we see that the only agreement that can be observed among languages is *between members of the same subfamily* (e.g., Portuguese–Spanish, etc.). The more phylogenetically distant the languages, the lower their correlation is (allowing for statistical errors). This hints at a possible answer to the gender-origin question.

That all Indo-European languages evolved from a common ancestor is indisputable. It is plausible to assume that this ancestor language employed a gender system, possibly one with a semantic basis. But what could have caused its modern descendants to assign genders such as masculine and feminine to inanimate objects? And how can a "pure" system (I mentioned Tamil as an example in the introduction) evolve into the modern chaos and disagreement?

The answer some authors have given to these questions is that the origin of gender is purely formal: some suffixes of sex-differentiable nouns acted as attractors, and created the genders in a purely formal, non-semantic way (Brugmann, 1899). This leaves open the question of what caused sex-differentiable nouns, rather than any other category, to become attractors. Another possible answer is that in some languages the initially semantic neuter gender was lost, and the void was filled by masculine and feminine genders being assigned to previously neuter nouns. Such a process can be observed today in Russian, where neuter nouns are only 13% of the total, and loanwords entering the language go primarily to the masculine gender, but also to the feminine (Corbett, 1991:317). This hypothesis does not take into account languages that retain the neuter gender, and still assign masculine and feminine genders to inanimate objects (German, Greek, etc.).

An alternative hypothesis is that masculine and feminine assignments to inanimate objects existed even in the original Indo-European ancestor. Although such assignments seem nonsensical today, they might have "made sense" in the remote past, at least among the few speakers of the ancestor language, based on animistic conceptions of the world. It could have appeared natural to a particular culture that, for example, a stone is of female sex. However, as the original language evolved, ideas about the stone's sex changed, too. Since there is no objective way to agree on something like the sex of a stone, the "opinions" among descendant languages evolved differently. What we observe today appears as a purely formal and arbitrary assignment, since the original "reasons" have been lost. One prediction of this hypothesis is that gender evolution in such languages should be traceable through a weak agreement between phylogenetically proximal languages. I believe the present work supports this implication, although further investigation of the hypothesis is clearly needed.

Acknowledgements

I would like to thank my advisors, Douglas Hofstadter, for bringing B&S's paper to my attention and motivating my work; and Michael Gasser, for his valuable comments and suggestions. Also, Nathan Basik, for his careful proofreading of the text.

References

Boroditsky, Lera and Lauren A. Schmidt (2000). *"Sex, Syntax, and Semantics"*. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. Philadelphia, PA.

Brugmann, K. (1899). "Das Nominalgeschlecht in den indogermanischen Sprachen". *Internationale Zeitschrift für allgemeine Sprachwissenschaft*, no. 4, pp. 100–9.

Corbett, Greville (1991). *Gender*: Cambridge University Press.

Dixon, R. M. W. (1972). *The Dyirbal Language of North Queensland*. Cambridge: Cambridge University Press.

Hockett, Charles F. (1958). *A Course in Modern Linguistics*. New York: Macmillan.

Imai, Mutsumi and Dedre Gentner (1997). "A cross-linguistic study of early word meaning: universal ontology and linguistic influence". *Cognition*, vol. 62, no. 2, pp. 169–200.

Konishi, Toshi (1993). "The Semantics of Grammatical Gender: A Cross-Cultural Study". *Journal of Psycholinguistic Research*, vol. 22, no. 5, pp. 519–534.

Mathiot, M. and M. Roberts (1979). "Sex roles as revealed through referential gender in American English". In M. Mathiot (ed.), *Ethnolinguistics: Boas, Sapir and Whorf Revisited*, 27 pp. 1–47. The Hague: Mouton.

Oxford English Dictionary (1993). *The New Shorter Oxford English Dictionary*: Oxford University Press.

Whorf, Benjamin L. (1956). *Language, Thought, and Reality: selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Appendix A: Words Examined

The 84 words in four categories are listed below. For the abbreviations used for the 14 languages see Table 1 (in text). The codes of gender values are as follows: -1 for masculine, 0 for neuter, 1 for feminine. Any intermediate values reflect the fact that more than one assignment was possible for a noun (e.g., "sea" in German and Spanish), or more than one noun of differing gender corresponded to the same concept. Blanks indicate that I could not obtain the appropriate gender (or the word is not native to the language). The last column (En) presents the average assignments of 20 American English monolinguals. Words marked with a star (*) were disambiguated for subjects who were asked to assign a gender as follows: table (furniture); chair (furniture); fork (utensil); bridge (over river); paper (a sheet of); bed (furniture); key (locking a door); watch (measuring time); star (on sky); gold (metal); power (of ideas, of wealth); revolution (of people).

	Fr	It	Pt	Sp	Du	Ge	Ic	Ir	Pl	Ru	Se	Al	Gr	Ku	En
Artifacts															
door	1	1	1	1	1	1	1	-1	0	1	1	1	1	-1	.00
wall	-1	0	1	1	-1	1	-1	-1	1	1	-1	-1	-1	1	-.10
table*	1	1	1	1	1	-1	0	-1	-1	-1	-1	1	0	1	.47
chair*	1	1	1	1	-1	-1	0	1	0	-1	1	1	1	0	-.20
spoon	1	-1	1	1	-1	-1	0	1	1	1	-1	1	0		.60
fork*	1	1	-1	-1	1	1	0	-1	-1	1	1	-1	0		.16
knife	-1	1	1	-1	0	0	-1	1	-1	-1	-1	1	0	1	-.50
car	1	1	-1	-1	1	0	0	-1	-1	1	1	1	0		-.50
house	1	1	1	1	0	0	0	-1	-1	-1	1	1	0	0	-.16
bridge*	-1	-1	1	-1	1	1		-1	-1	-1	-1	1	1		-.20
pistol	-1	1	1	1	0	1	1	-1	-1	-1	-1	1	0	1	-.70
book	-1	-1	-1	-1	0	0	0	-1	1	1	1	-1	0	0	.16
paper*	-1	1	-1	-1	0	0	-1	-1	-1	1	0	1	0	1	.20
bed*	-1	-1	1	1	0	0	1	1	0	1	-1	-1	0	1	.47
hammer	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0		-1.00
key*	1	1	1	1	-1	-1	-1	1	-1	-1	-1	-1	0	1	-.58
hat	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	0	1	-.30
shirt	0	1	1	1	0	0	1	1	1	1	1	1	0	-1	-.40
watch*	1	-1	-1	-1	0	1	-1	-1	-1	-1	-1	1	0		-.50
pencil	-1	1	-1	-1	0	-1	-1	-1	-1	-1	1	-1	0	-1	-.20
Natural Objects															
sky	-1	-1	-1	-1	1	-1	-1	1	0	0	0	-1	-1	1	.60
sun	-1	-1	-1	-1	1	1	1	1	0	0	0	-1	-1	0	-.10
moon	1	1	1	1	1	-1	0	1	-1	1	-1	1	0	1	.20
star*	1	1	1	1	1	-1	-1	1	1	1	1	-1	0	1	.40
tree	-1	1	1	-1	-1	-1	0	-1	0	0	0	1	0	1	-.30
sea	1	-1	-1	0	1	0	-.5	1	0	0	0	-1	1	1	.20
river	-1	-1	-1	-1	1	-1	0	1	1	1	1	-1	0	-1	.26
thunder	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	1	-1.00
rain	1	1	1	1	-1	-1	0	1	-1	-1	1	-1	1		.70
forest	1	1	1	-1	0	-1	-1	1	-1	-1	1	-1	0	0	-.30
boulder	-1	-1	1	1	1	-1	-1	1	1	1	1	-1	1		-1.00
mountain	1	1	1	1	-1	-1	0	-1	1	1	1	-1	0	-1	-.68
lake	-1	-1	-1	-1	0	-1	0	-1	0	0	0	-1	1	1	.50
air	-1	1	-1	-1	1	1	0	-1	0	-1	-1	-1	-1	0	.20
wind	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	-1		.50
earthquake	-1	-1	-1	-1	1	0	-1	-1	0	0	-1	-1	-1	1	-.40
stone	1	1	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	-.80
flower	1	-1	1	1	1	1	0	-1	-1	-1	-1	1	0	1	.90
gold*	-1	-1	-1	-1	0	0	0	-1	0	0	0	-1	-1	-1	-.10
water	1	1	1	1	0	0	0	-1	1	1	1	-1	0	1	.58
island	1	1	1	1	0	1	1	0	1	-1	0	-1	0	1	-.05
fire	-1	-1	-1	-1	0	0	-1	1	-1	-1	1	-1	1	-1	-.70

	Fr	It	Pt	Sp	Du	Ge	Ic	Ir	Pl	Ru	Se	Al	Gr	Ku	En
Abstr. Ideas															
justice	1	1	1	1	1	1	0	-1	1	1	1	1	1	1	-.50
freedom	1	1	1	1	-1	1	.5	1	1	1	1	1	1	1	.00
democracy	1	1	1	1	1	1	0	-1	1	1	1	1	1		-.30
idea	1	1	1	1	0	1	1	-1	1	1	1	1	1	1	.20
group	-1	-1	-1	-1	-1	1	-1	-1	1	1	1	-1	1	1	.20
anger	1	1	1	1	1	-1	1	1	-1	-1	0	-1	-1	1	-.70
surprise	1	1	1	1	1	1	1	-1	0	-1	0	1	1		.60
question	1	1	1	1	1	1	1	1	0	-1	0	1	1	1	.26
hunger	1	1	1	1	-1	-1	0	-1	-1	-1	1	1	1	1	-.37
power*	-1	-1	-1	1	1	1		1	1	1	1	1	1	1	-.70
love	-1	1	-1	-1	1	1	1	-1	1	1	1	1	1	1	.79
revolution*	1	1	1	1	1	1	1	1	1	1	1	-1	1		-.70
friendship	1	1	1	1	1	0	-1	1	1	0	1	1	1	1	.60
war	1	1	1	1	-1	-1	0	-1	1	1	-1	1	-1	.3	-.89
religion	1	1	1	1	-1	1	1	-1	1	1	1	1	1	0	.30
answer	1	1	1	1	0	1	0	-1	1	-1	-1	1	1	1	.05
happiness	-1	1	1	1	0	0	1	-1	0	0	1	1	1	1	1.00
sadness	1	1	1	1	0	1	0	-1	-1	1	1	1	1	1	.70
attack	1	-1	-1	-1	-1	-1	1	-1	-1	0	-1	-1	1	1	-.90
defense	1	1	1	1	.5	1	1	1	1	1	1	1	1	1	-.60
Animals															
cat	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	.58
dog	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-.80
horse	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	-1	0	-1	-.10
lion	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0		-.90
elephant	-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-.60
snake	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	0	-1	-.90
tiger	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-.50
antelope	1	1	-1	-1	1	1	1	-1	1	1	1		1		.10
ant	1	1	1	1	1	1	-1	-1	1	-1	-1	1	0	1	.00
fly	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.30
butterfly	-1	1	1	1	-1	0	0	-1	-1	1	-1	1	1	1	.90
bee	1	1	1	1	1	1	.5	1	1	1	1	0	1		.50
bird	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	0	-1	.60
wolf	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-.90
fox	-1	1	1	0	-1	-1	.3	-1	-1	1	1	1	1	-1	-.20
fish	-1	-1	-1	-1	0	-1	-1	-1	1	1	1	-1	0	-1	.37
sparrow	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	0		.50
penguin	-1	-1	-1	-1	0	-1		1	-1	-1	-1	-1	-1		.20
chimp.	-1	-1	-1	-1	0	-1	-1	-1	0	-1	0		-1		-.30
bear	0	0	-1	-1	-1	-1	-.5	-1	0	-1	-1	-1	1		-.80
spider	1	-1	1	1	1	1	1	-1	-1	-1	-1	1	1		.10
whale	1	1	1	1	-1	0	-1	-1	-1	-1	-1	1	1	-1	-.60

Recovering Context After Interruption

Jerry L. Franke (jfranke@atl.lmco.com)
Jody J. Daniels (jdaniels@atl.lmco.com)
Daniel C. McFarlane (dmcfarla@atl.lmco.com)
Lockheed Martin Advanced Technology Laboratories
1 Federal Street, A&E-3W, Camden, NJ 08102 USA

Abstract

As information systems become more complex and present an increasingly rich amount of information to users, interruptions present an ever larger hurdle to operational efficiency. User interface technologies intended to support increased user-control of the transitions between computer-based tasks can help mitigate that obstacle. We present a three-pronged approach that combines dynamic interruption coordination support with context review mechanisms to aid user navigation through interruptions. These mechanisms are implemented within a spoken dialogue interface system for a radio-based human-software agent military logistics task.

Introduction

Modern information technologies continue to successfully deliver ever more powerful information products. This increase in power, however, can support the user in performing tasks quickly and accurately only if users can exploit this increased information flow for their own needs. People have a limited capacity for information management that directly affects the quality of their decision-making in stressful real-world tasks. If the ever-increasing information stream is not properly managed, these human capacities could become overloaded. The net result is that the adoption of a new information technology can actually cause an overall decrease in mission performance.

New, more powerful information technologies may increase the volume of important information delivered to decision-makers, but at the same time increase the frequency of interruptions of those decision-makers, degrading their information management capacity. The number of alerts that interrupt users affects how they manage their limited attentional cognitive resources. An interrupting alert causes users to switch from their current task to the new alert task. After completing the alert task, users must switch tasks again to resume what they had been doing prior to the interruption. The cognitive demands of these context switches increase the effective workload of users, which in turn increases the probability of mental mistakes.

For example, Foushee and Helmreich (1988) found that the disruptive effects of interruptions have caused flight errors in commercial airline flights, sometimes resulting in fatal crashes. Human interruption is also a recognized problem in domains such as Navy command and control systems (Osga, 2000) and flightdeck or cockpit systems (Barnes, 1990; Adams and Pew, 1990; Adams et al., 1995).

The literature is rich with descriptions of the cognitive limitations people have relative to resuming tasks after being interrupted. Miyata and Norman (1986) give a general overview of the topic, discussing foregrounded and backgrounded activities and how interruptions are the standard way people switch between tasks in multitasking. Liu and Wickens (1988) discuss task interference and the effect of task type in human multitasking. McFarlane (2002) provides an in-depth review of the published relevant theory and proposes both a definition of human interruption and a taxonomy for classifying human interruptions.

Other studies investigate the causes of the disruptive nature of interruptions. McLeod and Mierop (1979) examine the effect of task similarity for manual tasks. Zijlstra and Roe (1999) found that the frequency of interruptions in an office environment affects the length of delay for people resuming the main task. Latorella (1998) found a modality interaction effect between how interruptions are presented (aurally or graphically) and the type of task that cockpit crew members perform (aural or graphical); different combinations of interface solution and task type resulted in different kinds of adverse effects on crew behavior. Linde and Goguen (1987) found that differences in how cockpit crews interact with each other affect their ability to successfully handle interruptions.

The objective of human alerting technology is to cancel the negative effects of human interruption and allow users to exploit the benefits of greater information volume for making better decisions. Human alerting mechanisms are being integrated within a broad range of commercial and military applications. These include announcement mechanisms for relatively less important systems like email, telephone, voicemail, internet instant messaging, chat rooms, automated help systems (like Microsoft's "Clippy"), computer-based tutor-

ing, and shopping agents. These applications also include many mission-critical systems including military command and control (C2), aircraft flightdeck control, power plant operations, spacecraft control centers, and real-time targeting sentinel-agent systems. McFarlane and Latorella (2002) present an in-depth discussion of the scope and importance of human interruption for HCI design.

Approach

There are three basic strategies for improving human performance on an interruption-laden multitask: (1) training (Hess and Detweiler, 1994); (2) selection of users (Joslyn and Hunt, 1998; Joslyn, 1995); and (3) user interface design. Due to the constraints of our real-world applications, we have focused our approach on the last option.

Our objective is to support efficient task recovery after interruption. It is useful to divide many user interface design approaches for human interruption into three phases. The pre-interruption phase prepares the user to transition from the main task to the interrupting task. The mid-interruption phase generally focuses on the user's transition to the interrupting task and includes the user's efforts and ability to maintain situational awareness of the main task while working on the interrupting task. The post-interruption phase sees the user return and reorient to the context of the original task that was interrupted.

Our approach has three parts, matching each of the three interruption phases.

Pre-interruption

Before the actual interruption takes place, the interface should give the user support for quick rehearsal of the current task before switching context to handle the interruption. Gillie and Broadbent (1989) noted that rehearsal may have potential for aiding human interruption in user interface design. Storch (1992) suggests that rehearsal may be useful in diminishing the negative effects of interruption after obtaining unexpected results in experiments unrelated to rehearsal. Detweiler et al. (1994) describes two experiments related to early warnings of interruptions that indicate that providing warnings is only marginally useful if the interruption task has a low memory load and is dissimilar to the main task while providing warnings can be extremely useful if the interruption task has a high memory load and is similar to the main task.

To allow the user to rehearse before interruption, some cue must precede or accompany the incoming alert. This cue helps to differentiate between the main task and interrupting task contexts and can take many forms, such as a visual flash, an audible beep, or a vibration. Because our particular applications involve a spoken dialogue interface, we have differentiated incoming alerts from the current task by having the interface use a different voice. For example, the interface may use a female voice while participating in dialogue related to the user's current task, then switch to a male voice when introducing the interruption to the user. This cue gives the user the opportunity to register the alert, allowing the user to rehearse the context of the main task before continuing into the interruption.

Mid-interruption

When the interruption occurs, the interface should support user control of context switching and help the user maintain situational awareness of backgrounded tasks. This switch can take many forms. McFarlane (2002) conducted a theory-based experiment that compared four basic alternative solutions to the problem of how to coordinate human interruption in computer user interfaces. These four solutions are: (1) interrupt immediately and get it over with; (2) provide negotiation support so that the user controls the timing and exact context of switching between tasks; (3) provide intelligent mediation that brokers the onset of interruption tasks on behalf of the user; and (4) the use of scheduled interruption time cycles so that interruptions only occur during set times or contexts. Of these four solutions, negotiation was measurably the best approach for all kinds of user performance, except in cases where even small differences in the timeliness of handling interruption tasks are critical (either the current task is too important to allow distraction by the negotiation process, or the interrupting task is too important to wait for the negotiation to be completed).

Our approach involves the intelligent, automated selection of interruption strategy on a case-by-case basis. Our selection criteria is based on a dynamic automated assessment of the relative importance between the current task and the interrupting task. If the interrupting task is mission critical compared to the current task, the user is interrupted immediately. If the current task is critically important compared to the interrupting task, the alert is held until the user is finished with the current task (that is, it's scheduled for the next cognitive break). In all other cases, the interruption is negotiated.

To further aid the user in assessing relative task importance, we vary the default option in negotiation. That is, if the interrupting task appears to be slightly more important than the current task, the default option for the user is to accept the interruption. If the interrupting task is not deemed to be of higher importance, the default option for the user is to defer the interruption until the next cognitive break. Table 1 presents the full interruption strategy selection process for a three-valued priority system.

Interruption Task	Current Task		
	high	medium	low
high	Negotiated, default defer	Negotiated, default interrupt	Interrupt immediately
medium	Negotiated, default defer	Negotiated, default defer	Negotiated, default interrupt
low	Defer interruption until cognitive break	Negotiated, default defer	Negotiated, default defer

Table 1: Interruption Strategy Selection based on relative priority

Post-interruption

After the interruption is complete and the user transitions back to the original, interrupted task, the interface should provide recovery support to the user. That is, it should provide mechanisms to aid the user in recalling the context of the interrupted task, helping the user return more quickly to that previous task. Malin et al. (1991) state that user interfaces should be designed to reorient users to previously interrupted activities when they try to resume them. In their work, a simple log of relevant recent decisions is made easily available to the user for reference.

Our approach to context recovery involves providing the user commands that query the interface about aspects of the previous task. In a spoken dialogue system, this takes the form of meta-dialogue, with possible queries like "Where was I?" or "What was I last working on?" The user can also ask questions specific to the task, such as inquiring which supplies have been ordered so far in a requisition application.

Finally, the user can request a full progress review of the interrupted task. This provides a complete replay mechanism to the user, catching the user up to previous task context quickly and in detail. In a spoken dialogue system, this takes the form of requests for a summary of the task progress to-date.

Implementation

As a testbed for our approach to intelligent alerting and interruption management, we applied our techniques to a spoken dialogue interface. We have implemented a number of speech applications following the Listen, Communicate, Show (LCS) paradigm(Daniels, 2000). LCS systems integrate mixed-initiative spoken dialogue interaction with mobile intelligent agents to provide a natural, robust interface to information systems.

In most domains, users can use LCS to command agents to persistently monitor information sources for specific information events. When these events occur, the agents inform the LCS interface, which calls and alerts the user. If the user is currently engaged in another task, this agent-initiated conversation can result in an interruption.

The spoken dialogue portion of an LCS system is built upon the Galaxy architecture developed at MIT (Seneff et al., 1999). Galaxy supports distributed, plug-and-play systems in which specialized servers are coordinated through a centralized communication hub. LCS systems contain servers specialized for speech and natural language processing, a dialogue manager to direct the system's side of the conversation with the user, and an agent server for communicating and coordinating with the agent system.

Originally, when LCS monitor agents would notify the user, they would communicate to the dialogue manager directly through the agent server. The dialogue manager, which contained limited control mechanisms for interruption, would interject the interrupting alert at the next available moment in the dialogue. This would ensure that the user would not be interrupted mid-utterance, but does not take into account the effects of interruption on the user's cognitive state.

To integrate our new interruption techniques, we added several new servers to the LCS architecture (see Figure 1 for illustration). The priority server ascertains the relative priorities of the current and interrupting tasks. The dialogue manager keeps the priority server informed of the task in which the user is engaged, while the agent server communicates the priority of incoming alerts.

The interruption server selects the interruption strategy most appropriate for the relative priority determined by the priority server. Once the interruption strategy is determined, the interruption server supervises as the system enacts the strategy. If the interruption is deferred, the interruption server tracks it to make sure that the alert is eventually delivered.

Because negotiated interruptions require interaction about the interruption (rather than about the interrupting task itself), we implemented a dialogue manager to drive this interaction in a domain-independent manner. The negotiation manager controls the system's part of the negotiation process and

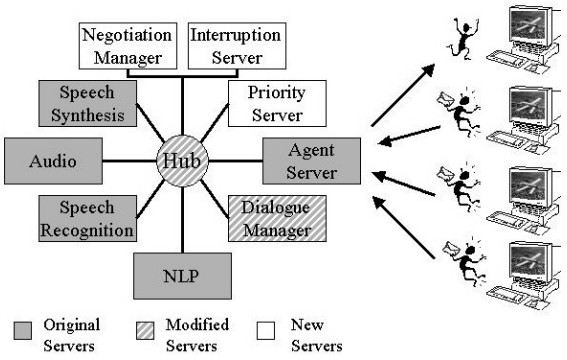


Figure 1: The ability to task multiple agents to perform persistent tasks (such as monitoring information systems) is a strength of LCS systems. However, since multiple agents may return results at the same time, advanced methods for handling interruptions were required. Some enhancements to the standard LCS architecture were required to implement these new interruption mechanisms.

coordinates with domain-specific dialogue managers to ensure that the system speaks to the user in a reasonable, focused manner.

In addition to constructing the new servers, we made several enhancements to the already existing LCS infrastructure. We implemented the meta-dialogue for post-interruption context recovery by adding logic for meta-dialogue control to the domain-specific dialogue managers. We also added an intuitive pre-interruption cue, programming the hub to have the system use voices for interruptions that are different from the voice used for the interrupted task.

Application

We applied the new interruption techniques to an LCS domain that supports Marines in managing requests for supplies using regular military radio protocols. This application was originally developed as part of the Small Unit Logistics (SUL) Advanced Concept Technology Demonstration (ACTD) program. The spoken language interface assists a Marine user in placing, modifying, deleting, or checking the status of a supply request. The SUL system also supports the creation of monitor agents to track requests and notify the user when either the status of the request changes or if the agent observes that the request hasn't been given attention over a set period of time.

In its original implementation, the SUL system would accomodate these returning notification activities by waiting until a break in the current conversation before providing any notification results to the Marine, regardless of the priority of either the notification activity or the current task. By allowing the SUL system to break into an ongoing conversation with important news, we can create a spoken dialogue interface that more realistically emulates radio protocols. However, this feature brings with it all the challenges associated with interruptions that have been discussed throughout this paper.

IMMEDIATE: *System:* `Break! Break! MAGTF-5, this is BSSG1. Alert! Urgent Rapid Request 1738 has changed status to be canceled. over`
NEGOTIATED - INTERRUPT: *System:* `Break! Break! MAGTF-5, This is BSSG1. Alert about Immediate Rapid Request 1738...Accept now? over`
NEGOTIATED - DEFER: *System:* `Break! Break! MAGTF-5, This is BSSG1, Alert about Routine Rapid Request 1738...Defer now?`
SCHEDULED: [No interaction until the user ends the current conversation. Then the system contacts the user.] *System:* `MAGTF-5, MAGTF-5, this is BSSG1, over.`

Figure 2: Example opening utterances for each interruption strategy. In all cases, a voice different from the one the user had been listening to was used for the alert. Note that when an interruption occurs, the user is explicitly informed of the new task's priority to support the user's decision to switch tasks.

To support interruption strategy selection, we established a priority comparision scheme based on the priority field of each logistics request. For a SUL request, there are three priorities: routine, immediate, and urgent. We mapped these priorities to the low, medium, and high priority scheme described earlier in Table 1. We used the interruption strategy selection method described in this paper to govern delivery of agent alerts. Figure 2 shows examples of how interruptions would be presented to the user for each strategy.

To support post-interruption context review, we implemented two sets of meta-commands, relying on radio protocol to guide us. In the first case, the system repeats just its most recently stated utterance from the prior conversation. For the SUL domain, the proword (that is, a military procedure word) "Read back" is used. In the second case, the system reiterates all information it has been given about the current task. Figure 3 shows an example in which the user has requested more than just the prior system utterance. For this, the proword "Read back my request" is used.

In addition, we implemented dialogue that allows the Marine user to examine specific parts of a supply request by querying the interface specifically about

that part. For example, the user might ask, "Who is the point of contact for this request?" or "How many grenades did I order?" This provides the user complete control in returning to the context of the interrupted supply request. Similar dialogue supports the user in orienting quickly to interrupting alerts about other supply requests as well.

The SUL spoken dialogue system, with alerting enhancements, has been demonstrated successfully multiple times in operational settings. The enhanced LCS alerting infrastructure is being used as the basis for several more applications that will be field tested in the near term.

Future Work

We are working toward several enhancements of the current LCS interruption mechanisms. In each case, the enhancements build upon a core capability present in the current system.

Our current use of overall task priority to select interruption strategy assumes that a coarse-grained decision is sufficient. A more-informed decision would result from a finer-grained knowledge of where the user is in the current task. For example, in the SUL domain, the system is programmed with knowledge of the information that is necessary to fully complete a supply request. With its programmed knowledge of the request process, the system should be able to ascertain how close the user is to the beginning or end of completing the task, or if the user is in the middle of clarifying a particular step in the task process.

We are also investigating adding multiple modalities, where possible, as a method to further differentiate interrupting tasks from the current task, taking advantage of modality effects like those described in (Latorella, 1998). While our use of different voices helps, the use of another modality to cue interruption may prove more helpful to the user in mitigating cognitive disruption. This approach will be implemented and tested in the field.

In addition, we plan to add finer control of context review. Currently, our system provides for review of prior context in one of two forms: either the most recent system utterance or the entire set of known information items that the system has. While this is quite useful, with long, complex tasks, an intermediate level of detail might be preferred. We plan to construct and test dialogue methods for giving the user that finer control.

Conclusions

The disruptive effects of human interruption by information systems is a serious concern, particularly in high stress situations such as military operations. There are technical solutions available at all points during the interruption process to help mitigate this problem. By alleviating this cognitive disruption, we can help the user move from task to task more quickly, resulting in potentially more efficient, less error-prone work by the user.

Acknowledgements

The authors would like to thank the other members of the Recovering Context After Interruption project, including Dan Davenport, Frank Davis, James Denny, Steve Knott, Dan Miksch, Mike Orr, Kathleen Stibler, Mike Thomas, and Ben van Durme. We would also like to thank Dylan Schmorrow, Program Manager of the DARPA Augmented Cognition program. This work was funded under ONR contract N00014-01-0455.

References

Adams, M. J. and Pew, R. W. (1990). Situational awareness in the commercial aircraft cockpit: a cognitive perspective. In *IEEE/AIAA/NASA 9th Digital Avionics Systems Conference*.

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors*, 37(1):85–104.

Barnes, V. E. (1990). Cockpit distractions: Precursors to emergencies. In *Proceedings of the Human Factors Society 34th Annual Meeting*.

Daniels, J. (2000). Integrating a spoken language system with agents for operational information access. In *Proceedings of Innovative Applications of Artificial Intelligence (IAAI-2000)*, Austin, TX.

Detweiler, M. C., Hess, S. M., and Phelps, M. P. (1994). Interruptions and working memory. Technical report, The Pennsylvania State University, Department of Psychology. Unpublished.

Foushee, H. C. and Helmreich, R. L. (1988). Group interaction and flight crew performance. In Wiener, E. L. and Nagel, D. C., editors, *Human Factors in Aviation*. Academic Press, Inc., San Diego, CA.

Gillie, T. and Broadbent, D. E. (1989). What makes interruptions disruptive? a study of length, similarity, and complexity. *Psychological Research*, 50(4):243–250.

Hess, S. M. and Detweiler, M. C. (1994). Training to reduce the disruptive effects of interruptions. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*.

Joslyn, S. (1995). *Individual Differences in Time Pressured Decision Making*. PhD thesis, University of Washington.

System: ⟨domain voice⟩ `Delivery time is oh four hundred hours. Awaiting delivery location...`
System: [alert voice] `Break! Break! MAGTF-5, this is BSSG1. Alert about Immediate Rapid Request 1376. Accept now? over.`
User: `Affirmative BSSG1.`
System: [alert voice] `Request 1376 has changed status from New to be Cancelled.`
User: `What was it requesting?`
System: [alert voice] `The supply for Rapid Request 1376 is 5000 five point five six ball rounds and 400 gallons of water.`
User: `Roger BSSG1, anything else?`
System: [alert voice] `Negative MAGTF-5. BSSG1 out.`
System: ⟨domain voice⟩ `MAGTF-5, This is CSSOC. Back to your request. Delivery time is oh four hundred hours. Awaiting delivery location...`
User: `Read back my request.`
System: ⟨domain voice⟩ `Roger, MAGTF-5. I read back new rapid request. Precedence is routine. Supply is 100 dodic alpha five five five and 200 cases of MREs. Delivery time is oh four hundred hours. Awaiting delivery location...`

Figure 3: Example of an interruption with recovery back to the prior converstaion. The user queries the system about the details of the interrupting task's request for quick orientation to the new task. Returning to the interrupted task, the user can request a full summarization of the request in its current state.

Joslyn, S. and Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied*, 4(1):16–43.

Latorella, K. A. (1998). Effects of modality on interrupted flightdeck performance: Implications for data link. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*.

Linde, C. and Goguen, J. (1987). Checklist interruptions and resumption: A linguistic study. Technical report, National Aeronautics and Space Administration, Ames Research Center, Moffett Field, CA. NASA Contractor Report NASA-CR-177460.

Liu, Y. and Wickens, C. D. (1988). Patterns of task interference when human functions as a controller or a monitor. In *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*, pages 864–867.

Malin, J. T., Schreckenghost, D. L., Woods, D. D., Potter, S. S., Johannesen, L., Holloway, M., and Forbus, K. D. (1991). Making intelligent systems team players: Case studies and design issues, vol. 1, human-computer interaction design. Technical report, National Aeronautics and Space Administration, Washington, DC. NASA Technical Memorandum 104738.

McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *HCI*, 27. to be published.

McFarlane, D. C. and Latorella, K. A. (2002). The scope and importance of human interruption in HCI design. *HCI*, 27. to be published.

McLeod, P. and Mierop, J. (1979). How to reduce manual response interference in the multiple task environment. *Ergonomics*, 22(4):469–475.

Miyata, Y. and Norman, D. A. (1986). Psychological issues in support of multiple activities. In Norman, D. A. and Draper, S. W., editors, *User Centered System Design*, pages 265–284. Lawrence Erlbaum Associates, Hillsdale, NJ.

Osga, G. A. (2000). 21st century workstations: Active partners in accomplishing task goals. In *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*.

Seneff, S., Lau, R., and Polifroni, J. (1999). Organization, communication, and control in the galaxy-ii conversational system. In *Proceedings for Eurospeech '98*, Budapest, Hungary.

Storch, N. A. (1992). Does the user interface make interruptions disruptive? a study of interface style and form of interruption. Technical report, Lawrence Livermore National Laboratory. UCRL-JC-108993.

Zijlstra, F. R. H. and Roe, R. A. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, 72:163–185.

Four Problems with Extracting Human Semantics from Large Text Corpora

Robert M. French and Christophe Labiouse
Quantitative Psychology and Cognitive Science
Psychology Department, University of Liege, Belgium
{rfrench, clabiouse}@ulg.ac.be

Abstract

We present four problems that will have to be overcome by text co-occurrence programs in order for them to be able to capture human-like semantics. These problems are: the intrinsic deformability of semantic space, the inability to detect co-occurrences of (esp. distal) abstract structures, their lack of essential world knowledge, which humans acquire through learning or direct experience with the world and their assumption of the atomic nature of words. By looking at a number of very simple questions, based in part on how humans do analogy-making, we show just how far one of the best of these programs is from being able to capture real semantics.

Introduction

"You shall know a word," wrote J. R. Firth in 1957, "by the company it keeps." This idea, in one form or another, underlies the statistical study of the co-occurrence of lexical items in large text corpora. This burgeoning field of research has been made possible to a large extent by the ready availability of vast databases of text that can be automatically scanned by computer. While we certainly do not dispute the value of the statistical study of large text corpora, we take issue with the claim that lexical co-occurrence alone can capture real-world semantics. We focus on four main problems with co-occurrence analysis programs:

- they do not take into account the intrinsic deformability of semantic space due to context-dependence;
- the cannot detect co-occurrences of abstract structures, especially when they are highly distal;
- they lack of essential world knowledge, which humans acquire through learning or direct experience with the world;
- they assume that words are "atomic" entities.

These issues are ones that will have to be effectively dealt with by text analysis techniques in order for them to capture even elementary human semantics. These points will be examined in the context of human analogy-making.

The remainder of this paper is organized as follows. We begin by discussing the relationship between analogy-making and concept meaning and show why current co-occurrence programs would have so much difficulty with this broader view of concept meaning. We then consider one of the best recent co-occurrence programs, PMI-IR (Turney, 2001a) and show just how far this program is from being able to plausibly respond (i.e., in a human-like manner) to even the simplest possible analogies.

The intrinsic deformability of semantic space

We will take issue with one of the main principles underlying LSA (Landauer and Dumais, 1997), HAL (Lund & Burgess, 1996) and other programs based on lexical analysis of large corpora – namely, that "The meaning of a word can be thought of as a location in semantic space and the dimensionality of that space and the location of any word within it can be recovered from estimates of the distance between word pairs." (Fletcher & Linzie, 1998). The implication is that words have stable, fixed locations in semantic space. While this is obviously not entirely false, this principle overlooks the fact that these locations in semantic space are *highly context dependent*. They not only can, but *must* be able to move considerably in semantic space depending on the context in which they are to be used.

Consider a very simple example. A "claw hammer" would, under most circumstances, be close in semantic space to terms like "ball-peen hammer," "hit," "pound," "nail," "saw" and, even, "club." However, if, while nailing a floor, you suddenly have a back itch, the "claw" part of the hammer will likely become much more salient as a back-scratcher, rather than a nail-remover. Your realization that you can use the hammer as a back-scratcher temporarily moves the object in semantic space much closer to "back-scratcher," "itch," etc., than when it is perceived only as an object with which one can drive in nails. This "relocation" of the meaning of a word/concept in semantic space based on context is at the very heart of analogy-making, of perceiving one object as an instance of another class of objects (Chalmers, *et al*, 1992; Hofstadter, 1995). It is therefore essential to any algorithm that claims to be able to automatically extract word meaning from very large text corpora And it is precisely this ability to relocate in semantic space in a context-dependent manner that is currently beyond the reach of all co-occurrence techniques.

In short, while co-occurrence techniques may plausibly situate a word in semantic space with respect to its *average* usage, this is not sufficient to capture the context-dependent shifts in word meaning required to understand even the simplest analogies on which so much of our cognition is based. For this we need to somehow extract, at the very least, abstract relational information concerning the word.

Detection of (distal) abstract structures

Acquiring the semantics of a particular word allows us to rate the quality of associations between that word and other words. To plausibly claim that a program has acquired, or even partially acquired, the semantics of a word means that it should give word-association ratings that are at least approximately similar to those given by humans (French, 1990). We will use this rating technique to judge the performance of text co-occurrence programs.

There are (at least) two different bases for these associations, even if this distinction is not always easy to characterize (Chalmers, French, & Hofstadter, 1992). To say, "John is a real beanstalk," refers to largely "surface" attributes of John and beanstalks — namely, they are both tall and thin. On the other hand, when we say "John is a real wolf with the ladies," we don't mean John grows long gray hair around women and bites them, but rather that his *relation* with women is socially predatory, analogous to a wolf's relation of physical predation with its prey. The first analogy is largely *attributional,* based essentially on common surface features (in this case, the attributes "tall" and "thin") of John and beanstalks, whereas the latter analogy is primarily *relational*, based on a mapping between John's behavioral interactions with women and wolves behavioral interactions with prey. The first kind of association can be captured by co-occurrence techniques, whereas the latter — the basis of virtually all deep analogy-making (Gentner, 1983) — is still well beyond the reach of these techniques.

Incorporating semantic information

An equally important difficulty involves the unavailability to these programs' of crucial semantic information that cannot be acquired merely by examining word co-occurrences. In two of the examples below this lack of crucial contextual knowledge — that fathers are always male in one example, and the fact that there is an undeclared war going on between the Israelis and the Palestinians in the other — causes the particular text co-occurrence analysis program under consideration to fail completely in responding to the simplest questions involving word meaning. Humans acquire this information through direct experience with the world or through explicit learning, whereas these programs currently have no way of acquiring it. The point is that, when making judgments about word meaning, people — unlike co-occurrence programs — make use of a wealth of relational and semantic information that is unrelated to word co-occurrence.

Words are not atomic entities

Consider an example of a "subcognitive" question from French (1988, 1990) involving the rating of a neologism. "On a scale of 1 (awful) to 10 (excellent) please rate:

- *Flugly* as the name of glamorous Hollywood actress,
- *Flugly* as the name of an accountant in a W. C. Fields movie."

Humans, of course, can do these two ratings without difficulty: *Flugly* is a decidedly lousy name for a glamorous Hollywood actress and a fine name for an accountant in a W. C. Fields movie or a teddy bear. But how do we "know" this, since you have never seen the word *Flugly* before? You know, at least in part, that *Flugly* doesn't work for the name of a glamorous actress because of its *component parts* (French, 1990). In particular, it contains an unpleasing-to-the-ear guttural "g," to say nothing of the syllable "ug" or the entire word "ugly." Similarly, we rate it as a good name for an accountant in a W. C. Fields' movie because, in our mind's ear, we hear him pronouncing the name as "Flugleeee." This requires phonetic information acquired by having heard W. C. Fields' unique manner of speaking (or having heard others imitating this manner) and by the fact that various components of *Flugly*, namely, "ly," can be transformed into a drawling "leeee."

The point is that words contain parts that contain crucial information that contributes to the overall meaning of the word. Co-occurrence programs are currently insensitive to this information. And it is not clear that by extending their analyses to the letter or syllable level that i) there would not be a problem of combinatorial explosion and ii) that this would be an appropriate way to acquire this information.

PMI-IR

In the examples that follow, we will consider the performance of one recent program, PMI-IR (Turney, 2001a, b), that, according to its author, outperforms all other current programs on the most widely used benchmark for programs that attempt to extract word meaning from large text corpora. This benchmark is their performance on the standard synonym selection tasks that are part of the Test Of English as a Foreign Language (TOEFL) and the test of English as a Second Language (ESL).

The co-occurrence technique used by PMI-IR is one of a family of "Pointwise Mutual Information" (PMI) techniques developed by Church & Hanks (1989) and Church *et al.* (1991). In order to calculate the conditional probability scores on which it bases its choice of the correct synonym, the program queries 350 million pages of English text indexed by the AltaVista search engine. The most sophisticated version of PMI-IR is able to make use of local (proximal) context in order to correctly answer questions such as, "Every year in the early spring farmers [tap] maple syrup from their trees (drain; boil; knock; rap)." As Peter Turney, the author of PMI-IR, points out, "the problem word *ta*p, out of context, might seem to best match the choice words *knock* or *ra*p, but the context *maple syrup* makes *drain* a better match for *ta*p" (Turney, 2001b). The program factors in the context provided by "maple syrup" to correctly answer this question.

The program does, indeed, perform impressively on the synonym recognition task. According to Turney, the

program produced the following results on the standard TOEFL and ESL synonym recognition task:

> "The task of synonym recognition is, given a problem word and a set of alternative words, choose the member from the set of alternative words that is most similar in meaning to the problem word. PMI-IR has been evaluated using 80 synonym recognition questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym recognition questions from a collection of tests for students of English as a Second Language (ESL). On both tests, PMI-IR scores 74% . . . For comparison, the average score on the 80 TOEFL questions, for a large sample of applicants to US colleges from non-English speaking countries, was 64.5% (Landauer and Dumais, 1997). . . . Latent Semantic Analysis (LSA), another statistical technique, scores 64.4% on the 80 TOEFL questions (Landauer and Dumais, 1997)."

Three examples

In what follows we use a word-rating technique from French (1988, 1990) and similar to standard similarity judgment techniques used to study how word meanings are represented (see, for example, Rips, Shoben, & Smith, 1973). The key idea is that these simple questions require non-local context for their answers (French & Labiouse, 2001).

Rating lawyers

Our first example involves the rating of *lawyers* as various other entities.

> "Rate on a scale of 1 (terrible) to 10 (excellent) rate *lawyers* as: horses, fish, telephones, stones, sharks, cats, flies, birds, slimeballs, kangaroos, robins, dogs, and bastards."

We applied the PMI-IR search technique described in Turney (2001b) using the Alta-Vista search engine and found that it gave the *lowest* (i.e., poorest) ratings to "Lawyers as slimeballs" (1.06) and "Lawyers as bastards" (1.15), the latter being roughly equivalent PMI-IR's rating of "Lawyers as kangaroos" (1.17)! We then asked a group of 26 undergraduates at Willamette University (Oregon) to also do these ratings. These results (Figure 1) are much more in line with one might expect for humans with a clear understanding of the semantics of the word "lawyer" — namely, lawyers are judged (fairly or unfairly) to be most like slimeballs, bastards, dogs and sharks, and least like telephones, kangaroos, and birds. PMI-IR, on the other hand, judges lawyers to be most like computers, cats, and telephones and least like slimeballs, bastards, kangaroos and robins. Lawyers as sharks or fish are judged to be equally bad. A comparison of human vs. PMI-IR results can be seen in Figure 1. In short, it is amply clear that even for this straightforward question about lawyers, the human semantics of "lawyer" does not even vaguely resemble the semantics extracted by PMI-IR.

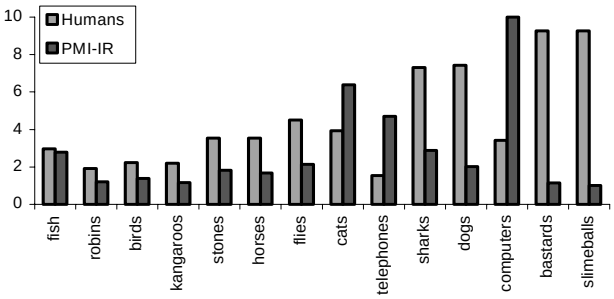


Figure 1. A comparison of PMI-IR and human data. The two profiles are clearly very different.

We also found that PMI-IR gave an extremely high rating to "Lawyers as children," higher, in fact, than any of the choices tested in Figure 1. Clearly, something is wrong here: first, lawyers cannot even *be* children (something which PMI-IR has no way of knowing) and, even metaphorically, it just doesn't seem right to us.

Rating the plausibility of Jewish/Palestinian ministers' names

Next we used PMI-IR to judge how good various first names would be for an Israeli or a Palestinian minister. We chose ten traditional Jewish names (Uri, Ariel, Moshe, Yitzhak, Yehudi, David, Samuel, Benjamin, Shimon, and Zeev) and nine traditional Arab names (Saddam, Usama, Ahmed, Mohammed, Salah, Amin, Khalil, Ashrawi, and Yasser). We asked two separate questions, each processed independently by the program. The first was "How good is X [one of the names, e.g., *Ahmed*) as the name of an Israeli minister?" All nineteen names were rated for this question. Then a second question was asked: "How good is X [again, one of the 19 names] as the name of a Palestinian minister?" All 19 names were rated for this second question. We then compared the ratings for each name for the two questions to determine their degree of correlation.

Once again, PMI-IR fails rather spectacularly: for example, it considers *Yasser* to be almost as good a first name for an Israeli minister as for a Palestinian minister! Similarly, *Ariel* is judged to be the best name, out of all ten Jewish names and all nine Arab names, for either an Israeli minister or a Palestinian minister . The results for the other names are shown in Figure 2.

Why does the program rate *Yasser* as a highly probable name for an Israeli minister and *Ariel* as highly probable for a Palestinian minister? The reason is simple: Because the program is concerned *only* with the co-occurrence of words, in this case the words *Yasser*, *Ariel*, *Israeli*, *Palestinian* and *minister*. The fact that Israel and Palestine are currently waging an undeclared war is known to PMI-IR only through higher than normal co-occurrences of war-related words and words like *Israel, Palestine*, *intifada*, etc. It knows nothing *about* wars, about their causes and effects, about their effects on societies and individuals in those societies, about hatred, about destruction, about refugees, about Israel, about Palestine, etc. *ad infinitum.* It knows only that sometimes these words co-occur with higher

frequency than others. The complete absence in PMI-IR of this deep relational structure in which the words that it encounters (and concepts these words represent) are embedded is precisely why PMI-IR fails to convincingly answer even the simplest questions that require deeper relational structure and knowledge to be answered plausibly.

So, to return to our example, in the context of the current crisis in the Middle East and of cultural specificities of first-names, good names for Palestinian ministers should be perceived as bad names for Israeli ministers and vice-versa. PMI-IR is, as we have said, unaware of the cultural context surrounding these questions. Specifically, PMI-IR is ignorant of the obvious (to us) cultural fact that some first names are typically Jewish while others are typically Arab and the relation of that cultural fact to the currently perceived inappropriateness of Palestinian ministers with Jewish names and vice-versa. So, according to PMI-IR, the appropriateness of a name for a Palestinian minister correlates almost perfectly (+0.98) with the appropriateness of the same name for an Israeli minister (See Figure 2).

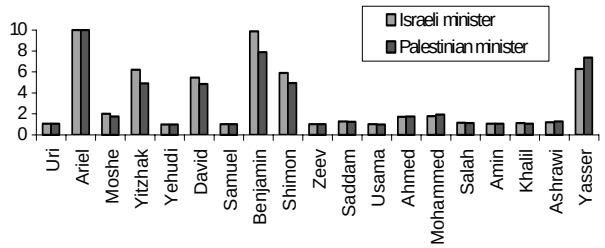


Figure 2. For two separate questions: "How good is X as the name of an Israeli minister?" and "How good is X as the name of a Palestinian minister?" PMI-IR produces an almost a perfect correlation between the appropriateness of a given name as either that of an Israeli or a Palestinian minister.

Rating names of mothers and fathers

Finally, we decided to pick an example, simple in the extreme and far removed from politics and current events, that relied on a very specific piece of contextual information that would be available to all humans but not to a word co-occurrence analysis. We compared PMI-IR's answers to the following two questions: How good is X [a first name] as the name of a father?" and "How good is X [the same first name as in the first question] as the name of a mother?" For each question we asked PMI-IR to rate ten very common men's names (John, William, Stuart, Peter, Robert, Jack, Gary, Steve, Albert, and Michael) and ten very common women's names (Barbara, Mary, Patricia, Linda, Susan, Jennifer, Karen, Nancy, Elizabeth, and Dorothy).

When judging the appropriateness of a particular name as the name of a father (or mother), humans partly rely on a simple fact that the program does not have — namely, that fathers are invariably men, while mothers are invariably women. Consequently, humans will *necessarily* rate women's names lower than men's names for the question: "How good is X as the name of a father?" Not so PMI-IR. The program concludes that "John" is the best name out of all twenty names for a father *and for a mother*. It rates "Mary" as being a very good name for a father or for a mother. Ditto for the name "William." As in the above example, the appropriateness of a particular name for a father correlates essentially perfectly (+0.99) with the appropriateness of that same name for a mother! (Figure 3)

Once again, the program fails because extracting co-occurrences of words in a large corpus of text is simply not good enough to answer questions that require abstract contextual knowledge or experience. Again, the problem is that PMI-IR has neither abstract rules nor world experience that it can rely on. And since, in any text where the word "father" occurs, the word "mother" will generally not be far away, PMI-IR fails completely on this simple rating task.

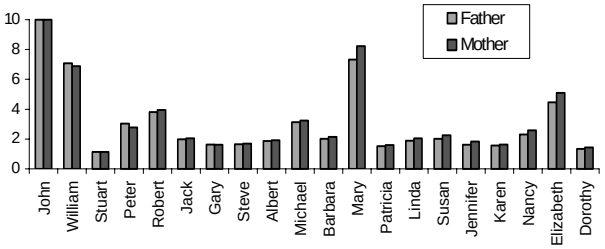


Figure 3. Two questions were asked: "How good is X as the name of a father?" and "How good is X as the name of a mother?" Lacking all context about what "fathers" and "mothers" actually are, PMI-IR produces an almost a perfect correlation between the appropriateness of the names, male or female, for a father or for a mother!

Why PMI-IR works so well on the synonym selection task

Given PMI-IR's poor performance on the simple examples above, how could it be so good in the synonym selection task, a task that would seem to require a relatively sophisticated semantic understanding of words in order to be done successfully? In what follows we will briefly examine why this program, and other similar programs, most notably LSA, are able to perform so well on this task, in spite of their inability to do the examples above.

The author of PMI-IR claims that his program can do better on the TOEFL and ESL synonym tests than any other current computer program (Turney, 2001a,b). This is believable and reasonable. Turney illustrates PMI-IR's performance on the synonym-finding task with the word *levy* (as in "to levy taxes"). Four choices are proposed — *imposed, believed, requested, correlated* — and the program chooses one of them as the best synonym based on how often that word is close to "levy" in many Web pages. The reason for PMI-IR's success does, indeed, reflect the semantics of the word under consideration, but is tied most directly to the stylistic reasons for which we use synonyms — viz., so

as not to repeat the same word too often in a given text or, especially, in the same paragraph. This purely *stylistic* constraint imposes the proximity of synonyms, which is detected by PMI-IR.

Assume you are writing an article to be put on a Web page about some *blunder* that occurred. In describing this blunder, you are aware that it is bad style to repeat the word *blunder* over and over again in your text, so you resort to synonyms, such as *failure*, *mishap*, *mistake*, *slip*, *bungle*, *mess*, and so on. This obviously produces co-occurrences of *blunder* and *mistake*, of *blunder* and *slip*, etc., and this is precisely what PMI-IR detects. A *blunder* IS (to a first approximation) a *mistake*, which IS a *slip*, etc. These words all have approximately the same dictionary definitions. In other words, the features that describe them are largely identical. This is what we called above *attributional similarity*. The point is that we can expect attributionally similar words, if only for stylistic reasons, to occur close to one another in a text. Hence, PMI-IR's excellent performance on this task.

This technique can, indeed, incorporate proximal context, as in the example of the word *tap* in the context of "maple syrup." But most analogical association involves abstract context derived from examples that, if they exist at all in the text corpus, may well be separated by millions of pages from the word under consideration. It is an open question in the field of computational analogy-making as to how this abstract relational structure might be stored and indexed fluidly enough to be accessible for later retrieval in a wide variety of contexts (see Chalmers, French, & Hofstadter, 1992, for a detailed discussion of this issue), but one thing is clear: it is *not* accessible to programs that rely only on local word co-occurrence to produce their semantics.

And, to be fair, this is one way in which humans learn attributionally similar words/concepts. But there is much more to "semantic similarity" than surface similarity.

To reiterate, *relationally (or metaphorically) similar* words require a great deal more than the detection of attributional similarity and physically proximal context. Consider rating a *banana split* as *medicine* (French, 1988, 1990). The number of times that these two items will occur together in any text anywhere is now, and will forever be, infinitesimally small compared to the other associations involving banana splits or medicine. For programs that extract semantics only from text corpora this poses a serious problem, referred to as the problem of data sparseness (Dagan et al., 1994). But the problem is unavoidable. *Of course* the number of Web pages containing the terms "banana split" and "medicine" will be vanishing small because is it not a common association at all, but it remains a perfectly valid, readily understandable one that we can judge without difficulty because we understand it *in relation to our experience with the world*, i.e., to facts like the doctor bringing us a bowl of ice-cream after we have had our tonsils out, with our mother taking us for a sundae to pick up our spirits when our junior high school safety poster was eliminated from the city competition, etc.

In other words, describing one word in terms of another usually involves much more than the above kind of "blunder-mistake-mishap-slip" synonym searching. It involves mentally placing the both words in a variety of *relational* as well as attributional contexts (that can shift fluidly) and converging on a context that fits both words (for detailed discussions of this see: Chalmers, French, & Hofstadter, 1992; Hofstadter, 1995; etc.) If both words fit that context very well, then we give the association a high rating. The more difficult it is to converge on an appropriate context for both words, the lower the rating.

PMI-IR, however, is incapable of extracting these all-important relational and contextual characteristics of situations. Specifically, for questions of the form, "Rate X as a Y," the program is incapable of grasping the relational structure in which each of the words is embedded and then of mapping those two structures onto one another in order to determine the relational similarity of the words.

Conclusions

While we acknowledge the impressive performance on certain lexical tasks of programs that employ co-occurrence analyses on large text corpora, our contention is that these programs lack the capabilities necessary to acquire real (i.e., human) semantics. This paper must not be read as a criticism of these methods per se, but rather as an incentive for researchers to develop new techniques that can incorporate more of the mechanisms by which we humans acquire semantics. These requirements go well beyond the often-cited problems of the lack of syntactic knowledge (Perfetti, 1998) and conceptual disambiguation (Landauer & Dumas, 1997). We have pointed to four problem areas for these programs, areas in which we believe future research should be focused. These areas are i) the ability to cope with the context-dependent deformability of semantic space, ii) the detection of co-occurrences of abstract structures, especially similar, but distal, abstract structures, iii) the means of providing the programs with essential world knowledge, and iv) the elimination of the assumption of words as "atomic" entities. In other words, we maintain that to know a word in a manner even approximately equivalent to how we humans know it, requires far more than merely knowing the "company it keeps."

In short, while the area of text analysis of large corpora is a fascinating and promising one, we believe that in order for real, human semantics to emerge from these techniques, the problems raised in this paper will have to be squarely confronted and overcome.

Acknowledgments

This work was supported in part by grant HPRN-CT-1999-00065 from the European Commission. Christophe Labiouse is supported by a Belgian NFSR Research Fellowship. The authors would also like to thank Jim Friedrich at Willamette University, Oregon,

for his help in conducting the informal survey cited in the text.

References

Chalmers, D. J., French, R. M. and Hofstadter, D. R. (1992). High-level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *J. of Experimental and Theoretical and Artificial Intelligence*, *4*(3), 185-211.

Church, K.W., and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association of Computational Linguistic*s, pp. 76-83.

Church, K.W., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Uri Zernik (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexico*n. New Jersey: Lawrence Erlbaum, pp. 115-164.

Dagan, I., Pereira, F. & Lee, L. (1994). Similarity-based estimation of word co-occurrence probabilities. *Proceedings of the 32nd Annual Meeting of the Assoc. for Computation Linguistics,* 272-278.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysi*s, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-195*9, London: Longman (1968).

Fletcher, C. & Linzie, B. (1998). Motive and Opportunity: Some Comments on LSA, HAL, KDC, and Principal Components. *Discourse Processes, 25*(2&3), 355-361.

French, R.M. (1988). Subcognitive Probing: Hard Questions for the Turing Test. *Proceedings of the Tenth Annual Cognitive Science Society Conference*, Hillsdale, NJ: LEA. 361-367.

French, R.M. (1990). Subcognition and the Limits of the Turing Test. *Mind*, *99*(393), 53-65.

French, R. M. and Labiouse, C. (2001). Why co-occurrence information alone is not sufficient to answer subcognitive questions. *J. of Experimental and Theoretical Artificial Intelligence.*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155-170.

Hofstadter, D. R. and the Fluid Analogies Research Group (1995). *Fluid Concepts and Creative Analogies*, New York, NY: Basic Books.

Landauer, T. & Dumais, S. (1997). A solution to Plato's Problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments and Computers, 2,* 203-208.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. Discourse Processes, 25, 363-377.

Rips, L. J., Shoben, E. J. & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1-20.

Turney, P.D. (2001a). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001*), in press.

Turney, P. D. (2001b). Answering subcognitive Turing Test questions: A reply to French. *J. of Experimental and Theoretical Artificial Intelligence.*

Yarowsky, D. (1992). Word sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 454-46.

The Importance of Starting Blurry: Simulating Improved Basic-Level Category Learning in Infants Due to Weak Visual Acuity

Robert M. French, Martial Mermillod,
Quantitative Psychology and Cognitive Science
Psychology, U. of Liège, Belgium
{rfrench, mmermillod}@ulg.ac.be

Paul C. Quinn
Psychology, Washington & Jefferson College
Washington, PA, USA
pquinn@washjeff.edu

Alan Chauvin,
Psychology, University of Grenoble, France
chauvin@lis-inpg.fr

Denis Mareschal
Psychology, Birkbeck College,
London, UK
d.mareschal@bbk.ac.uk

Abstract

At the earliest ages of development, perceptual maturation is generally considered as a functional constraint to recognize or categorize the stimuli of the environment. However, using a computer simulation of retinal development using Gabor wavelets to simulate the output of the V1 complex cells (Jones & Palmer, 1987), we showed that reducing the range of the spatial frequencies from the retinal map to V1 decreases the variance distribution within a category. The consequence of this is to decrease the difference between two exemplars of the *same* category, but to increase the difference between exemplars from two *different* categories. These results show that reduced perceptual acuity produces an advantage for differentiating basic-level categories. Finally, we show that the present simulations using Gabor-filtered input instead of feature-based input coding provide a pattern of statistical data convergent with previously published results in infant categorization (e.g., Mareschal & French, 1997; Mareschal et al, 2000; French et al, 2001).

Background

This paper builds on earlier work by Quinn, Eimas, and Rosenkrantz (1993), Mareschal and French (1997), Mareschal, French, and Quinn (2000) and French, Mermillod, Quinn, and Mareschal (2001). Quinn et al. (1993) reported the following categorization asymmetry. Infants familiarized with a number of exemplars of cats show significantly increased interest when subsequently tested on an exemplar of a novel dog compared to a novel cat. However, if the infants are first familiarized with images of dogs and then tested on a novel dog and a novel cat, there is no significant difference in interest between the two test stimuli. Mareschal and French (1997) and Mareschal et al. (2000) attributed this to the greater variance of the "dog" stimuli set compared to the "cat" stimuli set, the interpretation being that the Dog category largely subsumed the Cat category. Thus, an infant familiarized on the less variable category, Cat, would, in general, view an exemplar of a dog as a novel stimulus, whereas an infant familiarized on the more variable category, Dog, would tend to perceive a cat exemplar as simply belonging to the already-familiar Dog category. This, we claimed, explained the asymmetric levels of attention that Quinn et al. (1993) had observed. To further test this hypothesis, French et al. (2001) artificially reversed the inclusion relationship by carefully selecting breeds of dogs that were relatively similar (i.e., low variance) and highly variable breeds of cats. The connectionist computer model predicted, and the experimental results with infants subsequently confirmed (French et al., 2001), a reversal in the categorization asymmetry observed by Quinn et al. (1993).

However, one outstanding question remained. Even though, intuitively, the variability of the Cat category appears to be less than that of the Dog category, how could one be *sure* of this in any quantifiable way? Mareschal and French (1997) and Mareschal et al. (2000) handled this as follows. They originally selected ten features common to both cats and dogs (head length, head width, eye separation, ear separation, ear length, nose length, nose width, leg length, vertical extent, and horizontal extent) and measured the values associated with these features for each of the photos of the 18 Cat exemplars and 18 Dog exemplars used in their experiments. Even though this choice of features was based on experimental data where infants typically look at the head and face region of the stimulus when they observe an animal (Quinn & Eimas, 1996; Spencer, Quinn, Johnson, & Karmiloff-Smith, 1997), how could we be sure that the set of perceptual features that we had chosen corresponded to those features to which the infants were actually attending? Further, the claim was that at 3-4 months of age infants were not making use of previously acquired perceptual information (i.e., prior categorical knowledge of dogs or cats or, for that matter, ears, noses, legs, etc.); rather, they were simply relying on statistical pattern recognition. Under these circumstances, using a set of measurements of specific high-level perceptual

features to characterize the input seemed, if not

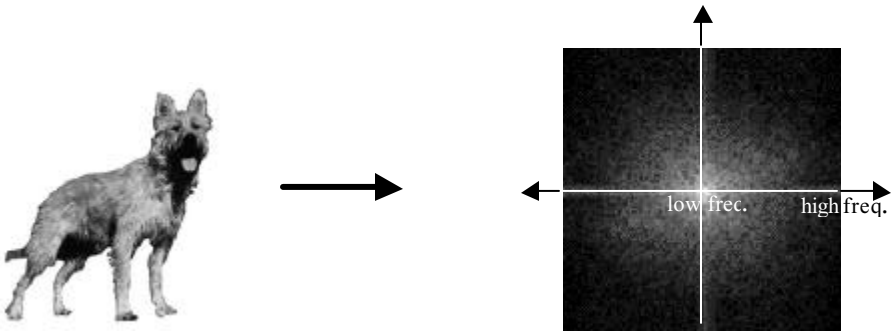


Figure 1: Transformation of the original image into a spatial-frequency map

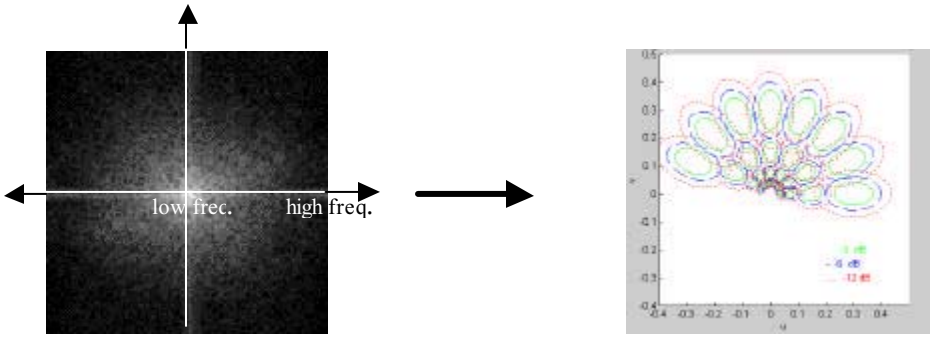


Figure 2. Once we have the map of spatial frequencies, we "cover" this map with spatial-frequency ovals along various orientations of the image. (Each of the ovals are normalized to have approximately the same energy.)

necessarily incorrect, at least somewhat inappropriate.

We therefore decided to attempt to examine this problem in a more neurobiologically plausible manner, one which sidestepped the difficulties inherent in selecting and measuring various perceptual features of the cat and dog stimuli. The dog/cat stimuli used in the simulations reported in this paper were those used in French et al. (2001), all of which had been normalized to have approximately the same size.

Organization of the present paper

We will attempt to answer two questions in the present paper.

The first is: Can we avoid the use of explicit feature coding in our autoencoder model of infant categorization and replace this coding with Gabor-filtered input known to have a neurophysiological counterpart in the infant visual system? We will show that this can, indeed, be done successfully.

The second issue that we will address starts from the well-known fact that the 3-4 month old infant visual system is not sensitive to high spatial frequency information (Banks & Salapatek, 1981; Dobson & Teller, 1978). However, instead of this being a disadvantage for the infant, we will show that, somewhat counter-intuitively, this low visual acuity is actually an *advantage* in learning basic-level categories. The claim is that high spatial frequency information in the input signal produces an "information overload" in the infant cognitive system, adding information that is not necessary for correct categorization but that must, nonetheless, still be processed. In other words, when the infant is attempting to learn basic-level categories, high spatial frequency information in the input is very much like noise (Turkewitz & Kenny, 1982; Turkewitz & Kenny, 1985) and, as such, the less there is, the better.

Spatial frequency maps

It is well known that different columns in V1 are sensitive to different ranges of spatial frequencies (De Valois & De Valois, 1988; Tootell, Silverman, & De Valois, 1981). A scene reconstructed from only low spatial frequency information (i.e., with fine details blurred out) appears to us to be blurry. On the other hand, an image composed of high spatial frequencies would show *only* the fine details and would have no global perspective (rather like seeing many individual trees, but having no sense of the global entity, a forest). In any case, in order to have an optimal perception of a scene, we need the entire range of spatial frequencies. Therefore, by means of a 2D

Fourier transform, we first decomposed each of the images in the stimulus set into its component spatial frequencies and plotted them on a spatial frequency map (see Figure 1).

We then covered the frequency diagram with a "flower-petal" arrangement of 26 oval spatial frequency areas ("filters") corresponding to various orientations emanating from the center of the spatial-frequency diagram (Figure 2). Gabor functions were used to simulate the 2D spatial and spectral structure of simple cells in visual primary cortex. (Jones & Palmer, 1987; Jones, Stepnowski, & Palmer, 1987). The smaller petals near the center of the map encompass the low frequencies, while the larger ovals further from the center group together high spatial frequencies. For each of these 26 filters, we calculate an "energy" value based on the local energy spectra, thereby simulating the activity of V1 complex cells (Sakaï & Tanaka, 1999). This value determines the importance of that particular filter. If there are many spatial-frequency points that fall in a particular oval, it is given a high energy value; few points in a particular oval mean a low energy value.

Recall that in prior experiments and simulations (Mareschal et al., 2000; French et al., 2001), the dog/cat stimuli were characterized by a vector of ten values, with each value corresponding to a particular "high-level" feature. Now, instead of using ten features, we characterize each of the images by a vector of 26 values, each of which corresponds to the weighting of a group of spatial frequencies along various orientations of the image.

Visual acuity in infants

We know that the visual acuity in infants is not the same as that of adults (Banks & Salapatek, 1981; Dobson & Teller, 1978). In particular, infants do not perceive high-spatial frequencies (i.e., fine details), or perceive them only poorly. Certain authors (Turkewitz & Kenny, 1982; Turkewitz & Kenny, 1985) have claimed that, rather than being a problem, this reduced visual acuity may actually improve perceptual efficiency by eliminating the "information overload" caused by too many extraneous fine details likely to overwhelm their cognitive system. An implication is that basic-level category learning may be facilitated by reduced visual acuity.

In both of the simulations below we removed most of the high spatial frequencies from the input given to the autoencoder network that was used in Mareschal and French (1997), Mareschal et al. (2000), and French et al. (2001). This was done by weighting the contribution of each of the spatial frequencies according to a normal distribution (with the low spatial frequencies at the center) and cutting off all spatial frequencies above 7.1 cycles/degree. The spatial frequencies are Gaussian-filtered in such a way that spatial frequencies above 3-4 cycles/degree contribute very little to the input vector associated with each image; the cut-off of 7.1 cycles/degree completely removes the highest spatial frequencies.

Overview of the simulations

The 26-16-26 autoencoder network used in the two simulations presented in this paper is based on a simple encode-compare-adjust principle (Sokolov, 1963; Charlesworth, 1969; Cohen, 1973): When an infant sees a perceptual stimulus, this stimulus is encoded as an internal representation, which is continually compared to the external stimulus and adjusted to match it. As long as there is a significantly large discrepancy between the internal representation and the external stimulus, the infant continues to look at the external stimulus. As this discrepancy falls, the infant becomes less interested in the external stimulus. In the autoencoder model, this is equivalent to the network's correctly generalizing on output to match the network input (i.e., if the error on each of the 26 outputs is less than 0.5). In particular, we will use this criterion of generalization to measure the network performance on the category-learning task in Simulation 2.

In the simulations reported here we hope to establish two claims – namely:

i) Simulation 1: that the use of a vector of 26 weighted spatial-frequency values, as described above, does, indeed, produce autencoder network results that are similar to those produced by infants tested on the same images and

ii) Simulation 2: that the reduced visual acuity produced by largely eliminating high-spatial frequency information from the input (i.e., "blurry" vision) actually significantly *improves* the network's ability to categorize the images presented to it.

Simulation 1: The adequacy of Gabor-filtered spatial-frequency input

In the first simulation we used the dog/cat stimulus set used in French et al. (2001). These authors used an encoding technique developed in Mareschal and French (1997) and Mareschal et al. (2000) in which 10 features of the animal images were measured and used as input to a 10-8-10 autoassociative network. Using feature-based input to this autencoder, we obtained categorization results that qualitatively matched experimental data with infants. In contrast, in the present simulation, we decomposed each image into a vector of values consisting of the energy values from each Gabor filter for a given orientation and spatial frequency. These values correspond, at least approximately, to what V1 neurons are known to "perceive."

Each value of the 26-element vector represents an "energy" level associated with that particular spatial frequency. For this simulation, frequencies above 3-4 cycles per degree of visual arc are given a very low

energy value (very high frequencies, i.e., those above 7.1 cycles/degree are simply removed), which means that they contribute very little to the overall input vector (i.e., they contribute very little to the overall characterization of the image). The removal of this high spatial-frequency information was done to simulate the reduced visual acuity of 3-4 month old infants (Courage & Adams, 1995).

The simulation reported here was done on a standard 26-16-26 feedforward backpropagation autoencoder network (learning rate: 0.1, momentum: 0.9, Fahlman offset: 0.1). The stimulus set and the training regime was identical to that used in French et al. (2001). (It is important to recall that in French et al., 2001, the Dog stimuli were selected to be the less varied category, while the exemplars making up the Cat category were chosen to be considerably more varied than the dogs.)

Networks were trained in batches of 2 patterns for a maximum of 250 epochs. This simulated familiarization with pairs of pictures for a fixed period before being presented with a new familiarization pair. All results were averaged over 100 runs.

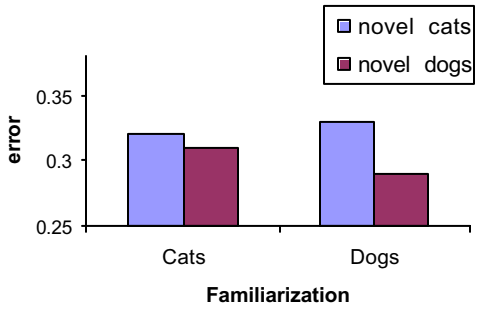


Figure 3a: Network generalization errors on novel cats/dog exemplars as a function of familiarization category.

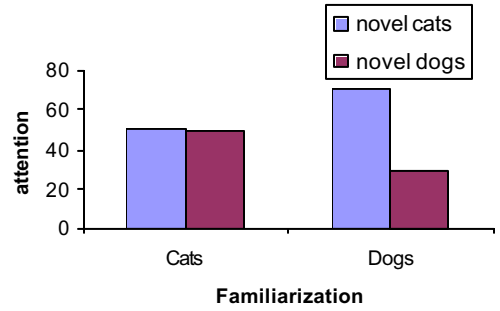


Figure 3b: Corresponding results for 3-4 month old infants

Figure 3a shows the model's generalization error to novel exemplars of cats and dogs as a function of whether they were trained on cats (the broad category) or on dogs (the narrow category). Networks trained (i.e., familiarized) with cats show very little difference in error (hence predict little difference in infant looking times) when tested with a novel cat or a dog. In contrast, networks originally trained with dogs show significantly more error ($F(1, 198)=13.4$ $p<0.0005$) when tested with a novel cat than when tested with a novel dog (suggesting a preference for looking at a novel cat vs. a novel dog). Figure 3b shows the corresponding attentional asymmetry in 3-4 month old infants, as reported in French et al. (2001).

These simulation results using Gabor-filtered spatial frequency data allow us to conclude that the use of this type of spatial frequency data produces a reasonable fit to data. Most importantly, this result allows us to circumvent the thorny issue of using a particular set of "high-level" feature measurements (ear length, eye separation, etc.) to characterize the images used in the simulations.

Simulation 2. Improved categorization with reduced visual acuity

Does the autoencoder model of infant categorization (Mareschal & French, 1997; Mareschal et al., 2000) show improved categorization performance (at least on the dog/cat basic-level category images used in French et al., 2001) when "reduced acuity" input is used compared to "full acuity" input? The answer is that categorization performance is, indeed, enhanced, as we will show below.

To reiterate, the key idea of this simulation, which at first blush seems rather counter-intuitive: categorization performance for basic-level categories (Rosch et al., 1976) should be better *without* high spatial frequency information. This information is rather akin to noise in the input since, while it does indeed add information to the signal, it is not needed for accurate basic-level categorization. This extraneous information thus makes it more difficult (for the infant or for the network) to make use of the lower spatial frequency information that is, in fact, essential to basic-level categorization.

We used the same network as in Simulation 1, with an identical parameter set. We first ran the network (100 independent runs) with input data that contained all of the spatial-frequency information in the images. We then ran the network again (100 independent runs) with input data from which most of the high-spatial frequency information had been removed, as described above. The network was trained for 250 epochs on the training stimuli, as in Simulation 1.

As can be seen in Figure 4, whether the network was trained on Cats or Dogs, whether it was tested on novel dogs or novel cats, its categorization performance is significantly poorer when the input signal also contains high spatial frequency information compared to input with the high spatial frequencies removed.

It is also important to note that in the reduced visual acuity condition, we continue to see a significant difference in error (corresponding to

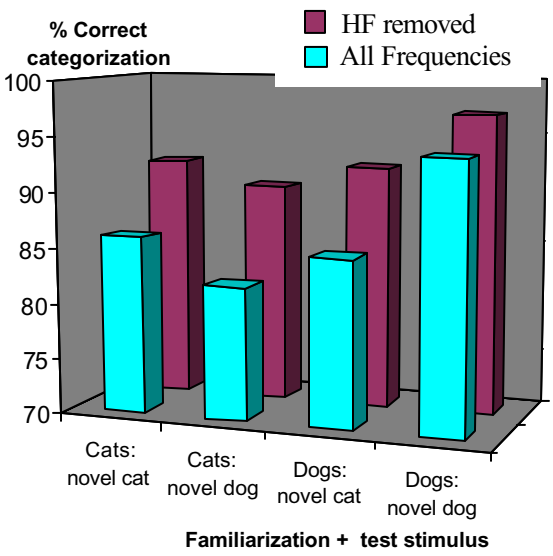


Figure 4. More information is not always better information, at least for basic-level categorization. The addition of high spatial frequency information makes correct basic-level categorization *more* difficult for the network.

attention in infants) when the network is trained first on dogs (in these experiments, the narrow category) and then sees a new cat, compared to when the network is first trained on cats (the broad category) and then sees a new dog.

Basic-level categories and incrementally increasing cognitive load

It is important to note that reduced acuity should improve categorization learning in the case of basic-level categories, but not subordinate-level categories. To see why this would be true we need to refer to Rosch et al.'s (1976) definition of "basic-level" categories. This level of categorization is the level for which the ratio of between-category variance to within-category variance is the highest. In other words, between-category variance is high with respect to within-category variance, which is generally relatively low. Within-category variance increases as fine-grained details of category exemplars increase. But these finer details are revealed only by the high spatial frequencies. For this reason, a decreased visual acuity that consists of partially or completely removing high-spatial frequency information, will decrease within-category variance and leave between-category variance largely unchanged. This would improve the learning of basic level categories, but would make it difficult, if not impossible, for 3-4 month old infants to learn categories that depend on high spatial frequency information. This applies, in particular, to subordinate-level categories.

Having *already learned* a certain number of basic-level categories under conditions of reduced visual acuity, when the high spatial frequency apparatus does begin to come on-line at around 7 to 8 months of age (Kellman & Arterberry, 1998), the infants will be in a better position to then do more refined (i.e., subordinate-level) category learning. Thus, rather than having to confront all of the information associated with a particular category at once, the limitations of visual acuity of the infants' immature visual system first helps the infant to distinguish broader categories. Once these have been learned (or partially learned), then their visual/cognitive apparatus is then ready to build on this knowledge by incorporating the fine-grained details, perceived through high spatial frequency perception, that characterize subordinate expert-level categorization. The overall results of the simulations are thus consistent with a differentiation-driven view of early category development (Quinn & Johnson, 1997, 2000).

Furthermore, these results are entirely consistent with Archambault, Gosselin, & Schyns (2000), who showed that basic-level categorization seems to be more resistant to changes in viewing distances than that of subordinate-level categorization. This is because of the fact that as an object recedes from the viewer, information about details (i.e., high spatial frequency information) is lost, whereas low-spatial frequency information is not. Since basic-level categorization is largely based on the latter, we would expect more resistance to change of this type of categorization compared to subordinate-level categorization, where features are, indeed, essential.

A Prediction of the Model

A simple prediction emerges from these results. By manipulating the amount of high-frequency information in test images, it should be possible to vary infants' responses to these items after familiarization on a standard set of basic-level categories. So, for example, consider the Dog/Cat stimuli from the experiment by Quinn et al. (1993), in which the Dog category largely subsumes the Cat category. Under normal circumstances when infants are familiarized with cats, then shown a novel dog and a novel cat, they devote significantly more attention to the novel dog than to the cat. But were we to choose a novel dog and a novel cat whose differences were based largely on high spatial frequency information, we would expect the previously observed novelty preference to disappear, even if for us, adults, the two animals were quite different, one clearly being a dog, the other, clearly a cat.

Conclusion

In an extension of work done by Mareschal & French (1997), Mareschal et al. (2000) and French et al. (2001), we have been able to show that there is no need to use feature-based characterizations of the stimuli presented to the encoder network. Autoencoder results using Gabor-filtered input

corresponding approximately to the set of frequencies that the human visual system is known to use also produce a good approximation to categorization results in infants. We have also modeled a rather counter-intuitive learning advantage for basic-level categories that arises from reduced acuity input. Finally, based on the results of our autoencoder model of infant categorization and on the results we obtained using reduced acuity input, we have suggested experiments that might be performed on infants to further examine the validity of this model.

Acknowledgments

This work was supported in part by Grant HPRN-CT-2000-00065 from the European Commission to R. French, ESRC (UK) Grant R000239112 to D. Mareschal, and by Grant BCS-0093600 from the National Science Foundation to P. C. Quinn.

References

Archambault, A., Gosselin, F., & Schyns, P. (2000). A natural bias for the basic level? *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, NJ:LEA, 585-590.

Banks, M.S., & Salapatek, P. (1981). Infant pattern vision: A new approach based on the contrast sensitivity function. *Journal of Experimental Child Psychology, 31,* 1-45.

Charlesworth, W. R. (1969). The role of surprise in cognitive development. In D. Elkind & J. Flavell (Eds.), Studies in cognitive development. Essays in honor of Jean Paiget, pp. 257-314, Oxford, UK: Oxford University Press.

Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19,* 157-180.

Courage M.L., Adams R.J. (1995). Infant peripheral vision: the development of monocular visual acuity in the first 3 months of postnatal life. *Vision research, 36,* 1207-1215.

De Valois, R.L., De Valois K.K. (1988). *Spatial Vision*. Oxford University Press. New York.

Dobson, V., & Teller, D. Y. (1978). Visual acuity in human infants: A review and comparison of behavioral and electrophysiological studies. *Vision Research, 18,* 1469-1483.

French R. M., Mermillod M., Quinn P. C. & Mareschal D. (2001). Reversing category exclusivities in infant perceptual categorization: simulations and data. *Proceedings of the 23th Annual Cognitive Science Society Conference,* LEA, 307-312.

Jones, J.P. & Palmer L.A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J Neurophysiol. 58(6),* 1187-211.

Jones, J.P., Stepnoski A. & Palmer L.A. (1987). The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *J Neurophysiol.* 58(6): p. 1212-32.

Kellman, P. J., & Arterberry, M. E. (1998). *The cradle of knowledge: Development of perception in infancy*. Cambridge, MA: MIT Press.

Mareschal, D., & French, R. M. (1997). A connectionist account of interference effects in early infant memory and categorization. In *Proceedings of the nineteenth annual conference of the Cognitive Science Society* (pp. 484-489). London: Erlbaum.

Mareschal, D., French, R., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology, 36,* 635-645.

Quinn, P. C. & Johnson (2000). Global before basic object categorization in connectionist networks and 2 month-old infants. *Infancy, 1(1),* 31-46.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of Experimental Child Psychology, 63,* 189-211.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. *Journal of Experimental Child Psychology, 66,* 236-263.

Quinn, P., Eimas, P., & Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3- and 4-month-old infants. *Perception, 22,* 463-475.

Rosch E., Mervis C.B., Gray D.G., Johnson D.M. & Boyes-Braem P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Sakaï K. & Tanaka S. (1999). Spatial pooling in the second-order spatial structure of cortical complex cells. *Vision Research*, 40, 855-871.

Solokov, E. N. (1963). Perception and the conditioned reflex. Hillsdale, NJ: LEA.

Spencer, J., Quinn, P. C., Johnson, M. H., & Karmiloff-Smith, A. (1997). Heads you win, tails you lose: Evidence for young infants categorizing mammals by head and facial attributes (Special Issue: Perceptual Development). *Early Development and Parenting, 6*, 113-126.

Tootell, R. B., Silverman, M. S., De Valois, R. L. (1981). Spatial frequency columns in primary visual cortex, *Science*, 214, 813-815.

Turkewitz G., Kenny P. A. (1982). Limitations on input as a basis for neural organization and perceptual development: a preliminary theoretical statement *Developmental Psychobiology, 15(4),* 357-368.

Turkewitz G., Kenny P. A. (1985). The role of developmental limitations of sensory input on sensory/perceptual organization. *Journal of Developmental and Behavioral Pediatrics: JDBP, 6(5),* 302-306.

Modelling the Development of Dutch Optional Infinitives in MOSAIC

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.UK)
Julian Pine (JP@Psychology.Nottingham.Ac.UK)
Fernand Gobet (FRG@Psychology.Nottingham.Ac.UK)
School of Psychology, University Park, Nottingham
NG7 2RD United Kingdom

Abstract

This paper describes a computational model which simulates the change in the use of optional infinitives that is evident in children learning Dutch as their first language. The model, developed within the framework of MOSAIC, takes naturalistic, child directed speech as its input, and analyses the distributional regularities present in the input. It slowly learns to generate longer utterances as it sees more input. We show that the developmental characteristics of Dutch children's speech (with respect to optional infinitives) are a natural consequence of MOSAIC's learning mechanisms and the gradual increase in the length of the utterances it produces. In contrast with Nativist approaches to syntax acquisition, the present model does not assume large amounts of innate knowledge in the child, and provides a quantitative process account of the development of optional infinitives.

The Optional Infinitive Stage

One phenomenon which has received considerable attention in the area of syntax acquisition is the so-called *Optional Infinitive (OI) stage* (Wexler, 1994, 1998). Children in the OI stage of development use a high proportion of (root) infinitives, that is, verbs which are not marked for tense or agreement. In English, root forms such as *go*, or *eat* are infinitive forms, whereas *ate* or *goes* are marked for tense and agreement + tense respectively. Verbs which are marked for agreement or tense are known as *finite* verbs. (Technically, infinitives are a subclass of the class of *non-finite* verb forms, which also includes past participles and progressive particles).

Another feature of the OI stage is that children often omit subjects from their sentences. That is, children will produce utterances such as *throw ball* from which the subject (*I*) is absent. While the proportion of infinitives is (considerably) higher than for adult speech, children in the OI stage do show competence regarding other syntactic attributes of the language. Typically, children will not make errors in the basic verb-object order. English-speaking children, for instance, will say *throw ball*, but not *ball throw*. One puzzling feature of the OI stage is that children produce both inflected and uninflected forms in contexts requiring the inflected form, but do not produce finite forms in nonfinite contexts. The fact that children use both inflected and uninflected forms shows that it is not the case that they simply don't know the inflected forms.

The optional infinitive stage has been shown to occur in many different languages, which can differ considerably in their underlying syntactic properties, and children do show competence regarding these syntactic properties. Different languages also differ with respect to how pronounced the OI stage is. Since most verb forms in English are not distinguishable from non-finite forms, it is relatively difficult to distinguish optional infinitives from grammatically correct utterances. In other languages (e.g. Dutch), the number of unambiguously finite forms is larger, and as a result the optional infinitive stage is more pronounced.

Wexler (1998) has proposed a Nativist account of why children in the optional infinitive stage produce a large number of non-finite forms. In accordance with Chomsky's theory of Universal Grammar (Chomsky 1981), he theorizes that children in the optional infinitive stage actually know the full grammar of the language. The only thing they do not know is that Agreement and Tense are obligatory. This approach accounts for the fact that children produce both correct finite forms and incorrect (optional) infinitives. It also explains why children rarely produce other types of errors. Finally, its great strength is that it unifies across languages where children clearly use optional infinitives despite differences in their underlying grammar. However, there are also a number of problems with Wexler's account.

Firstly, Wexler's theory does not give a process account of developmental change in the use of optional infinitives. He assumes this to be due to *maturation*.

Secondly, the theory makes very limited quantitative predictions. It only predicts that the optional infinitive stage occurs, and that children will stop making optional infinitive errors at some point. It makes no specific predictions regarding the time course of this development, or related changes in other attributes.

Thirdly, the theory assumes a large amount of innate knowledge in the child (the theory assumes that the child does not know that inflection is obligatory, but otherwise knows the full grammar of the language).

An obvious alternative to Wexler's theory is that children learn the grammar of a language through exposure to that language. Wexler discounts this kind of learning-based approach on the grounds that the grammar is too difficult to learn, that the optional infinitive stage lasts too long (years), and that, although children produce both correct and incorrect forms, when they use finite forms, they use them correctly (Wexler, 1994).

In this paper, we aim to show that the dynamics of the optional infinitive phenomenon can be simulated using a simple learning mechanism which performs a distributional analysis of naturalistic input. Earlier versions of the model have already been shown to simulate the basic optional infinitive phenomenon in both English (Croker, Pine & Gobet, 2001) and Dutch (Freudenthal, Pine & Gobet, 2001). Whereas the earlier versions modelled one specific stage in development, the present model aims to simulate the *developmental change* that is apparent in the use of optional infinitives.

There are a number of reasons for choosing Dutch as the target language. Firstly, as was mentioned, in adult speakers' Dutch, unambiguous finite forms are far more frequent than they are in English. In English, in the present tense, only the third person singular can be distinguished from the infinitive form. In Dutch, the first, second and third person singular are unambiguously finite. If, for instance, an English speaking child produced *I throw ball*, it would be unclear whether the verb *throw* was an infinitive form. The Dutch equivalent *ik gooi bal* would be classified as a finite form, because *gooi* is different from the infinitive *gooien*. Thus, the number of unambiguously finite forms is larger in Dutch than in English. (This suggests that developmental change in the use of optional infinitives is likely to be more pronounced in Dutch than it is in English, which makes the simulation of Dutch child language more informative as a modelling exercise.) A second reason for using Dutch is that detailed data regarding this development are available. Wijnen, Kempen & Gillis (2001) have analysed the corpora of two Dutch speaking children and have shown that the proportion of root infinitives decreases from around 90% to roughly 10% between the ages 1;6 and 3;0. By comparison, root infinitives are used in less than 10% of adults' utterances. Wijnen et al. concluded that the frequency of occurrence of optional infinitives in the child's speech was related to frequency, and utterance position, as well as lexical transparency.

A third reason for choosing Dutch as the target language is that Dutch grammar is relatively complex when considering finiteness of verb forms. Dutch is what is known as an SOV/V2 language. This means that the verb in Dutch can take one of two positions, depending on its finiteness. A non-finite verb takes the sentence final position, whereas finite verbs take the second position. Therefore, in the sentence

Ik gooi een bal (1)
(I throw a ball)

the verb *gooi* (*throw)* is finite and takes second position. In the construction

Ik wil een bal gooien (2)
(I want a ball throw/ I want to throw a ball)

the verb *gooien* is a non-finite form, and takes sentence final position. (The auxiliary *wil* is finite and takes second position.) In English, which is an SVO language, verb position is not dependent on the finiteness of the verb. If a model is to learn from the distribution of naturalistic speech input, then the production of a large number of infinitives while respecting the overall grammar would appear to represent a greater challenge in Dutch than in English.

MOSAIC

MOSAIC (Model of Syntax Acquisition In Children) is an instance of the CHREST architecture, which in turn is a member of the EPAM (Feigenbaum & Simon, 1984) family of models. CHREST models have successfully been used to simulate novice-expert differences in chess (Gobet & Simon, 2000), as well as several phenomena in language acquisition (Jones, Gobet & Pine, 2000a, 2000b; Croker, Pine & Gobet, 2001, 2002; Freudenthal, Pine & Gobet, 2001, 2002). We will now give a brief description of MOSAIC. A more detailed description of the model can be found elsewhere in this volume (Freudenthal, Pine & Gobet 2002). The model we have used in these simulations is identical to the one that Freudenthal et al. (2002) used for the simulation of a different phenomenon (Subject Omission) in another language (English).

The basis of the model is a discrimination net, which is used to store the input that is fed to the model. The network is an n-ary tree which is headed by a root node. Utterances that the model sees are encoded by sequences of nodes in the network.

The model encodes the fact that word *a* has been followed by word *b* in the input by creating a node for word *b* under the node for word *a*. The fact that word *a* has preceded word *b* is similarly encoded. Fig. 1 may illustrate the basic MOSAIC network. Apart from the standard links between words that have followed each other in utterances previously encountered, MOSAIC also employs *generative links*. Generative links connect nodes that are distributionally similar. When two nodes (phrases) have a high likelihood of being preceded and followed by the same words in the input, a generative

link is created between them. Since distributionally similar phrases are likely to belong to the same word class, generative links that develop end up linking clusters of nodes that represent different word classes. The induction of word classes on the basis of co-occurrence statistics is the only mechanism that MOSAIC employs for representing syntactic rules. The main importance of generative links lies in the generation of utterances from the model. In generation, words that share a generative link can be substituted, thus allowing the model to generate novel utterances. Again, the reader is referred to Freudenthal, Pine & Gobet (2002) for details regarding generation. One point worth mentioning here is that the model will only output utterances that contain an end marker (i.e. where the utterance final phrase has occurred in a sentence final position in the input). Several authors have suggested that sentence final position is particularly salient, and that children are more likely to produce utterances that have occurred in sentence final position (Shady & Gerken, 1999; Naigles & Hoff-Ginsberg, 1998).

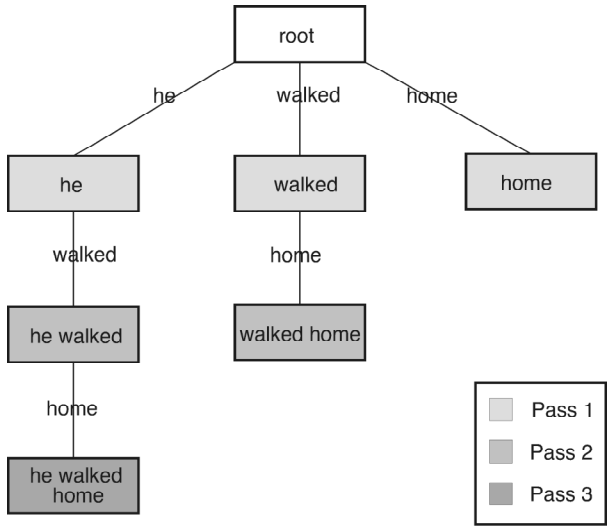


Fig. 1: MOSAIC learning an input

The model we used for these simulations is an extension of that used in Freudenthal, Pine & Gobet, (2001), which simulates the children's performance in Dutch at one specific point in time. This version of the model has also been shown to produce both root infinitives and correct inflected forms in English (Croker, Pine & Gobet, 2001). The main difference between this and the previous version of the model is that the present model learns much more slowly. By using a slow learning rate, and iteratively feeding input to the model and analysing its resulting output, we were able to model consecutive stages of development. In the previous version, a word was encoded on the first occasion it was seen, which resulted in a model with an MLU (Mean Length of (output) Utterance), that was comparable to that of a child that has passed the OI stage. In the present version, the probability of creating a node is dependent on the size of the network (a measure of the linguistic knowledge or vocabulary size of the child), and the length of the phrase that is being encoded. More specifically, the probability of creating a node is given by the following formula:

$$NCP = \left(\frac{*nodes_in_net*}{50{,}000} \right)^{length_phrase}$$

It will be apparent from the formula above that the probability of creating a node is very low if the network is small (i.e., the number of nodes in the net is low). As the number of nodes in the net grows, this probability will increase. A second point to note is the occurrence of the length of the phrase (number of words) in the exponent. This has the effect of lowering the probability of creating nodes that encode longer phrases. The value 50,000 has been chosen somewhat arbitrarily. Its main role is to ensure that the difference in node creation probability for short and long utterances decreases as a function of the size of the net. As the number of nodes in the net approaches 50,000 (a typical number for a *saturated* model given the Dutch input used here), the base number in the formula approaches one, and thus the weight of the exponent diminishes. One additional remark must be made about this formula: phrases that occurred in utterance final position (i.e., contained an *end marker*), were treated differently from other utterances in that their length (for calculation of the NCP) was decreased by 0.5. This constitutes an *end marker bias* in learning, rather than at production. It has been argued that utterance final phrases are learned more easily than non-utterance final phrases (Wijnen, Kempen & Gillis, 2001).

The Simulations

The data that were simulated were taken from Wijnen, Kempen & Gillis (2001). Wijnen et al. analysed two Dutch corpora of child and adult speech (the corpora of Matthijs and Peter and their mothers). The corpora consisted of transcribed tape recordings of speech between mother and child. For Matthijs, the recordings were made between the ages 1;9 and 2;11. For Peter they were made between 1;7 and 2;3. The children's MLU (Mean Length of Utterance) ranged from 1 to roughly 3. Wijnen et al. analysed the corpora with respect to the presence of the optional infinitive phenomena in both the mother's and the children's speech. On the basis of the children's data, four developmental stages were identified, and the proportion of finite, non-finite and discontinuous finites

(see below) was assessed. Since the corpora that Wijnen et al. analysed are available in the CHILDES data base (MacWhinney & Snow, 1990), we had access to the same corpora, and used these (maternal corpora) as input for the model.

In order to compare the output of the model to the children's speech, we ran the input through the model several times. After each run of the model, we generated output, and compared the MLU of the model with the child's MLU in the developmental stages that Wijnen et al. identified. We then selected for further analysis those output files that most closely matched the children's MLU for the four developmental stages. The actual analysis performed was similar to that of Wijnen et al. Firstly, we selected those utterances that contained one or more verb forms. We then classified these utterances as finite, non-finite or discontinuous finite. In doing so, we used the following criteria:

- An utterance is considered *non-finite* if it contains only non-finite verb forms.
- An utterance is considered *finite* if it contains only finite verb forms.
- An utterance is considered a *discontinuous finite* if it contains both a non-finite, and a finite form (e.g. a finite auxiliary).

There were some small differences from Wijnen et al.'s analysis. The most notable difference is that Wijnen et al. removed all forms resembling imperatives, starting with the early two word stage. When coding actual speech, this is relatively easy to do, since context allows one to disambiguate. Since the model's output does not provide this context, the classification remains somewhat ambiguous. We therefore decided not to remove forms resembling imperatives.

Results

Figure 1 shows the data and the simulations for Matthijs and Peter. The model shows a considerable drop (around 50%) in the proportion of non-finites for both input sets. For the children, the corresponding drop is 80-85%. Given the fact that we are using naturalistic input to model the development of children's speech, and the fact that we used an identical model for both children (i.e. no parameters were adjusted) we consider

Fig. 2a: Data for Matthijs

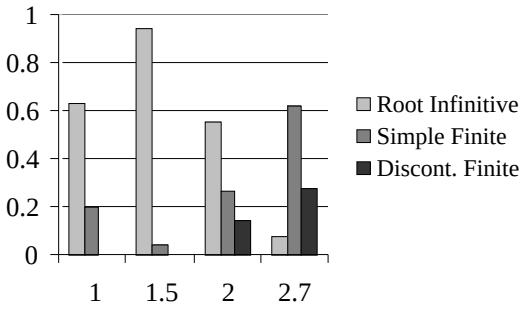


Fig. 2b: Model for Matthijs

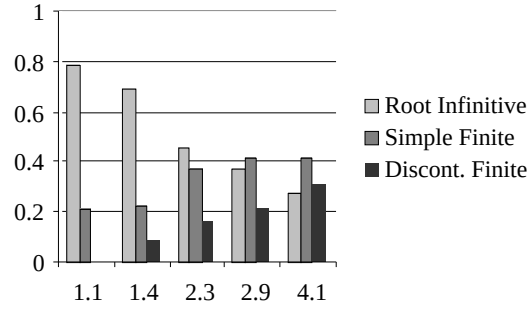


Fig. 2c: Data for Peter

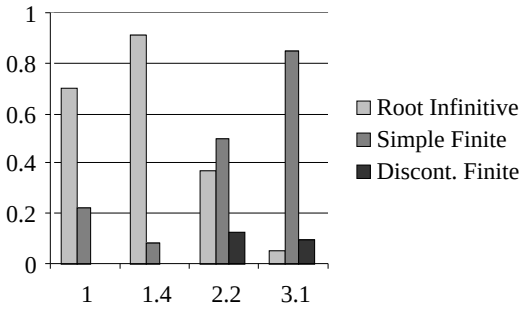


Fig. 2d: Model for Peter

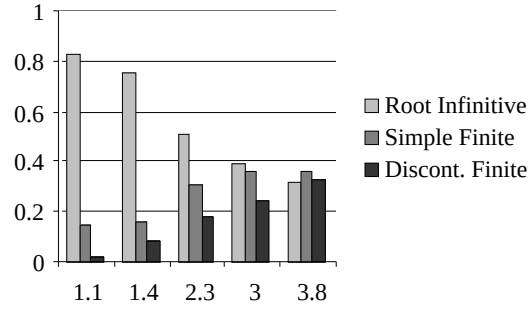


Figure 2: Distribution of root infinitives and (discontinuous) finites as a function of MLU for Matthijs, Peter, and their respective model.

this figure promising. (Note however, that we report five rather than four data points for the models. The last data point reflects an MLU larger than that for the children in the final stage, and is included to show that the proportion of non-finites continues to decrease.)

What mechanism is responsible for this drop in the model's output? The thing to note is that non-finite forms take sentence-final position in Dutch, and that the model is biased towards generating (and encoding) phrases that occurred in sentence-final position. The formula for calculating the node creation probability ensures that early on, the model will encode relatively short utterances that occurred in sentence-final position. If these utterances contain a verb, it will (in Dutch) most likely be a non-finite form. These non-finite forms may have been part of an auxiliary + verb construction (e.g. *He wants to build a house*). Since the model can generate partial utterances, it can learn the root infinitive *build a house* from this (discontinuous) finite form. Therefore, a high proportion of non-finite forms is expected in the early stages of the model's development. As the model sees more and more utterances, the number of nodes in the net will increase, and the probability of creating a node will also increase. As a result, longer and longer utterances will be encoded in the network. As the encoded utterances increase in length, they will be more likely to include words that occur early in the utterance. Since finite forms take second position in Dutch, the number of finite forms will increase as the model starts generating longer utterances. Note that this also means that root infinitives will slowly be replaced by discontinuous finites. Where the model may have output the root infinitive *build a house* early on, it will be able to output the discontinuous finite *he wants to build a house* as the size of the net increases.

Table 1: Proportion of correct Object-Verb orderings for the model as a function of finiteness (averaged over developmental phase).

	Finites	Non-Finites
Matthijs	.94	.91
Peter	.96	.93

Given that the model simulates the basic optional infinitive phenomenon, we now need to assess whether it conforms to the other criteria of the optional infinitive stage. Tables 1 and 2 show the proportion of correct verb placement and the position of the object relative to the verb. It is evident, that, in the majority of cases, the model uses the correct placement, indicating that it is sensitive to basic Dutch grammar.

The fact that the model gets the basic word order right in the majority of the cases is perhaps not very surprising. After all, the input that the model learns from has the correct word order. This is not a trivial result however, as the fact the children correctly produce the correct word order has been taken as evidence by Wexler (1994, 1998) that the child knows the actual grammar.

Table 2: Proportion of correct verb placement for the model as a function of finiteness (averaged over developmental phase).

	Finites	Non-Finites
Matthijs	.85	.95
Peter	.88	.97

Though these results are very promising, especially considering the fact that we are using naturalistic input to simulate actual children's speech, some issues require attention. For both children, the proportion of non-finites is underestimated for stage 2, and overestimated for the later stages. Possible causes for the underestimation in the early stages may lie in the fact that Wijnen et al. removed forms resembling imperatives as of stage two (which may also explain the relatively low proportion of non-finites in stage one in the data). We did not do this. This underestimation may be exacerbated by the fact that the model produces relatively few utterances early on, thus making it relatively sensitive to small changes. A second, possibly more likely cause may be that there are additional factors that cause the high proportion of non-finites in the children. Wijnen et al. claim, on the basis of a regression analysis, that frequency of occurrence alone is not enough to explain the high incidence of non-finite forms. They suggest that non-finite forms are learned more easily and attribute this to lexical transparency. Since MOSAIC does not employ any semantics, we cannot model this effect. Regarding the later stages, one possible cause for the overestimation is the fact that MOSAIC has a limited ability to unlearn. That is, at any stage, when the model generates output, it will generate all the utterances it can. Thus, once the model has learnt to generate *he wants to build a house*, it will also (still) generate *build a house*.

Mechanism for change

The model shows a drop in the proportion of non-finites of roughly 50%. We can now ask ourselves what has caused this change. Two possible explanations come to mind. Firstly, as the model learns, the MLU of the generated utterances increases. As explained earlier, if the generated utterances adhere to Dutch grammar, an increase in the proportion of finites is expected. A second possible cause lies in the proportion of generated (rather than rote learned) utterances. As the model's MLU increases, so does the proportion of generated utterances. This may result in a

disproportionate growth in the number of finite utterances. (Since finite forms are more frequent, a relatively large proportion of the generated utterances contain finite verbs.) While a regression analysis showed that the increase in MLU alone explained 90% of the variance in the proportion of finite utterances, and the proportion of generated utterances explained an additional 6%, the correlation between generativity and MLU was relatively large, which might decrease the sensitivity of this analysis. We therefore assessed the proportion of non-finites in rote utterances only. This increased the proportion of non-finites in the last stage by 10% for Peter's model, and by 20% for Matthijs' model. Apparently, the role of generativity is greater than the regression analysis suggests.

Conclusions

The model described in this paper clearly captures the development that is evident in Dutch children's use of infinitive verb forms. In doing so, the model provides both a process model, and a quantitative account of this transition. Furthermore, it shows that a considerable portion of the drop in non-finite forms can be explained by a learning mechanism that emphasizes utterance final phrases, and an increase in MLU, although the process is likely to be augmented by other considerations (as witnessed by the relatively poor fit for the very early and late stages). While it does not solve the learnability problem, and as such is probably too simplistic a model of syntax acquisition, the present simulations clearly show that the Optional Infinitive phenomenon does not, in itself constitute evidence for the innateness of syntactic knowledge. As such, it supports the suggestion that children's sensitivity to the distributional characteristics of their linguistic environment may aid them in learning their native language. In order to further test this suggestion, it will be necessary to assess to what extent the present findings generalise to other languages. This may also suggest possible extensions to the model.

Acknowledgements

This research was funded by the Leverhulme Trust under grant number F/114/BK.

References

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Croker, S., Pine, J.M., & Gobet, F. (2000). Modelling optional infinitive phenomena: A computational account. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling.* Veenendaal: Universal Press.

Croker, S., Pine, J.M. & Gobet, F. (2001). Modelling children's case-marking errors with MOSAIC. In E.M. Altmann, A. Cleeremans, C.D. Schunn & W.D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling.* Mahwah, NJ: LEA.

Feigenbaum, E.A. & Simon, H.A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336.

Freudenthal, D., Pine, J. & Gobet, F. (2001). Modeling the optional infinitive stage in MOSAIC: A generalisation to Dutch. In E.M. Altmann, A. Cleeremans, C.D. Schunn & W.D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling.* pp. 79-84. Mahwah, NJ: LEA.

Freudenthal, D. Pine, J. & Gobet, F. (2002). Subject omission in children's language: The case for performance limitations in learning. *This Volume.*

Gobet, F. & Simon, H.A. (2000). Five seconds or sixty: Presentation time in expert memory. *Cognitive Science*, *24*, 651-682.

Jones, G., Gobet, F. & Pine, J.M. (2000a). A process model of children's early verb use. In L.R. Gleitman & A.K. Joshu (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. pp. 723-728. Mahwah, N.J.: LEA.

Jones, G., Gobet, F. & Pine, J.M. (2000b). Learning novel sound patterns. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling* (pp.169-176). Veenendaal: Universal Press.

MacWhinney, B. & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17*, 457-472.

Naigles, L. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs. Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, 25, *95-120*.

Shady, M. & Gerken, L.(1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language*, 26, *163-176*.

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivation in child grammar. In N. Hornstein & D. Lightfoot (Eds.), *Verb Movement*. Cambridge: Cambridge University Press.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua*, *106*, 23-79.

Wijnen, F. Kempen, M. & Gillis, S. (2001). Root infinitives in Dutch early child language. *Journal of Child Language*, *28*, 629-660.

Subject Omission in Children's Language: The Case for Performance Limitations in Learning

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.Uk)
Julian Pine (JP@Psychology.Nottingham.Ac.Uk)
Fernand Gobet (FRG@Psychology.Nottingham.Ac.Uk)
School of Psychology, University of Nottingham
University Park, Nottingham, NG7 2RD UK

Abstract

Several theories have been put forward to explain the phenomenon that children who are learning to speak their native language tend to omit the subject of the sentence. According to the pro-drop hypothesis, children represent the wrong grammar. According to the performance limitations view, children represent the full grammar, but omit subjects due to performance limitations in production. This paper proposes a third explanation and presents a model which simulates the data relevant to subject omission. The model consists of a simple learning mechanism that carries out a distributional analysis of naturalistic input. It does not have any overt representation of grammatical categories, and its performance limitations reside mainly in its learning mechanism. The model clearly simulates the data at hand, without the need to assume large amounts of innate knowledge in the child, and can be considered more parsimonious on these grounds alone. Importantly, it employs a unified and objective measure of processing load, namely the length of the utterance, which interacts with frequency in the input. The standard performance limitations view assumes that processing load is dependent on a phrase's syntactic role, but does not specify a unifying underlying principle.

Subject Omission

Children who are acquiring English often produce sentences with missing subjects, like those shown below.

Hug Mummy
Play Bed
Writing Book
See Running

While these examples clearly do not adhere to adult English grammar, many contemporary theories of child language assume that children produce their sentences on the basis of an abstract grammar. Theories differ with respect to how much the hypothesized grammar differs from the adult grammar. According to the *pro-drop hypothesis* (Hyams, 1986; Hyams & Wexler, 1993), children represent a grammar that is different from the adult grammar in that it allows *null subjects.* In this respect, children's grammar resembles that of adult Italian and Spanish speakers. Other authors have argued that children actually possess the correct adult grammar, but drop subjects because they have difficulty expressing the (correct) underlying form due to some kind of processing bottleneck (L. Bloom, 1970; L. Bloom, Miller & Hood, 1975; Pinker, 1984; P. Bloom 1990; Valian, 1991). Thus, a child producing an utterance is thought to represent a grammatically correct underlying structure, but, due to performance limitations, some elements have a lower probability of being expressed than others.

A number of phenomena have been cited as evidence for the performance limitations view. P. Bloom (1990) showed that, in utterances with a subject, the length of the Verb Phrase (VP) is shorter than it is in utterances without a subject. The load associated with the provision of a subject is thought to decrease the likelihood of expressing a longer verb phrase. Along similar lines, the length of the VP is greater when the subject is a pronoun, than when it is a noun. This is thought to result from the fact that pronouns are phonetically shorter, and the fact that non-pronominal subjects may be longer than pronominal ones. L. Bloom (1970) has also found that subject omission is more likely in negated sentences or in sentences with relatively new (unfamiliar) verbs. Presumably, the load associated with negation and novel verbs is such that it induces subject omission.

While the performance limitations view makes sense from an information-processing point of view, it is not very precise in its predictions (Theakston, Lieven, Pine & Rowland, 2001). Performance limitations accounts also tend to be rather ad hoc in nature. Given the imprecise nature of performance limitations, it becomes all too easy to posit a greater processing load whenever the provision of a certain element leads to a greater likelihood of the omission of another, especially when there is an interaction with frequency. Furthermore, it is not clear whether an explanation of the patterns in the data requires a limitation in production coupled with full knowledge of a language's grammar (as the performance limitations view typically has it). In fact, as Theakston et al. point out, performance limited *learning of lexical items* (independent of syntactic complexity) may well give rise to the same pattern of

results without the need to assume a full representation of the grammar, and a different processing load for various types of grammatical roles. The present paper aims to test these claims by seeing to what extent a performance limited distributional analysis of naturalistic input can account for the pattern of omission and provision of grammatical categories that is found in children's speech. To this end, we aim to simulate the effects that P. Bloom (1990) attributes to performance limited production. We will now introduce the model we have used for these simulations.

MOSAIC

MOSAIC (Model of Syntax Acquisition In Children) is an instance of the CHREST architecture, which in turn is a member of the EPAM family of models. CHREST models have successfully been used to model phenomena such as novice-expert differences in chess and computer programming. In language acquisition, MOSAIC has been applied to the modelling of the use of optional infinitives in English and Dutch, the learning of sound patterns and the Verb Island phenomenon. Due to space limitations, we refer the reader to another paper in this volume for the relevant references (Freudenthal, Pine & Gobet, 2002).

The basis of the model is a discrimination net, which can be seen as an index to Long-Term Memory. The network is an n-ary tree, headed by a root node. Training of the model takes place by feeding utterances to the network, and sorting them (see Figure 1). Utterances are processed word by word. When the network is empty, and the first utterance is fed to it, the root node contains no test links. When the model is presented with the utterance *He walked home*, it will create on its first pass three test links from the root. The test links hold a key (the test) and a node. The key holds the actual feature (word or phrase) being processed, while the node contains the sequence of all the keys from the root to the present node. Thus, on its first pass, the model just learns the words in the utterance. When the model is presented with the same sentence a second time, it will traverse the net, and find it has already seen the word *he*. When it encounters the word *walked* it will also recognize it has seen this word before, and will then create a new link under the *he* node. This link will have *walked* as its key, and *he walked* in the node. In a similar way, it will create a *walked home* node under the primitive *walked* node. On a third pass, the model will add a *he walked home* node under the *he walked* chain of nodes. The model thus needs three passes to encode a three-word phrase with all new words. (For expository purposes, here we assume that a node is created with a probability of 1. As is explained under *learning rate*, this probability is actually lower and dependent on a number of factors). Figure 1 shows the development of the net through the three presentations of the sentence.

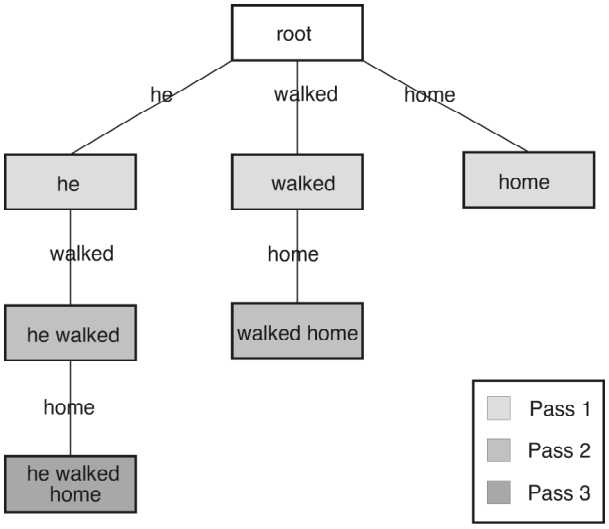


Figure 1: MOSAIC learning an input.

As the model sees more input, it will thus encode longer and longer phrases. Apart from the standard test links between words that have followed each other in utterances previously encountered, MOSAIC employs *generative* links that connect nodes that have a similar context. Generative links can be created on every cycle. Whether a generative link is created depends on the amount of overlap that exists between nodes. The overlap is calculated by assessing to what extent two nodes have the same nodes directly above and below them (two nodes need to share 10% of both the nodes below and above them in order to be linked). This is equivalent to assessing how likely it is that the two words are preceded and followed by the same words in an utterance. Since words that are followed and preceded by the same words are likely to be of the same word class (for instance Nouns or Verbs), the generative links that develop end up linking clusters of nodes that represent different word classes. The induction of word classes on the basis of their position in the sentence relative to other words is the only mechanism that MOSAIC uses for representing syntactic classes.

The main importance of generative links lies in the role they play when utterances are generated from the network. When the model generates utterances, it will output all the utterances it can by traversing the network until it encounters a terminal node. When the model traverses standard links only, it produces utterances or parts of utterances that were present in the input. In other words, it does *rote* generation. During generation, however, the model can also traverse generative links. When the model traverses a generative link, it can

supplement the utterance up to that point with a phrase that follows the node that the current node is linked to. As a result, the model is able to generate utterances that were not present in the input. Figure 2 gives an example of the generation of an utterance using a generative link.

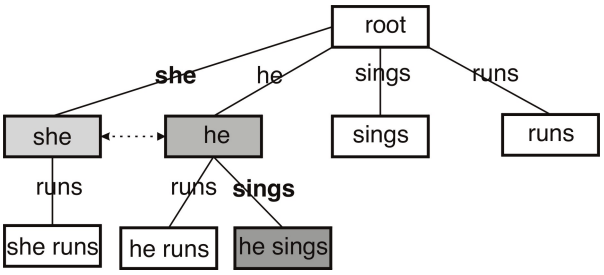


Figure 2: Generating an utterance. Because *she* and *he* have a generative link, the model can output the novel utterance *she sings*. (For simplicity, preceding nodes are ignored in this figure.)

Learning Rate

MOSAIC does not simply learn all the utterances it encounters. The probability of the creation of a node is dependent on the size of the net and the length of the utterance it encodes. This has the effect of making the learning process frequency sensitive. If an utterance is seen more often, it has a higher probability of being created. Finally, phrases that occur in an utterance final position in the input (have an *end marker*) have a higher probability of being encoded. The precise formula governing learning rate is given elsewhere in this volume (Freudenthal, Pine & Gobet, 2002).

Performance Limitations in MOSAIC

The only performance limitations in MOSAIC are the following:

- Frequency: high frequency items have a higher likelihood of being encoded, and thus feature in longer utterances
- Short phrases have a higher likelihood of being encoded than long phrases
- Utterance final phrases have a higher likelihood of being encoded.
- An utterance will only be produced (generated) if its final phrase has occurred in sentence final position in the input.

It may be appropriate to point out that these performance limitations are plausible from general theorizing in the cognitive psychology and learning literature. Huttenlocher et al. (1991) provide evidence for the effect of frequency on vocabulary learning. Evidence for the importance of sentence final position has been provided by Naigles & Hoff-Ginsberg (1998) and has been attributed to prosodic highlighting of the sentence final position (Shady & Gerken, 1999). In contrast to the standard performance limitations view, processing load in MOSAIC does not vary as a function of grammatical role. Also note that the version of MOSAIC used for these simulations is identical to that which Freudenthal, Pine & Gobet (2002) used for the simulation of the optional infinitive phenomenon in Dutch. No free parameters were fitted to obtain these results.

Subject Omission in MOSAIC

MOSAIC creates utterances without subjects because the model can output partial utterances, provided that the utterance final element has occurred in a sentence final position in the input. As a consequence, constituents that take a position early in the sentence, have a higher probability of being omitted than those that take a position further *downstream*. Since the subject takes first position in English, it has the highest likelihood of being omitted. However, this prediction is not tied to the English language. MOSAIC would generate utterances with omitted subjects in all languages that have the subject as the first element in their underlying word order.

Method

In order to simulate the data presented by P. Bloom (1990), two MOSAIC models were trained using corpora of maternal speech available in the CHILDES database (MacWhinney & Snow, 1990). We used the files of Anne and Becky. The mean length of utterance (MLU) in the output generated from the models was 2.87 for Anne's model, and 3.41 for Becky's model. In line with Bloom's analysis, we limited our analysis to utterances which could not be interpreted as imperatives. This is necessary as subjectless sentences in English are grammatical as imperatives (e.g. *Put it down*). Bloom selected a list of *nonimperative verbs* and *past tense verbs* for his analysis. Since these verbs cannot be used in an imperative form, sentences which contain a verb from these lists, and do not contain a subject, are true examples of subject omission. Tables 1 and 2 give the lists of verbs that were used for these analyses.

Table 1: Nonimperative verbs used for analysis

Care	Laugh	Miss
Cary	Laughs	Need
Fall	Like	See
Falls	Live	Sneeze
Forget	Lives	Want
Grow	Love	Wants
Know	Loves	

In line with Bloom's analysis, we removed from our samples all questions, all utterances that contained the words *not* or *don't*, all utterances where the verb was not used in a productive way, and all utterances where the target verb was part of an embedded clause.

Table 2: Past tense verbs used for analysis

Ate	Fixed	Saved
Bit	Folded	Saw
Bought	Forgot	Sent
Broke	Found	Sharpened
Brought	Gave	Spilled
Came	Goed	Stepped
Caught	Ironed	Stopped
Closed	Left	Thought
Cooled	Lost	Throwed
Covered	Made	Took
Cried	Melted	Tored
Drinked	Pee-peed	Tripped
Dropped	Pulled	Turned
Dropt	Rode	Washed
Falled	Said	Went
Fell	Sat	Wrote

Table 3 gives the data for three children that Bloom reports and the two simulations (Anne's and Becky's model). It can be seen that for the children, the Verb Phrase length in utterances with a subject is shorter than in utterances without a subject. It can also be seen that MOSAIC readily simulates this result, and the size of the effect is quite comparable to that in the children. The difference in verb phrase length is statistically significant for both Anne's model ($t(330) = 4.82$, $p < .001$), and Becky's model ($t(314) = 4.64$, $p < .001$).

Table 3: Mean length of Verb Phrases in sentences with and without subjects

Child	With Subject	Without Subject
Adam	2.33	2.60
Eve	2.02	2.72
Sarah	1.80	2.46
Anne's Model	2.14	2.76
Becky's Model	2.58	3.31

MOSAIC obtains this result because the probability of learning an item in MOSAIC is dependent only on its frequency and length, and not on its grammatical role. There is thus no reason (apart from differences in frequency), why sentences with subjects should, on average, be longer (or shorter) than those without. The fact that verb phrases in utterances with subjects should be longer than verb phrases in utterances without a subject is a straightforward consequence of this fact.

A second analysis performed by Bloom was to look at the length of the verb phrase as a function of the type of subject (no subject, pronoun or non-pronoun). The reasoning was that, since the processing load of a subject is higher than that of a missing subject, and the processing load of a non-pronoun subject is higher than that of a pronoun (since the pronoun is both phonetically shorter as well as shorter in word length), this should again result in length effects on the Verb Phrase. The results of this analysis are shown in table 4.

Table 4: Mean length of Verb Phrase as a function of subject size

	No Subject	Pronoun	Non-Pronoun
Adam	2.60	2.55	2.25
Eve	2.75	2.30	2.00
Sarah	2.45	1.90	1.50
Anne's Model	2.76	2.45	1.60
Becky's Model	3.31	2.93	1.67

Again, it is clear that MOSAIC has no difficulty in simulating these results (though the size of the effect in MOSAIC appears to be slightly larger than in the children that Bloom analysed). The difference in verb phrase length between utterances with a pronoun subject and those with a non-pronominal subject is statistically significant for both Anne's model ($t(64) = 3.45$, $p < .001$) and Becky's model ($t(104) = 4.40$, $p < .001$). There are two possible reasons why MOSAIC might simulate this result. Firstly, non-pronoun subjects are on average slightly longer than pronoun subjects. Pronouns are by definition one word long, while non-pronoun NP's can contain determiners and adjectives. In fact, Bloom indicates that the average non-pronoun subject for the children he analysed was 1.16 words long. Secondly, pronouns have a higher frequency of occurrence than non-pronominal subjects. In MOSAIC, this increases the likelihood that they will be learnt, and the likelihood that they will feature in longer utterances. We decided to test these two explanations in MOSAIC by performing the analysis on non-pronominal subjects of length one and greater separately. As it turned out, only a small proportion of the non-pronominal subjects had a length greater than one. For non-pronominal subjects of length one, the size of the VP was 1.58 for Anne's model, and 1.88 for Becky's model[1]. Both values are smaller than the VP length for pronoun

[1] One would expect the length of the verb phrase to increase when limiting this analysis to subjects of length 1. This is the case for Becky's model, but not for Anne's model. This is due to the fact that, for Anne's model, there were relatively few long non-pronominal subjects, but one of those that did occur had a particularly long verb phrase.

subjects. Given the low incidence of long non-pronominal subject in both these and Bloom's data, this clearly indicates that the lower complexity effect that Bloom attributes to the fact that pronouns are phonetically shorter, can be explained by frequency in the input. Note that MOSAIC does not employ a phonetic component. Phonetic differences can therefore not have contributed to MOSAIC's simulation of the effect.

The importance of frequency in the input as an explanation for the difference between pronouns and non-pronouns is also highlighted by a point made by Hyams & Wexler (1993). Though pronouns may be phonetically shorter, the process of assigning the referent to a (potentially ambiguous) pronoun may actually result in its processing load being higher, rather than lower. This would predict a shorter Verb Phrase length for pronominal than for non-pronominal subjects.

Subject versus Object Omission

It has often been shown that subjects are omitted more often than objects. In order to test how often objects are omitted, Bloom selected utterances which contain verbs that require an object, and calculated the proportion of object omission from these obligatory contexts. Table 5 shows this list of verbs.

Table 5: Verbs that take obligatory objects.

Bought	Ironed	Saved
Broke	Like	Saw
Brought	Love	See
Caught	Loves	Sharpened
Covered	Made	Thought
Drinked	Miss	Throwed
Fix	Need	Took
Folded	Pulled	Want
Found	Rode	Wants
Gave	Said	Washed

Table 6 compares the proportion of omitted subjects and objects from obligatory contexts (verbs from tables 1 and 2 for subjects, verbs from table 5 for objects). It can be seen that the proportion of subject omission is considerably higher than the proportion of object omission. The subject-object asymmetry was significant for both Anne's model (X^2 (1, N = 560) = 98.83, p < .001), and Becky's model (X^2(1, N = 548) = 125.97, p < .001). Bloom suggests several possible causes for this asymmetry. Firstly, it may be due to pragmatic factors. Since subjects typically convey given information, while objects convey new information, it may be more pragmatically appropriate to omit subjects when processing capacity is limited. A second possible cause might be that there is a 'save the heaviest for last' bias. This would result in subjects having a higher processing load than objects, and as a result, in them being omitted more often.

Table 6: Omission from obligatory contexts

	Subjects	Objects
Adam	57%	8%
Eve	61%	7%
Sarah	43%	15%
Anne's Model	64%	21%
Becky's Model	60%	14%

The explanation for the effect in MOSAIC is simple. As a result of MOSAIC's performance limitations, a constituent is less likely to be omitted when it occurs further toward the end of the sentence. Since subjects take first position, and objects usually come after the verb, the probability of omitting an object is smaller than the probability of omitting a subject. Bloom goes on to suggest that the hypothesized processing asymmetry should cause other differences between subjects and objects. For example, since pronouns exert less of a processing load, more pronouns will occur in subject position than in object position. Table 7 shows the relevant data, both for Bloom's analysis, and MOSAIC's simulations. Again, the assymmetry is significant for Anne's model (X^2 (1, N = 243) = 8.08, p < .01), and Becky's model (X^2(1, N = 292) = 27.53, p < .001).

Table 7: Proportion of overt pronominal Noun Phrases

	Subjects	Objects
Adam	41%	25%
Eve	36%	14%
Sarah	91%	33%
Anne's Model	47%	27%
Becky's Model	72%	40%

There is no specific reason why MOSAIC would predict this effect, but the pragmatic factors that Bloom mentions may well explain this result. Subjects tend to convey given information, and objects tend to convey new information. It certainly makes sense to introduce new information using a non-pronoun NP. The use of a pronoun requires the listener to resolve the referent of the pronoun. The use of a non-pronoun NP is usually less ambiguous, which aids the resolution process. In fact, several authors have argued that this is the preferred argument structure for English (Clancy, 2001). As such, it is not just a feature of child language, but is actually the preferred structure in adult language. The fact that MOSAIC simulates this result is simply a

reflection of the fact that it mimics the distribution of the input.

Conclusions

MOSAIC clearly simulates all the results that Bloom reports. MOSAIC is not an ad hoc model of subject omission, as it has already been shown to account for several phenomena in children's speech, and is firmly grounded in the CHREST/EPAM framework. Though MOSAIC has performance limitations, these reside mainly in the learning mechanism. Unlike the standard performance limitations view, MOSAIC does not assume full competence. In fact, MOSAIC has no built in knowledge regarding grammatical categories or roles. The effects arise in MOSAIC through a combination of performance limited distributional learning, and frequency sensitivity. Effects that are present in the input (such as a higher proportion of pronominal subjects than objects), are mimicked in MOSAIC's output because of the fact that it is a distributional analyser.

On a theoretical level, MOSAIC has two main strengths over the standard view of performance limitations. Firstly the definition of processing load in the standard view is somewhat ad hoc. If the provision of certain elements leads to a higher rate of subject omission, this is seen as evidence for a relatively high processing load of these elements. The actual reason for this high processing load then varies from effect to effect. Within MOSAIC, processing load is a function of the interaction of two objectively measurable variables: frequency in the input, and length of the phrase being encoded. When an item is more frequent in the input, it has a higher likelihood of being encoded, and therefore features in longer utterances that have a higher likelihood of being grammatical (i.e. including the subject). If two utterances have equal overall frequency, and one of the two includes a longer element (verb phrase), then some other element will necessarily be omitted. Since the subject is the first element in the sentence, this has a higher likelihood of being omitted.

Secondly, the standard performance limitations view assumes a large amount of innate knowledge in the child. For the simulation of these results, MOSAIC assumes no innate syntactic knowledge.

Acknowledgments

This research was funded by the Leverhulme Trust under grant number F/114/BK.

References

Bloom, L. (1970). *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA: MIT press.

Bloom, L., Miller, P. & Hood, L. (1975). Variation and reduction as aspects of competence in language development. In A. Pick, (ed.): *The 1974 Minnesota Symposium on Child Psychology*, Minneapolis: University of Minnesota Press.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry, 21, 491-504*.

Clancy, P. (2001). The lexicon in interaction: Developmental origins of preferred argument structure in Korean. In J.W. DuBois, L.E. Kumpf & W.J. Ashby (Eds), *Preferred argument structure: Grammar as architecture for function*. Amsterdam: John Benjamins.

Freudenthal, D. Pine, J. & Gobet, F. (2002). Modelling the development of Dutch optional infinitives in MOSAIC. *This Volume*.

Huttenlocher, J. Haight, W., Bryk, A. Seltzer, M. & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology*, 27, *236-248*.

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*, Dordrecht: Reidel.

Hyams, N. & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry, 24, 421-59*.

MacWhinney, B. & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17, 457-472.

Naigles, L. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs. Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, 25, *95-120*.

Pinker, S. (1984). *Language Learnability and Language Development*, Cambridge, MA: Harvard University Press.

Shady, M. & Gerken, L. (1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language*, 26, *163-176*.

Theakston, A.L., Lieven, E.V.M., Pine, J.M., Rowland, C.F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28, 127-152*.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40, 21-81*.

Does Positivity Bias Explain Patterns of Performance on Wason's 2-4-6 Task?

Maggie Gale (m.gale@derby.ac.uk)
University of Derby
Western Road, Mickleover, Derby DE3 9GX, U.K.

Linden J. Ball (l.ball@lancaster.ac.uk)
Department of Psychology, Lancaster University,
Lancaster, LA1 4YF, U.K.

Abstract

In the standard form of Wason's (1960) 2-4-6 task, participants must discover a rule that governs the production of sequences of three numbers. Studies typically show success rates of approximately 20%, which Wason attributed to a cognitive deficit that he labeled 'confirmation bias'. In Tweney et al.'s (1980) formally equivalent Dual Goal (DG) form of the task, however, success rates are at least double to those seen on the standard task. If this facilitated performance could be accounted for, then this would go some way toward explaining the normally low performance on the standard problem. The present experiment examined two competing accounts of the DG superiority effect: Evans' (1989) positivity bias explanation, and Wharton, Cheng and Wickens' (1993) goal complementarity theory. The experiment independently manipulated the number of goals that participants had to explore (a single goal vs. two complementary goals) and the linguistic labels used to provide feedback (DAX and MED vs. 'fits the rule' and 'does not fit the rule'). Results supported the goal complementarity account in that facilitation was evident in both DG conditions irrespective of the polarity of the feedback provided. We also discuss a novel finding: that it is the production of at least a single 'negative' triple that is most closely associated with task success.

Introduction

Poletiek (2001) summarises hypothesis testing as comparing internal thoughts with external facts in order to interact with the world. For example learning a language can be characterised as hypotheses testing as the learner utters sounds and observes the listener's reactions. Hypothesis testing, therefore, can be viewed as a fundamental mode of mental functioning, and for this reason is of considerable interest to psychologists and cognitive scientists alike.

One important experimental paradigm that has been employed extensively in order to study hypothesis-testing behaviour is the 2-4-6 task, introduced by Peter Wason in 1960. The 2-4-6 task is a deceptively simple rule discovery task, which Wason originally devised to investigate whether people conformed to the contemporary scientific norm of hypothesis testing, namely falsification (Popper, 1959). In the standard version of the 2-4-6 task, participants seek to discover a rule which generates sequences of three numbers (referred to as *triples*). They are initially given an example, conforming triple (2-4-6), and are then required to produce further triples which the experimenter classifies as either conforming to, or not conforming to, the rule. The to-be-discovered rule is 'any ascending sequence'. Participants produce triples until they are confident that they know the rule, at which point they announce it. Despite the seeming simplicity of the task, participants perform poorly, with typically only around 20% correctly announcing the rule on the first attempt, (e.g., Tukey, 1986; Wason, 1960; Wharton, Cheng & Wickens, 1993). Many of these incorrect announcements are a more restricted version of the rule, for example, 'numbers increasing by two'. It has been suggested (e.g., Wetherick, 1962) that the initial 2-4-6 exemplar lures participants into formulating such overly-restricted hypotheses. Participants then produce triples motivated by these hypotheses (e.g., 8-10-12), which always receive positive feedback, since they form a subset of the target rule. Faced with repeated confirmations of their hypothesis, participants seemingly become increasingly confident of its correctness until they announce it as the rule. It is clear that unless participants change their testing strategy they will never discover that although their hypothesis is sufficient, it is not necessary.

In his analysis of participants' performance on the task, Wason showed that solvers and non-solvers could be differentiated in terms of both the number of triples they produced (with solvers producing reliably more triples), and the type of triples generated (with solvers producing a higher proportion of triples which received negative feedback). Wason viewed the non-solvers' strategy of testing positive instances of their hypothesised rule as a cognitive failing, which he labeled 'confirmation bias'. However, Klayman and Ha (1987), in an elegant conceptual analysis of the underlying structure of the 2-4-6 task and its variants,

demonstrated that it is the *relationship* of the hypothesised rule to the target rule in the original task which causes participants' failure to discover the target rule. Klayman and Ha ague that what Wason regarded as a bias to seek confirmatory evidence could instead be conceptualised as a 'positive test strategy', which in certain circumstances is an effective method for yielding disconfirmations of a current hypothesis. Their essential point (cf. Wetherick's, 1962, argument noted earlier) is that in the standard task, the target rule ('any ascending sequence') has been deliberately designed to be more general than the hypothesis invited by the given triple, such that the application of a positive test strategy can never lead to the discovery of the target rule. For other target rule/hypothesis relationships, however, such as where the experimenter's rule (e.g., 'even numbers ascending by two and less than 10') is more restricted than the participant's initial hypothesis (e.g., 'numbers increasing by two'), then the implementation of positive testing would lead rapidly to falsification of the overly general initial hypothesis, and to accurate rule discovery.

Dual Goal Instructions

Tweney, Doherty, Worner, Pliske, Mynatt, Gross, and Arrkelin (1980) introduced a modified form of the task, in which participants were instructed to discover two rules, one called DAX, the other MED. The DAX rule governs triples of the traditional ascending type, all other triples are MEDs. Although formally equivalent to Wason's original task, this simple Dual Goal (DG) manipulation was seen to have a dramatic effect on success rates, with 60% of participants making a correct first announcement of the rule. This facilitated performance has been shown to be a robust finding that has been replicated many times (e.g., Farris & Revlin, 1989a, 1989b; Tukey, 1986; Wharton, Cheng & Wickens, 1993). Tweney et al., were at a loss to explain the facilitatory effect of the DG manipulation, although they felt that the explanation was somehow related to the way participants conceptualise the task, and how triples produced are related to their conceptualisation.

It has been noted that the DG manipulation has the effect of increasing the *number* of triples which are generated before rule announcement, and also the *variety* of triples (Gorman, Stafford & Gorman, 1987; Tukey, 1986; Tweney et al., 1980). Vallée-Tourangeau, Austin and Rankin (1995), in their replication and extension of DG instructional effects, formulated two measures of triple heterogeneity, namely *posvars* and *negtypes.* Posvars are triples which receive positive feedback but which do not increase by a constant. Thus, if the numbers that make up a triple are *a, b,* and *c*, a posvar is a triple in which *(b-a)* (*c-b*). Negtypes are the eight possible types of triples which receive negative feedback, (e.g., descending triples, identical-number triples, etc.). Vallée-Tourangeau et al. found that using these indices of triple heterogeneity, the DG manipulation led to increased production of both posvars and negtypes compared to Single Goal (SG) instructions. They interpreted this as being indicative of participants considering a wider range of hypotheses, although they did not directly test this claim. Whilst these observations of triple heterogeneity are interesting, they are largely descriptive, and do little to explain the facilitatory effect of DG instructions.

Evans (1989) proposed that poor performance on the standard task could be attributed to the operation of a general 'positivity bias', which is a form of selective processing that causes people to attend to positive rather than negative information. According to this proposal, facilitated performance using DG instructions is caused by the labeling of triples that 'do not fit the target rule' as MED, thereby avoiding a negative label, and hence counteracting participants' tendencies not to attend to this information. Evans (1989) argues that in the standard version of the 2-4-6 task, if a participant forms the hypothesis 'numbers ascending by equal intervals is right', they have logically also formed the hypothesis 'numbers not ascending by equal intervals is wrong'. They are, however, not aware of this alternative hypothesis, and therefore do not test it. In the DG manipulation, however, because participants are attempting to discover two rules, they test both DAX (correct) triples, and MED (incorrect) triples. Participants are more successful with the DG instructions since by carrying out positive tests of their MED hypotheses they are effectively carrying out negative tests of their DAX hypotheses, thus eliminating the overly restrictive hypotheses typically announced by the SG non-solvers. In summary, then, Evans argues that DG instructions facilitate performance by changing participants' representation of the task by creating a positive label for the previously negative 'does not fit' feedback.

Wharton et al. (1993) proposed a subtly different mechanism by which DG instructions improve performance. They invoke Klayman and Ha's (1987) proposal that a central feature of hypothesis testing behaviour is a tendency for individuals to adopt a positive test strategy which leads to the generation of triples that match their hypothesised rule. As we noted earlier, in the standard 2-4-6 task positive testing will never enable participants to discover the overly restrictive nature of their hypothesis as they will never generate a triple which lies outside of their hypothesis yet is still within the experimenter's target rule. With DG instructions, however, even though the exemplar triple for DAX suggests the same restricted hypothesis, the requirement to discover the second (MED) rule should encourage participants to form a second hypothesis (e.g., 'numbers ascending by intervals other

than two' are MED). On carrying out a positive test of the MED hypothesis, (e.g., 5-10-15) participants will (unexpectedly) receive DAX feedback, thus causing them to alter both their DAX and MED hypotheses. This sequence of events is repeated until satisfactory rules for DAX and MED are discovered. Thus, according to Wharton et al., it is the *complementary* nature of the two rules that leads to task success.

It is clear that Evans' (1989) positivity-bias account and Wharton et al.'s (1993) goal-complementarity account make very different predictions regarding performance on the 2-4-6 task. Evans' theory predicts that participants given positively-labeled DAX and MED feedback in relation to generated triples will perform better than those given a combination of 'fits the rule' (positively labeled) and 'does not fit the rule' (negatively labeled) feedback, and that this dissociation will be present irrespective of whether participants are asked to discover a single target rule or two complementary rules. In contrast, the goal-complementarity account proposed by Wharton et al. predicts that participants given the task of discovering two complementary rules will be more successful than those seeking a single rule, regardless of whether feedback is given as DAX/MED or 'fits'/'does not fit'. Previous studies of the facilitatory effect of DG instructions have always confounded these two variables, such that participants given DG instructions have always been given DAX/MED feedback, whilst those given SG instructions have always received 'fits'/'does not fit' or 'yes'/'no' feedback. The present experiment was designed to discriminate between the positivity-bias account and the goal complementarity theory, by manipulating these two factors independently. To this end, the participant's goal (i.e., discovery of one rule vs. discovery of two complementary rules) was systematically crossed with the linguistic label of the feedback (i.e., DAX/MED vs. 'fits the rule'/'does not fit the rule').

Method

Participants

Sixty undergraduates of varying backgrounds and ethnicity from the University of Derby took part in the experiment in exchange for course credits. They had not received any teaching on the psychology of reasoning before the experiment.

Design

A fully between participants design was employed that manipulated two factors: Goal (Single Goal vs. Dual Goal), and Linguistic Labeling (DAX/MED vs. Fits/Does Not Fit). Fifteen participants were randomly assigned to each of the four resulting conditions.

Procedure

Participants were tested in groups of up to four in a quiet laboratory. Standardised instructions were read to each group. Single Goal (SG) instructions referred to a unique rule: *'I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule'.* In what we subsequently refer to as the SG—Fits condition, participants were asked to discover the target rule by generating triples which they would be told either 'fitted' or 'did not fit' the rule that the experimenter had in mind. On the other hand, in what we refer to as the SG—DAX condition, participants were told that triples that fitted the rule were called DAX triples and those that did not fit the rule were called MED triples. It was explained to participants that on generating a triple they would be informed as to whether it was a DAX or a MED triple.

The Dual Goal (DG) instructions emphasised that there were two rules to be discovered: '*…your task is discover this rule, and also a second rule for categorising the triples that do not fit my rule*'. In the standard DG task (i.e., DG—DAX), participants were additionally informed that triples that fitted the rule were called DAX triples and those that did not fit the rule were called MED triples. They were instructed to produce further triples, which the experimenter would describe as either DAX or MED. In the DG—Fits condition, participants were told to generate triples which the experimenter would classify in terms of whether they 'fitted' or 'did not fit' the rule.

Participants in all conditions were given 2-4-6 as the example triple. All participants were provided with an answer sheet and were asked to write 2-4-6 on the first row, and either 'fits' or DAX in the feedback column, as appropriate. They were instructed that they could produce as many triples as they wished, and that when they were sure of the rule(s) they should write it (or them) on the answer sheet. In line with Gorman (1992), participants were allowed only one guess at the rule(s).

Results

Success Table 1 shows the frequency of correct and incorrect announcements by participants in each of the four experimental conditions.

Table 1: Frequency of correct announcements by condition.

Condition	N	Solvers	Non-Solvers
SG—DAX	15	3	12
SG—Fits	15	3	12
DG—DAX	15	12	3
DG—Fits	15	11	4

Table 2: Mean number of triples (and type of triples) produced by condition.

Condition	Total Triples	Posvars	Feedback	Negtypes
SG—DAX	7.6 (6.25)	0.33 (0.62)	0.73 (1.28)	0.33 (0.62)
SG—Fits	5.87 (2.75)	0.53 (1.36)	0.93 (1.58)	0.80 (1.42)
DG—DAX	10.27 (6.30)	1.13 (1.06)	2.67 (1.95)	1.20 (0.86)
DG—Fits	8.33 (3.22)	1.07 (1.10)	1.4 (1.24)	0.93 (0.80)

Note: SD in parenthesis.

Labeling of feedback (DAX/MED vs. Fits/Does Not Fit) seems to have had little effect on the likelihood of success; the proportions of solvers were similar between both the two SG groups and also the two DG groups. However more than three times the number of participants in the DG conditions than in the SG conditions announced the correct rule. A contingency table chi-square analysis was performed on the frequencies of correct and incorrect announcements (pooled over the labeling of feedback), and revealed a highly significant effect of the SG versus DG manipulation, $\chi^2(1) = 19.288$, $p < .001$.

Feedback Analyses were also performed to ascertain whether the manipulated factors had any effect on the type or number of triples produced (see Table 2). Again, the labeling of the feedback appeared to make little difference to the number or type of triples produced by participants. With regard to the SG versus DG instructions, however, there were significant main effects on three of the measures: number of triples produced, $F(1, 56) = 4.09$, $p < .05$; number of triples receiving negative feedback, $F(1, 56) = 9.1$, $p < .01$; and number of variable positive triples, $F(1, 56) = 5.86$, $p < .05$. The difference in the number of negtypes produced across SG and DG conditions also approached significance, $F(1, 56) = 3.96$, $p = .052$. There were no significant interactions for any of the measures.

Presence of Triple Types Although the analyses of triple type are interesting they do not give insight into the absolute importance of the production of the triple types. For this reason, it was decided to carry out further analyses in which the production of either a posvar or a negative triple was crossed with success on the task. In this way it would be possible to test whether the production of such triples is necessary for task success.

Table 3: Frequency of correct announcements by production of at least one posvar.

	Solvers	Non-Solvers	Total
Posvar produced	25	10	35
No Posvar produced	6	19	25
Total	31	29	60

A contingency table was, therefore, produced in which the production of *at least one* posvar was crossed with success (see Table 3). The table clearly demonstrates that the production of a single posvar is associated with success on the task, with four times the number of participants who produced a posvar making a correct announcement compared to those who did not produce one. A chi-square analysis confirmed the reliability of this observation, $\chi^2(1) = 13.137$, $p < .001$.

Table 4 shows a contingency table in which the production of at least one negative triple is crossed with success. Here the association is even more marked than in the case of the production of at least a single posvar, with there being only one instance of a participant who had not produced a negative triple correctly announcing the rule. In contrast, of the 34 participants who did produce a negative triple, 28 solved the task. A chi-square analysis revealed that these differences were highly significant, $\chi^2(1) = 36.363$, $p < .001$).

Table 4: Frequency of correct and incorrect announcements by production of a negative triple.

	Solvers	Non-Solvers	Total
Negative triple present	28	6	34
Negative triple absent	1	25	26
Total	29	29	60

Discussion

The results of the present experiment clearly support the goal complementarity account (Wharton et al., 1993) of the facilitatory effect of DG instructions on the 2-4-6 task. Evans' (1989) positivity-bias account, on the other hand, fails to find support in the evidence presented. The results show that DG superiority cannot be attributed to the re-labeling of negatively valenced 'does not fit' feedback as positive 'MED' feedback, as participants in the DG conditions performed significantly better than participants in the SG conditions, regardless of the nature of their feedback. This leads to the conclusion that the typically poor success rates on the standard form of the task cannot be accounted for by participants selectively attending to

positive information and thus ignoring a potentially informative set of triples. In relation to this point, the analyses of triple type and triple production show that participants in the DG condition produced a greater number and variety of triples. It could, therefore, be argued that it is not that SG instructions lead to selective processing of negative information, but rather that SG instructions do not promote the exploration of negative information in the first place (cf. Wharton et al., 1993).

The final set of analyses also revealed a hitherto unremarked phenomenon. It has long been noted that people who solve the 2-4-6 task tend to produce more triples as well as a greater proportion of negative triples (Wason, 1960). It has also been demonstrated more recently that solvers generate a greater variety of triples (e.g., Vallée-Tourangeau et al., 1995). What has not previously been shown, however, is that it is the production of at least a *single* negative triple that is so closely associated with success on the task. Indeed it remains possible that other indices of success such as the total number of triples produced or overall triple variety may well be mediating factors through which the critical negative triple is produced as a result of task manipulations. This is an area which would seem to require closer investigation.

The basic observation that negative-triple production is so closely related to task success, does, at first sight, appear rather paradoxical The point is, that given the typically overly-restrictive hypotheses which participants form, it seems intuitively obvious that it should be the production of the discriminatory *posvars* (rather than negative triples) that would be most strongly associated with task success. Although our results do indicate that posvar generation is significantly linked to correct initial rule announcements on the task, it remains striking that the production of negative triples is even more predictive of task success. Why might this be the case?

One possibility is that the production of a *descending* triple (and its associated MED or 'does not fit' feedback) somehow makes the general dimension of *ascending numbers* appear to be relevant to the target DAX or 'fits' rule. The concept 'descending' may have this effect by facilitating the establishment of a salient contrast class that promotes an insight into the potential scope of the target rule. Closer investigation of the precise role of negative triples in facilitating task success - perhaps through the invocation of clear contrast sets within the space of possible triples - would, therefore, appear to be essential. To achieve this a finer-grained system of codifying the triples that participants produce may be required.

In summary, the results of this study clearly support a goal complementarity account of facilitated performance using DG instructions on the 2-4-6 task. Participants in the DG conditions were more successful at the task than those in the SG conditions. The lack of effect with regard to the labeling of feedback would appear to undermine a standard positivity-bias account. Further work, however, is vital to understand the role that negative triples play in determining task success.

References

Evans, J. St. B. T. (1989). *Bias in human reasoning*: *Causes and consequences.* Hove: Lawrence Erlbaum Associates, Inc.

Farris, H., & Revlin, R. (1989a). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory and Cognition, 17,* 221-232.

Farris, H., & Revlin, R. (1989b). The discovery process: A counterfactual strategy. *Social Studies of Science, 19,* 497-513.

Gorman, M. E. (1992). Experimental simulations of falsification. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking: Vol. 1*. Hemel Hempstead: Harvester Wheatsheaf.

Gorman, M. E., Stafford, A., & Gorman, M. E. (1987). Disconfirmation and dual hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology* 39A, 1-28.

Klayman J., & Ha, Y-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Poletiek, F. H., (2001). *Hypothesis-testing behaviour.* Hove: Psychology Press.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Tukey, D. D. (1986). A philosophical and empirical analysis of subjects' modes of inquiry in Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 38A,* 5-33.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arrkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32,* 109-123.

Vallée-Tourangeau, F., Austin, N. G., & Rankin, S. (1995). Inducing a rule in Wason's 2-4-6 task: A test of the information-quantity and goal-complementarity hypotheses. *Quarterly Journal of Experimental Psychology, 48A,* 895-914.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14,* 129-140.

Wharton, C. M., Cheng, P. W., & Wickens, T. D. (1993). Hypothesis-testing strategies: Why two goals are better than one. *Quarterly Journal of Experimental Psychology, 46A,* 743-758.

A Connectionist Model of Planning via Back-chaining Search

Max Garagnani
Department of Computing
The Open University
Milton Keynes, MK7 6AA - UK
M.Garagnani@Open.ac.uk

Lokendra Shastri and **Carter Wendelken**
The International Computer Science Institute
Berkeley, CA 94704 USA
Shastri@ICSI.Berkeley.EDU
CarterW@ICSI.Berkeley.EDU

Abstract

A connectionist model for emergent planning behavior is proposed. The model demonstrates that a simple planning schema, acting in concert with two general purpose cognitive functionalities, namely, episodic memory and perception, can solve a restricted class of planning problems by backchaining from the goal to the current state. In spite of its simple structure, the schema can search for and execute plans involving multiple steps. We discuss how this simple model can be extended into a more powerful and expressive planning system by incorporating additional control and memory structures.

Introduction

Consider a classical planning problem, specified by an initial state, a goal state and a set of operators. A direct approach to solving this problem consists of searching the state space to find a 'path' between the initial and final states. Several symbolic planning systems adopting this approach in conjunction with the use of heuristics (Hoffman & Nebel, 2001; Haslum & Geffner, 2000; Bacchus & Teh, 1998) have recently shown notable improvements in efficiency on various benchmark problems.

Although a state space search algorithm is conceptually simple, it is not obvious how the data structures and control mechanisms required for the specification and execution of such an algorithm can be realized in a neurally plausible manner. In this paper we propose a connectionist model that exhibits a state-space search behavior. The model uses only a few simple control structures in conjunction with essential cognitive faculties, such as episodic memory, semantic memory, and perception.

Episodic memory (Tulving, 1995) refers to our ability to remember specific events and situations in our daily lives. The use of memory and experience in planning and reasoning has been investigated by several researchers (see (Waltz, 1995; Spalazzi, 2001) for useful accounts). Neurological and psychological data strongly suggests that episodic memory is distinct both in its functional characteristics and neural basis from other forms of memories such as semantic memory, memory for common sense knowledge, and procedural knowledge. It has been argued that events and situations in episodic memory are best viewed as *relational instances* that specify a set of bindings between the roles of a relational schema and objects that fill these roles in a given event or situation (Shastri, 2001b; 2002). We assume that a planning agent is capable of remembering past events such as "performing action A under conditions P lead to consequences C". Each episodic memory trace of this type can be represented as a triple of the form *Preconditions, Action, Consequences*, and we will refer to such triples as PAC memories (or events)[1].

Finally, we assume that the planning agent is capable of observing the current world state through perception. By this we mean that the agent can determine whether or not certain perceptually salient and directly observable relations hold in the world. For example, in the context of the classical blocks world scenario, this assumes that the agent can look at the table and determine whether or not a specific block is 'clear.'

A memory-based planning schema

Figure 1 shows the abstract structure of the proposed planning schema. In order to explain its behavior, let us describe the functionality of each of its components and their interactions using a simple example.

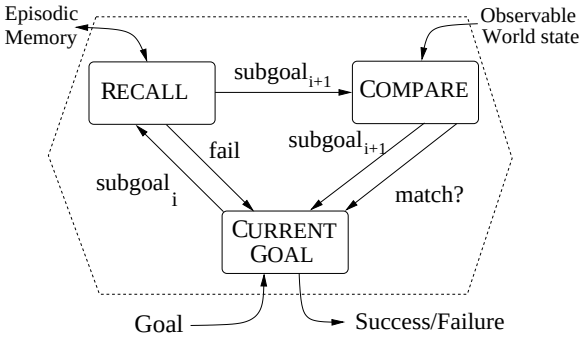


Figure 1: A block diagram showing the basic components of the planning mechanism.

Consider a planning problem with only two blocks ('A' and 'B') where the agent's goal G is to achieve `On(A,B)` (i.e. block A on block B) and the current world state is `On(B,A)`. Let us assume that the agent's episodic memory contains the two PAC events E_1 and E_2:

[1]Note that preconditions and consequences may contain multiple predicates.

E_1) In state P_1={On(B,A)},
action A_1= Unstack(B,A)
led to C_1={OnTable(A),OnTable(B)};
E_2) In state P_2={OnTable(A),OnTable(B)},
action A_2= Stack(A,B)
led to C_2={On(A,B)}.

(1) Initialize

The schema activity is initialized by conveying the goal G ($\equiv$ On(A,B) in our example) as input to CURRENT GOAL [2]. In the absence of any incoming activity from COMPARE, CURRENT GOAL simply passes control along with the goal (subgoal $_i = G$) to the RECALL component. The behavior of the CURRENT GOAL component in the presence of an input from COMPARE is different, and is described below.

(2) Recall

RECALL is activated when it receives the subgoal $_i$ (On(A,B), in the example) as input from CURRENT GOAL. The function of this component is to search episodic memory for an event wherein a specific *action* (say, A) performed under some specific *preconditions* (say, P) lead to a set of *consequences* (say, C) in which subgoal $_i$ is true. In our example, E_2 happens to be such an event. When a matching event is found, action A and preconditions P are recollected and become 'active'; P becomes the current focus of the agent's attention[3], and control is transferred to COMPARE, along with P as the current subgoal $_{i+1}$. In the example, subgoal $_{i+1}$ = P_2 ={OnTable(A),OnTable(B)}. If there are no events whose consequences 'match' the subgoal $_i$, the schema execution halts and signals a *failure*.

(3) Compare

The COMPARE block compares subgoal $_{i+1}$ with the current world state, which is assumed to be observable through perception. It returns a positive outcome *iff* subgoal $_{i+1}$ is true in the current state. In the example, the response is negative, as the world is still On(B,A). After the comparison, the outcome and the subgoal $_{i+1}$ (=P) are passed to CURRENT GOAL, which takes control of the activity and reacts as explained below.

(4) Repeat

If CURRENT GOAL receives a negative result from COMPARE, the following happens:

- the original goal G is no longer passed as input to RECALL, and ceases driving the activity of the schema;
- the set of preconditions P (=subgoal $_{i+1}$) becomes the new subgoal $_i$ and is passed to RECALL by the CURRENT GOAL component.

[2]It is assumed that this goal is represented in other networks outside the planning schema and communicated to the schema via controlled spreading activation.

[3]In order to achieve the original goal G, it suffices to achieve P and execute action A.

At this point, one loop is completed and the procedure repeats from step (2) with P as the current goal.

If CURRENT GOAL receives a positive input from COMPARE, the schema terminates returning *Success* and the control is given to an appropriate 'Action schema' that will carry out the action currently active in memory (A). If the original goal G is not achieved via the execution of this action, the planning schema is re-invoked and re-initialized with G. Note that we are not assuming the existence of a working memory which would allow the agent to dynamically store sub-goals during the planning process or to maintain active more than *one* PAC event at a time. Because of this, the proposed system is forced to 're-discover' parts of the same plan every time an action is executed, as described below.

Returning to the example, the result of COMPARE is negative and no action is performed: G is 'forgotten', while P_2 becomes the new subgoal $_i$ and is passed on to RECALL. The schema now queries the episodic memory by asking if {OnTable(A),OnTable(B)} has been achieved in the past: PAC event E_1 is recollected. The precondition P_1={On(B,A)} – required to perform Unstack(B,A) – becomes the new focus of attention and is compared with the current world state, producing a successful outcome: a chain of (two) PAC events connecting the goal to the initial state has been found, and the planning problem has been (potentially) solved. However, because of the absence of a working memory able to dynamically store goals and subgoals, all the agent can 'see' at this point is the last PAC event recollected. The currently active action (A_1=Unstack(B,A)), though, can and *should* be executed, since this will get the current state one step closer to the goal. After the positive outcome of COMPARE, the schema terminates returning '*Success*': the agent carries out the currently active action Unstack(B,A), and the new state of the world becomes {OnTable(A),OnTable(B)}.

After the action has been completed, the agent is 're-exposed' to the initial goal G (which has not been achieved yet), and the planning schema is re-invoked. The subsequent flow of activity is identical to the first part of the previous one, except that now P_2={OnTable(A),OnTable(B)} matches the current state of the world, and thus action A_2= Stack(A,B) is executed. This leads to achieving the original goal G, and the schema is no longer invoked.

The connectionist planning schema

The planning mechanism described above has been implemented using the representational machinery of SHRUTI, a neurally plausible structured connectionist architecture that demonstrates how a network of neuron-like elements can encode a large body of structured knowledge and perform a variety of inferences within a few hundred milliseconds (Shastri & Ajjanagadde, 1993; Shastri, 1999; Shastri & Wendelken, 2000).

SHRUTI suggests that the encoding of relational information (frames, predicates, and schemas) is mediated by

neural circuits composed of *focal-clusters*, and that the dynamic representation and communication of relational *instances* involves the transient propagation of *rhythmic* activity across these clusters. A role-entity binding is represented in this rhythmic activity by the *synchronous* firing of appropriate cells. Rules are encoded by links that enable the propagation of rhythmic activity across focal clusters, and a fact in long-term memory is a temporal pattern matching circuit.

In the past, SHRUTI's representational machinery has been used to encode commonsense knowledge (Shastri & Ajjanagadde, 1993), causal models (Shastri & Wendelken, 2000), as well as action schemas and reactive plans (Shastri, Grannes, Narayanan & Feldman, 1997) and decision-making (Wendelken, 2001).

A memory-based proto-planner

Consider the network structure depicted in Figure 2. This network fragment consists of two 'control' focal-clusters ACHIEVE and RECALL, two predicate focal-clusters On and OnTable, an action focal-cluster Unstack, two entity focal-clusters A and B, and a type focal-cluster Block. Typically, a focal-cluster contains several *control* and *role* nodes. For example, the focal-cluster ACHIEVE contains control nodes +, -, and ?, and role nodes I and G (the entity and type focal-clusters do not contain role nodes).

Role nodes within the focal-cluster of a predicate or schema provide a mechanism for expressing role (or parameter) bindings. In particular, a dynamic binding between a role and its filler is expressed by the synchronous firing of the role node and the focal-cluster of the object filling the role. A relational focal-cluster with bound role nodes designates a particular *relational instance*.

The *enabler* (?) node associated with a focal-cluster may be viewed as an "initiate query" or "initiate activity" node. In contrast, *collector* nodes (+ and –) associated with a focal-cluster indicate the outcome of a query or of other activity pertaining to the focal-cluster. In particular, the activation of the + (–) collector indicates a positive (negative) response to a query or signals a successful (unsuccessful) completion of some activity.

Inference occurs via the propagation of activity between focal-clusters. The links between the enabler and role nodes of interconnected focal-clusters allow queries posed in one focal-cluster to propagate to other focal-clusters: if role node R_1 is linked to role node R_2, the firing of R_1 induces synchronous firing in R_2, allowing dynamic binding propagation.

A query is communicated to a focal-cluster by activating its enabler node and binding its role nodes to appropriate role fillers. In Figure 2, the query "Can block B be placed on the table, given that B is on A?" is communicated by activating ?:ACHIEVE, and synchronizing the firing of ACHIEVE.I and +:On; the firing of ACHIEVE.G and +:OnTable; the firing of On.x, OnTable.x and ?:B; and that of On.y and ?:A. The activity of ACHIEVE propagates to the RECALL cluster, resulting in the query "Is there some action which led to OnTable(B) from the precondition On(B,A)."

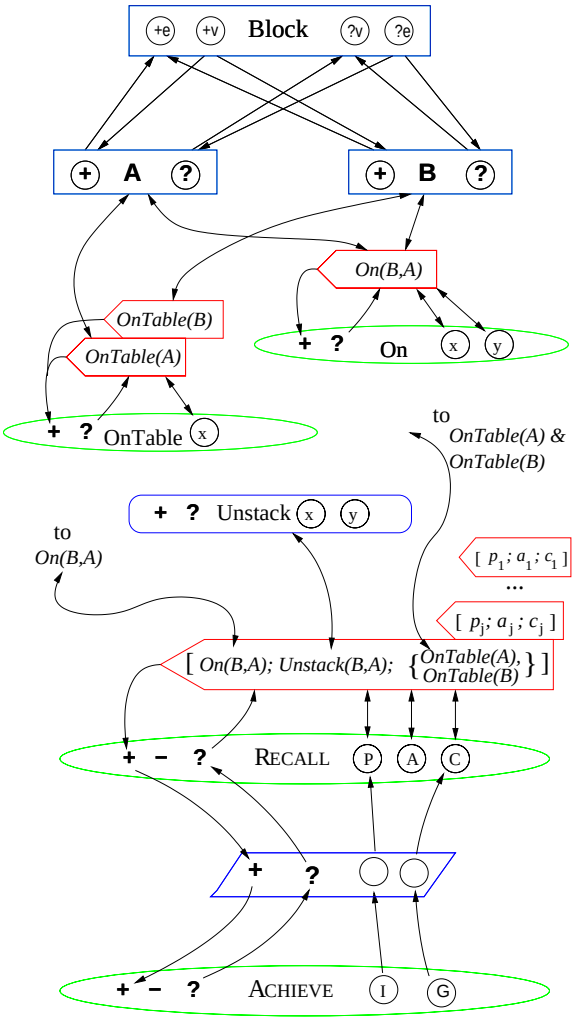


Figure 2: diagram showing network structure for a proto-planning mechanism.

Fact structures attached to a relational focal-cluster encode specific instances of that relation. If the query active at a focal-cluster matches an attached fact, the fact becomes active and, in turn, activates the positive collector of the relation's focal-cluster, binding (via synchronous firing) each of the relation's role nodes to the entity filling these roles in the fact. A neurally plausible model of how this might happen in the brain is described in (Shastri, 2001b).

In the case of RECALL, the facts structures attached to this focal-cluster represent episodic memories of specific PAC events. The activation of the query "Is there some action which led to OnTable(B) from the precondition On(B,A)" matches the memorized PAC event and leads to this event becoming active (i.e., recalled). This in turn results in the activation of +:RECALL and the synchronous firing of the unbound role RECALL.A with the representation of Unstack(B,A): the system re-

calls that performing `Unstack(B,A)` when `On(B,A)` was true lead to `OnTable(B)` being true.

An important feature of the system consists of its ability to treat a relational instance as a role-filler (e.g. `ACHIEVE.I` $\equiv$ `On(B,A)`). In order to support this requirement, SHRUTI allows for two levels of temporal synchrony. Bindings between standard role nodes and entity/type nodes are represented within a rapid *minor* oscillatory cycle, while bindings between specialized role nodes and relational instances are encoded within a slower *major* oscillatory cycle.

The simple schema described above, consisting of the `ACHIEVE` and `RECALL` clusters acting in concert with the episodic memory, can retrieve previously memorized '*if-then*' (PAC) tuples. Thus, this schema can be construed as a *proto-planner* capable of returning one-step 'plans.' The next section demonstrates how this schema can search for *sequences* of actions, therefore constituting the next 'stage of evolution' of this proto-planner.

The planning schema in SHRUTI

Figure 3 shows how the planning schema of Figure 1 has been implemented using SHRUTI's representational machinery. It is easy to see how the focal clusters of this schema can be mapped to the elements of Figure 1 (the CURRENT GOAL block has been realized with two clusters, `PLAN` and `SUBGOAL`). We shall use the same example adopted earlier to illustrate how the schema can perform a basic form of *planning as search*.

Let us assume that the memory of the agent (represented only abstractly in the figure) contains the two PAC events of the previous example, namely, E_1 = (P_1, A_1, C_1) and E_2 = (P_2, A_2, C_2), and that the PERCEPTION block, when queried with input P, activates the + or – collector depending on whether the event bound to P is true or false in the observed world state.

The schema is invoked by activating the `PLAN` cluster's enabler ('?') node and by binding its role node G to the relational instance expressing the current goal (`On(A,B)` in the example). After initialization, activity propagates upwards along links to clusters `SUBGOAL` and `RECALL`. After few major cycles, `?:RECALL` is activated, with role C firing in synchrony with the current goal {`On(A,B)`}. The PAC event E_2 matches the activity in the cluster and is retrieved. Consequently, `+:RECALL` becomes active, and the roles A and P are instantiated with actions A_2 = `Stack(A,B)` and relational instance P_2 = {`OnTable(A,B)`}, respectively (the clusters corresponding to predicate 'OnTable' and action 'Stack' are not shown in the figure). The activity of P reaches `COMPARE`. Since `OnTable(A,B)` is not true in the current world state, `-:Compare` becomes active. This leads to the inhibition of the links from `PLAN` to `SUBGOAL`, which blocks the propagation of the query `PLAN(On(A,B))` through the schema. Simultaneously, activity from `COMPARE` reaches `SUBGOAL`: the role node G starts firing in synchrony with P, which was temporally bound to the relational instance {`OnTable(A,B)`}. Hence, this becomes the new (sub)goal of the schema, and its focus of attention. Activity from `?:SUBGOAL` reaches `?:RECALL` again, while role node C starts firing in synchrony with G. This leads to the retrieval of PAC event E_1, and hence, the precondition P_1 = {`On(A,B)`} gets bound to role P and role A is bound to the action instance A_1 = `Unstack(B,A)`. These bindings are in turn propagated to `COMPARE`.

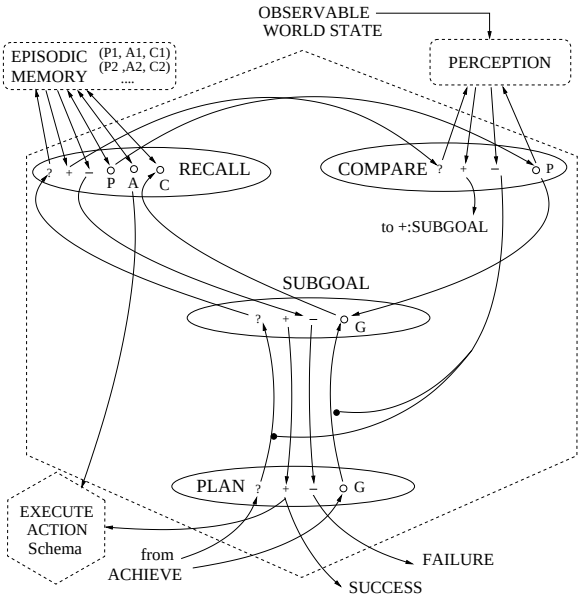


Figure 3: A diagram showing the connectionist structure of the planning schema.

The positive outcome of the comparison leads to the activation of `+:COMPARE`, which communicates to `PLAN` that the search has terminated successfully. After action A_1 is executed, the initial goal G={`On(A,B)`} (not yet achieved and still present in the system) causes the schema to restart. The subsequent flow of activation is identical to the initial part of the previous sequence, except that when P_2 = {`OnTable(A,B)`} is compared with the current state, the outcome is positive and the currently active action (`Stack(A,B)`) is executed. This achieves the goal G and terminates the activity.

Simulation results

The above planning schema has been realised and tested using the "SHRUTI Agent Simulator" software written in Java. The example described in the previous section has been used to test the functioning of the schema. Figure 4 shows the detailed trace of activation resulting from the actual simulation. Note how the diagram reflects closely the flow of activity described before, up to the first positive outcome of `COMPARE`.

Consider, for example, time point α of the diagram. Here, the PAC fact E_2 has just become active because of the initial query '`PLAN(On(A,B))`', which has been propagated upwards and has led to the query '`RECALL(`x`,`y`,On(A,B))`'. As a consequence,

the two preconditions {OnTable(A),OnTable(B)} are about to become active, and will be propagated to the COMPARE cluster, where they will be matched against the current state of the world[4].

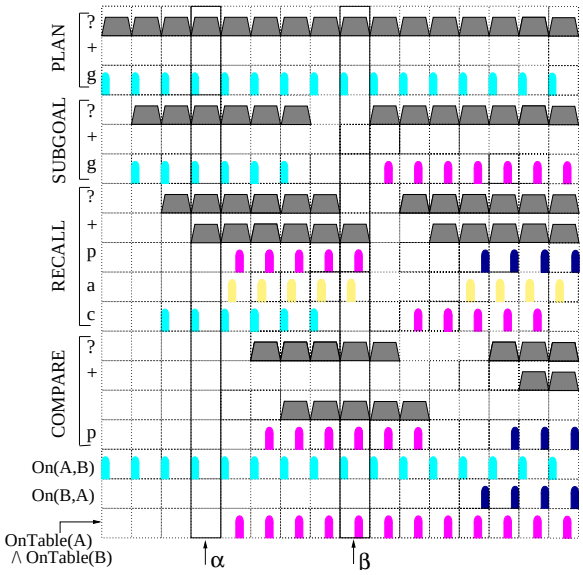


Figure 4: Node activation trace of the simulation.

Time point β is a snapshot of the situation immediately following the negative outcome of the (simulated) comparison. Notice how the ow of activity going from the PLAN cluster to the RECALL cluster has been interrupted by the negative outcome of COMPARE in order to allow the new goal {OnTable(A),OnTable(B)} to make its way through to cluster RECALL.

Discussion

The work described in this paper is part of a larger effort whose goal is to develop a neurally plausible architecture for reasoning, remembering, planning, and decision making. This paper presents progress along an important dimension of this ongoing effort. Perhaps the most interesting aspect of this work is the demonstration that general purpose cognitive faculties such as episodic memory, semantic memory and perception can be harnessed to produce a state-space search behavior and solve a subclass of planning problems.

The planning schema discussed in this paper is limited in a number of ways; however, as discussed below, this schema can be extended into a much more powerful and expressive planning system by incorporating additional control and memory structures, and by leveraging the full representational and expressive power of SHRUTI.

The proposed planning schema is susceptible to getting trapped in deadends. As the system searches for a path from the goal to the initial state, it can get caught in a state that subsumes a set of conditions D which do not match the consequent of any PAC event in memory. There is, however, a simple three part solution to this problem. First, the agent detects that it has reached a deadend state (this is signaled by the activation of -:RECALL. Second, the agent memorizes that this path leads to a deadend in the context of the current problem. It can do so by memorizing the following episodic memory trace: ìwhen trying to achieve the goal G, instantiating a subgoal D leads to a deadendî. [5] Third, the agent restarts the search and at each step in the search process retrieves both PAC and DEADEND events that match the current subgoal. Any retrieved PAC event that is counterindicated by retrieved DEADEND event is ignored. Since the memorization of deadends prunes the potential search space, with suf cient practice, the agent may memorize a large number of deadend events and carry out a highly ef cient search.

Another limitation of the proposed planning schema is that it needs to traverse the same paths through the state space several times during the course of nding a plan. However, if the agent could remember the path traversed from the current goal to the initial state, it would not have to rediscover the same plan subsequences many times over: plan execution would involve traversing the memorized sequence of PAC events only once (in the reverse order) and executing the actions associated with each PAC event in the sequence. Note that remembering such a path can be viewed as memorizing a sequence of PAC events. Learning of event sequences is a well-known property of episodic memory, but it remains to be seen how the process of such on-line memorization of event sequences can be fully integrated with the on-line retrieval of previous episodic memories. Working memory mechanisms can also play a complementary role in such on-line memorization. Our current research addresses the functioning of episodic memory (Shastri, 2001b; 2002) as well as that of working memory, and we hope that the development of powerful episodic and working memory models will directly bene t future work in the development of planning schemas.

Since the proposed planning schema operates within the SHRUTI architecture, the full range of knowledge representation and reasoning capabilities of SHRUTI can be leveraged during planning. This includes representing and reasoning with commonsense (semantic) knowledge, causal models, type hierarchies, context-sensitive prior probabilities of events and estimated utility/value of world-states. Thus, general purpose domain knowledge as well as planning speci c knowledge can be seamlessly combined to support planning involving not just memory retrieval, but also inference.

The functionality of the current planning schema is

[4]The perceptual task of verifying whether some conditions hold in the current world state was simulated by manually activating the + or − collector of the cluster ëCompareí as appropriate.

[5]The representational machinery required to encode such DEADEND ìeventsî is similar to that required to encode PAC events: like PAC events, the episodic memory trace of DEADEND events also involves role- llers that are partial state-descriptions, speci ed by sets of conditions.

also limited by its inability to make use of goal decomposition. Imagine that the agent is trying to nd a plan for the goal $G_1 \& G_2$ given the world state I and the two PAC events PAC(I', a_1, G_1) and PAC(I'', a_2, G_2) in memory. The planning schema described in this paper will be unable to solve the composite goal G_1 & G_2, even though it will be able to solve each of the subgoals G_1 and G_2 if presented individually[6]. In order to deal with goal decomposition, the schema must (i) recognize that it can solve one of the subproblems using one of the PAC facts, (ii) pick the subproblem to be solved, (iii) note down the subproblem that it is deferring for now, (iv) nd a solution to the selected subproblem, (v) shift attention back to the deferred subproblem, and (vi) solve the deferred subproblem. A connectionist implementation of this algorithm would require a more complex schema (control structure) than the one described in the previous sections, together with the ability to remember deferred goals. The memory of deferred goals can take the form of working memory (if deferred goals have to be remembered for a few seconds) or episodic memory (if the goals have to be remembered over longer time periods).

Another area of ongoing research of direct relevance to the work described here concerns the representation of complex action schemas and plans. In past work, we have shown that parameterized schemas capable of dealing with partially ordered actions, conditional actions, concurrent and iterative actions, as well as compositional and hierarchical actions can be encoded using SHRUTIís representational machinery (Shastri et al., 1997). This makes us con dent that the more complex control structures required for encoding more sophisticated planning schemas would not present an insurmountable problem.

A key issue that remains open is the learning of appropriate control structures. We are investigating this question within the frameworks of spike-timing dependent synaptic plasticity (Wendelken & Shastri,2000; Song, Miller & Abbott, 2000) and recruitment learning based on long-term potentiation (Malenka & Nicoll, 1999; Shastri, 2001a).

Acknowledgments

This work was partially funded by NSF grants 9720398 and 9970890.

References

Bacchus, F., & Teh, Y. W. (1998). Making forward chaining relevant. *Proceedings of the Fourth International Conference on AI Planning Systems (AIPS 1998)* (pp. 54ñ61).

Haslum, P., & Geffner, H. (2000). Admissible Heuristics for Optimal Planning. *Proceedings of the 5th Internat. Conf. of AI Planning Systems (AIPS 2000)* (pp. 140ñ 149). Breckenridge, Colorado: AAAI Press.

Hoffmann, J., & Nebel, B. (2001). The FF Planning System: Fast Plan Generation Through Heuristic Search. *Journal of Artificial Intelligence Research, 14*, 253ñ 302.

Malenka, R. C., & Nicoll, R. A. (1999). Long-term Potentiation - A Decade of Progress? *Nature, 285*, 1870ñ 1874.

Shastri, L. (1999). Advances in SHRUTI - a neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence, 11*.

Shastri, L. (2001a). A Biological Grounding of Recruitment Learning and Vicinal Algorithms. In J. Austin, S. Wermter & D. Wilshaw (Eds.), *Emergent neural computational architectures based on neuroscience*. Springer-Verlag.

Shastri, L. (2001b). A computational model of episodic memory formation in the Hippocampal system. *Neurocomputing, 38-40*, 889ñ897.

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences, 6(4)*, 162ñ168.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning. *Behavioral and Brain Sciences, 16(3)*, 417ñ494.

Shastri, L., Grannes, D., Narayanan, S. & Feldman, J. (1997). A Connectionist Encoding of Parameterized Schemas and Reactive Plans. In G. Kraetzschmar and G. Palm (Eds.), *Hybrid Information Processing in Adaptive Autonomous Vehicles*. Springer-Verlag.

Shastri, L., & Wendelken, C. (2000). Seeking coherent explanations - a fusion of structured connectionism, temporal synchrony, and evidential reasoning. *Proceedings of the Twenty-Second Conference of the Cognitive Science Society*. Philadelphia.

Song, S., Miller, K., & Abbott, L. (2000). Competitive Hebbian Learning Through Spike-Timing Dependent Synaptic Plasticity. *Nature Neuroscience, 3*, 919ñ926.

Spalazzi, L. (2001) A Survey on Case-Based Planning. *Artificial Intelligence Review, 16(1)*, 3ñ36.

Tulving, E. (1995) Organization of Memory: Quo Vadis? In M.S. Gazzaniga (Ed.), *The Cognitive Neuroscience*. MIT Press.

Waltz, D.L. (1995) Memory-based reasoning. In: M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*. MIT Press.

Wendelken, C., & Shastri, L. (2000). Probabilistic inference and learning in a connectionist causal network. *Proceedings of the Second International Symposium on Neural Computation*.

Wendelken, C. & Shastri, L. (2002). SHRUTI-agent: A structured connectionist model of decision-making. *Proceedings of the 24th Conference of the Cognitive Science Society*. Washington, D.C. August, 2002.

[6] That is, assuming that I' and I'' hold in state I, and that I'' also holds in the state resulting from performing action a_1 in state I.

Events Versus States: Empirical Correlates of Lexical Classes

Silvia Gennari (sgen@wam.umd.edu)
Cognitive Neuroscience of Language Laboratory
1401 Marie Mount Hall. University of Maryland, College Park, MD 20742

David Poeppel (dpoeppel@deans.umd.edu)
Cognitive Neuroscience of Language Laboratory
1401 Marie Mount Hall. University of Maryland, College Park, MD 20742

Abstract

Philosophers and linguists have claimed that verb meanings are divided into semantic types or superordinate categories that differ in internal conceptual structure. In particular, eventive verbs, which have internal causal structure are distinguished from stative verbs, which have no internal causal structure. In this paper, we explore the processing consequences of assuming that the lexical representations of verb meanings differ in the complexity of their internal representations. We conducted two experiments, a lexical decision task and a self-paced reading study, that investigated how verb types of different complexity are processed. We predicted that the conceptually more complex eventive verbs would take longer to process than stative verbs. In both experiments, this prediction was confirmed. This lends support to theories of verb concepts that propose classifications based on internal representations and shows that there are discrete and abstract conceptual categories in the domain of events.

Introduction

An important question in cognitive science concerns how word meanings (or lexical concepts) are internally represented. Although considerable progress has been made in the domain of nominal concepts since Rosch's studies, the nature and organization of verb concepts is less well understood. Early studies on verb meanings investigated whether verbs had internal semantic structure, as proposed in linguistic theories, but failed to find evidence supporting such a view (e.g., Fodor, Garrett, Walker, & Parkes 1980, Kintsch, 1974, Rayner & Duffy, 1986). For example, Rayner and Duffy (1986) measured the eye-fixation time on verbs during reading that were assumed to differ in internal complexity. They found no reading time differences corresponding to the semantic complexity of the verbs. This sort of finding, together with Fodor and colleagues' theoretical arguments (Fodor, 1975, Fodor, Fodor, & Garrett, 1975, Fodor & Lepore, 1998), was taken to support the view that verb meanings are atomic and lack internal structure. However, recent psycholinguistic studies challenge this view. Several sentence processing experiments have shown that lexical semantic properties such as selectional restrictions and verb-specific thematic roles (agent vs. patient) are quickly accessed by the processor when parsing syntactic ambiguities (e.g., Trueswell, Tanenhaus & Kello, 1993, Trueswell, Tanenhaus & Garnsey, 1994). More relevant to verb concepts per se, McRae, Ferretti & Amyote (1997) have shown that thematic roles have internal conceptual structure (as object categories do) and that their feature structure is quickly accessed by the parser when resolving syntactic ambiguities. Moreover, Ferretti, McRae & Hatherell (2001) have shown that verbs prime their typical agents, patients and instruments (e.g., *praying* primes *nun*). They argue that verbs activate event schemas or generalized situation based knowledge that facilitate accessing the meaning of their typical participants. Finally, McKoon & Macfarland (2000) have found processing correlates of two types of verb meanings, those that are conceptualized as either externally caused events (e.g., *break*) or internally caused ones (e.g., *grow*). These verb types are assumed to differ in internal lexical complexity, particularly in their causal components (see also Gentner 1981). Taken together, these findings suggest that there is some internal structure in verb meanings: thematic structure and event types.

The work presented here further investigates verb concepts, i.e., how verbs, which refer to events, are processed and internally represented. In particular, we ask whether there are verb-general concepts and structures beyond and above the idiosyncratic meanings of individual verbs. We follow numerous linguistic and philosophical studies (as in McKoon & Macfarland, 2000) in assuming common structural and causal properties across classes of verbs that define superordinate concepts. Thus, beyond the existence of typical agent-verb-relations (that between *nuns* and *praying*), there may be more abstract structural or conceptual properties that organize our knowledge of events stored in the lexicon.

The classification of verbs and their semantic properties has been the topic of numerous philosophical and linguistic studies (Vendler, 1967). Following traditional Aristotelian classes, these studies have argued that there is a typology of events underlying verb uses. Verb types appear to be universal (Smith, 1991) and are supposed to reflect the way speakers conceptualize the domain of events, i.e., the semantic/conceptual properties they assign to a particular actual occurrence. One general distinction typically made between verb meanings is, among others, that between states and events (Vendler 1967, Dowty, 1979, Taylor 1977, Bach, 1986, Verkuyl, 1993, Jackendoff, 1990, Rappaport-Hovav & Levin 1998). The distinction seems cognitively basic because it is grounded in causal properties: eventive verbs typically denote a cause and a change from an initial state to a resulting one (e.g. *write*, *destroy*), while stative verbs simply denote properties or stable relations between participants (e.g. *love*, *belong, contain*) (Dowty, 1979, Parsons 1990). The

distinction presupposes that verb lexical meanings have internal conceptual structures that differ in complexity: event lexical concepts have internal sub-components derived from their causal properties (e.g., the cause and the resulting state), while states lack any such components.

In this paper, we investigate how verbs denoting states and events are processed and represented. In particular, we ask whether eventive and stative verbs, which drastically differ in their semantic-conceptual complexity, are processed in ways consistent with their complexity. Two psycholinguistic experiments show that speakers' processing of verb meanings varies according to lexical semantic complexity, thus supporting the view that eventive and stative verbs are represented differently.

The Distinction between Verb Classes

Eventive and stative verbs are distinguished by semantic and syntactic properties. Syntactically, they differ in their ability to co-occur with certain adverbs and in certain constructions. These distributional restrictions are taken as tests that identify membership in one verb class or another. For example, stative verbs such as *deserve* are distinguished by their ability to occur with simple present in English but not present progressive, as in (1) and (2). Similarly, stative verbs cannot occur in nominalized constructions such as that in (3), cannot appear as complements of verbs like *force* as in (4), and cannot be modified by adverbial phrases as in (5). In contrast, eventive verbs such as *build* have the opposite distributions (for more tests, see Dowty 1979):

(1) Bill *is deserving** / *is building* something.
(2) Bill *deserves* / *builds** something.
(3) What Bill did was to *deserve** / *build* something.
(4) Bill forced Mary to *deserve** / *build* something.
(5) Bill *deserved** / *built* something in an hour.

The intuition behind (1)-(4) is that the participants of a stative eventuality are not causal or volitional agents. Rather, they are *experiencers*. The two classes thus involve different relations between their participants. The intuition behind (5) is that states persist in time while events have ending points or culminations. Examples of each verb type are given in Table 1.

Table 1: Verb examples

Events	*States*
enter	live
accuse	love
create	contain
give	know
build	despise
buy	constitute
betray	cherish

Semantically, the classes are distinguished by logical entailments. When each verb type occurs in sentences, they allow or disallow distinctive inference patterns. The entailments refer to causal and temporal properties of the construed eventuality as a whole. Stative sentences imply facts, i.e., they entail that they hold true for an indefinite period of time. In contrast, eventive verbs have the change-of-state entailment: they either entail a single change of state, as in (6) and (7) below (in Vendler's classification, achievement and accomplishment verbs), or sequence of changes as in (8) (activity verbs). These changes are caused by either an agent's single act or a series of actions that may be sustained for a while. But in contrast to states (which can persist on their own) events stop when their cause does so that they do not hold for indefinite periods.

(6) *x* killed *y* → *x* caused *y* to become dead
(7) *x* built *y* → *x* caused *y* to become existent
(8) *x* hammered *y* → *x* caused *y* to become hammered

The temporal entailments distinguishing each class are the counterpart of their causal properties. Because states have no internal (causal) structure, they are true at a given interval as well as at any subpart of this interval. More precisely, if a state is true at any interval *i*, it follows that it is true at all instants within *i,* as in (9).

(9) If Bill had a bike last week, he had a bike throughout the week.

This entailment is called temporal homogeneity. In contrast, events lack this property (Dowty 1979): if a change of state is true at any interval *i*, it follows that it is false at the initial part of *i*, (the initial state) and it is true at the final part of *i* (the final state) as in (10):

(10) If Bill wrote a letter in an hour, the letter was not written before the hour and was written right after it.

The entailment captures the fact that single-change events typically have sub-parts (the initial and resulting state) so that the event as a whole cannot be true until it is completed. This also holds for activity events at the level of each component change. It follows that events are not temporally homogeneous because their truth at a given interval does not hold at any sub-instant.

It is clear that whether one focuses on their temporal or causal properties, states are fundamentally different from events. While states lack internal causal structure and are temporally homogeneous, events have complex causal structures and are not temporally homogeneous.

The contrast between these verb classes has led linguists to propose conceptual lexical representations that capture their semantic properties and the relations between their participants. The representations are expressed via logical operators (Dowty 1979) or primitive

predicates (Rappart Hovav & Levin 1998, Jackendoff, 1990) such as CAUSE and BECOME, together with verb-specific lexical predicates that hold true of their arguments. Consider the representations of the following verbs:

write = *x* CAUSE(BECOME(*y be-written*))
break = *x* CAUSE(BECOME (*y be-broken*))
possess = *x* possess *y*
deserve = *x* deserve *y*

Because verbs of the same type have structurally similar representations, the distinction between events and states can be expressed in verb-general conceptual structures:

Events *x* CAUSE(BECOME(*y state*))
States *x* state *y* = *state(x, y)*

Eventive representations typically involve changes and causes, while stative representations simply involve a stative predicate that holds true of participants. In fact, stative predicates are component part of events, because changes include resultative states. This renders eventive verbs as semantically more complex than stative verbs.

Experimental Evidence

If the mental representations of verbs in fact differ due to their causal and temporal properties, this suggests that each verb type may involve differential processing cost depending on internal complexity. Representing the meaning of an eventive verb entails representing different alternative states of affairs such as the initial state and final state resulting from the agent's intervention. In contrast, representing the meaning of a stative verb implies representing a single state of affairs. If processing a verb implies accessing and processing its lexical meaning, more complex meanings should yield longer processing time. To test this empirical prediction, we conducted two psycholinguistic experiments. The first experiment was a visual lexical decision task. This task has been shown to be sensitive to top-down influences of meaning (see Balota, 1994, Balota, Ferraro & Connor, 1991) and several semantic effects such as abstract vs. concrete aspects of meaning have been reported (Blesdale, 1987, Eviatar, Menn & Zaidel, 1990, Paivio, 1991). The second experiment was a self-paced reading study, in which the reading time of verbs (occurring in sentences) was measured. Previous literature has shown that lexical complexity factors such as number of senses (Rayner & Duffy 1986) and type of verbs (McKoon & Macfarland 2000) have an effect on reading times.

Experiment 1: Visual Lexical Decision Task

In this experiment, we ask whether stative verbs are recognized faster than eventive verbs, given the hypothesized semantic complexity differences.

Materials 31 and 32 words were selected for each verb type (states and event) according to the semantic and syntactic criteria discussed in Dowty, 1979. The items were matched for word length, frequency (Associate Press Corpus, mean frequency for events = 2.40, and for states = 2.45), number of sense (WordNet database: events: 2.79; states: 2.59), number of orthographic neighbors (events: 2.23, states: 1.85) and argument structure. Verbs were transitive verbs (taking obligatory NP or PP complements), except for 6 intransitive verbs in each class, and were not ambiguous between noun and verb uses. Non-words (= 62) were possible words similar to real words. This favors deeper processing of words, because written form is not sufficient discriminator to perform the task (see Seidenberg, Petersen, MacDonald & Plaut, 1996).

In a pre-test study, imageability ratings (how easy is it to imagine/visualize the meaning of a word) were collected from another set of subjects to control for the possibility of this effect. We subsequently used the items' imageability ratings as a covariate in our analysis. The rationale for incorporating this factor derives from the observation that higher imageability ratings are associated with faster reaction times (James, 1975, Paivio, 1991, Strain, Patterson & Seidenberg, 1995). We used the instructions provided in Chiarello, Shears, & Lund's (1999) norming. Comparisons of these ratings across categories revealed that the categories differed, with eventive verbs being more imagaeble (mean for events = 4.21 in a scale from 1 to 7, mean for states = 3.25, p= .001).

Examples of test words are the following:

Events: betray, borrow, conquer, create, deduce, align, attract, devour.
States: adore, aspire, believe, belong, cherish, comprise, contain, deserve, detest.

Participants and Procedure 52 right-handed native speakers of English participated in this study, all students at the University of Maryland. For each word presented in the screen, participants decided whether it was a word of English. The experiment was carried out in G3 Macintosh computer running Psyscope. Words were presented at varying inter-trial time (500-1500 ms) on the center of the screen. Before each stimulus word or non-word, a fixation point was presented for 500 ms. The reaction time (RT) to each stimulus was automatically collected.

Results Analysis of covariance across items with RTs as dependent variable and imageability ratings as covariate revealed a significant main effect of imageability ($F(1,60)= 7.19$, $p= .009$), a main effect of word type ($F(1, 60)= 7.95$, $p= .006$) and no interaction. The overall word effect is represented in Figure 1, with bars representing standard error. Mean difference was small but reliable, (about 20 ms), because not all state/event pairs show big

differences. We also conducted an analysis across subjects with similar result ($F(51,1)= 40.21$, $p = .0001$).

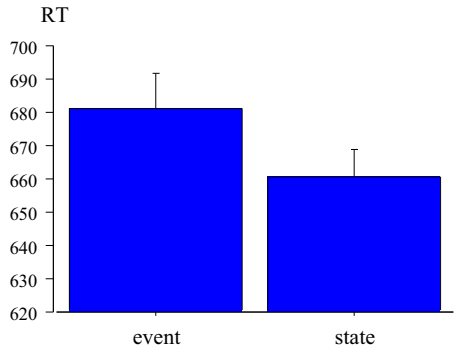


Figure 1: RTs as a function of word type

We interpret these findings as indicating that both imageability and meaning complexity have an effect on RTs. The lack of an interaction between the two main effects indicates that the word type effect does not depend on imageability. Overall these results support the prediction that stative verbs are recognized faster than eventive verbs, consistent with an effect of representational complexity.

Experiment 2: Self-paced Reading study

In this experiment, we asked whether stative verbs are processed faster than eventive verbs when integrated into previous information in the process of sentence comprehension. For this, we measured the reading time to verbs of the sentence stimuli.

Materials 84 sentences containing 42 stative and 42 eventive verbs (plus fillers) were selected. Test verbs and sentences were carefully matched for a number of variables known to affect reading times in context. First, the verbs were pair-wise matched for frequency and word length. Mean log-frequency for both states and events was 3.96 and the mean word length was 6.11 characters for events and 5.82 for states (Collins Cobuild corpus). Comparisons of these variables were not significant.

Second, we pair-wise matched the verbs used in the sentences by their number of syntactic arguments and preferred (most frequent) syntactic frames. This is because Shapiro, Nagel & Levine (1993), Shapiro, Zuriff, & Grimshaw (1987), Rayner and Duffy (1986), McElree (1993) and others have shown that argument structure complexity as well as preferred argument structure can have behavioral consequences in reading times. We used Schulte im Walde's (1999) electronic corpus based on syntactic analysis of the Bank of English to compute number of syntactic arguments and the frequency of syntactic frames. Each selected verb pair had roughly the same number of syntactic frames in which they can occur and for the most frequent frame, the same number of arguments. For example, *love* was matched with *build*, which have similar log frequencies. Proportions of corpus occurrences in different syntactic frames is the following:

love		*build*	
subj:obj	0.67	subj:obj	0.76
subj:to-inf	0.10	subj:obj:obj	0.05
subj:v-ing	0.03	subj:obj:for	0.04

Both verbs occur very frequently in transitive uses, and both verbs have 3 possible argument structures. For verbs like *believe* (equi-biased verbs), which have two equally frequent argument structure (NP, sentence complement), the two most frequent frames were matched for number of arguments. So, *believe* was matched with *report*, which have roughly the same frequent syntactic frames.

Third, test sentences were exactly alike up to the point of the verb, and in some cases, the sentences were completely alike except for the verb. This eliminates the possibility that factors associated with preceding words affect the reading time of the verbs. Examples of matched verb and sentence stimuli are the following:

(11) The young boy bullied his parents. (event)
The young boy adored his parents. (state)
(12)The retired musician built a house. (event)
The retired musician loved his daughter. (state)

Finally, we control for the plausibility relation between the subject NP and the verb. Because certain types of subjects may be more likely to appear with one or the other verb type, we obtained individuals' judgments rating the typicality of the subject-verb relation. We asked 50 students to rate how typical it was for a given subject NP to perform the corresponding action denoted by the verb (Trueswell, et al. 1994). The ratings were compared across word types and did not differ significantly ($t<1$). The mean rating for events and states were 5.51 and 5.60 respectively in a scale of 1 to 7.

Procedure 30 students at University of Maryland read sentences on the computer screen. After each sentence, participants answered a comprehension question. The words of the sentences were presented one-by-one and the participants pressed a key on the keyboard to see each word. Reading time for each word was recorded, though our interest was in the reading time of verbs.

Results comparison of reading times at the verb position revealed a word type effect both across subjects ($F(1, 29)=10.66$, $p=.003$) and items ($F(1,43)=8.9$ $p=.004$). Eventive verbs took longer to process than state verb (about 25 ms. difference). Figure 2 represents the mean reading times (and standard errors) for the nouns preceding the verb, the verb position and the next word.

The results are thus similar to those of the lexical decision task and strongly support our hypothesis of a semantic complexity difference between verb meanings. Semantic complexity in this experiment is clearly independent of syntactic behavior and argument structure complexity.

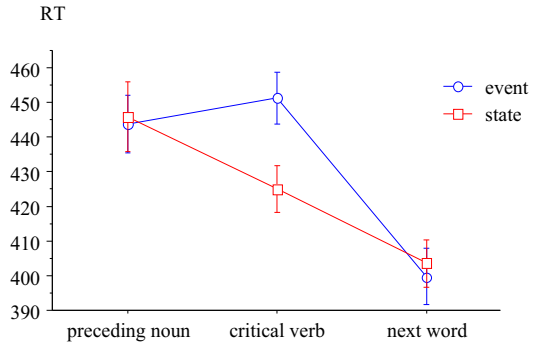


Figure 2: Reading times as function of word type and position

Discussion

The results of these experiments show that eventive verbs take longer to process than stative ones, thus supporting the view that the mental representations associated with them differ in semantic complexity. Computing the verbs' meanings involves differential processing cost, as suggested by the hypothesized complexity of conceptual representations. The distinction between these verb classes is rather abstract and is based on whether the verbs involve a change of states. To our knowledge, this is the first time that these verb classes are shown to have empirical correlates.

Note that the results cannot be attributed to syntactic differences among verbs (frames or number of arguments), as this factor was kept constant. Nor can they be explained as a consequence of expectations generated by the type of arguments with which verbs occur, a factor that has been often manipulated in studies of syntactic ambiguity resolution. In our sentence comprehension experiment, the same subject-arguments were used for both eventive and stative verbs and the plausibility relations between subject-argument and verb were equalized. Also, no such factor was present in the lexical decision task. Thus, the alternative interpretation of the results in which there is a processing difference but not a representational one does not seem plausible, as there is no apparent reason to expect a purely processing difference. We are inclined to conclude, then, that difference in processing cost are due to representational complexity differences between states and events and that such differences may rely on the causal vs. non-causal relations they establish between their participants.

However, the results are neutral as to whether representations such as *x CAUSE(BECOME(y state))* are accurate expressions of the internal representation of the verb meaning. These results only suggest that the representation of eventive verbs is more complex than that of stative ones, regardless of how the complexity is spelled out. Yet if the internal mental representation of verbs includes the type of relations that they establish between participants, it is possible that the complexity difference is due to causal features. In one case, the eventualities denoted involve changes and cause-effect relations (and therefore, agentive participants), while in the other case, they involve mere descriptions of facts. These are important cognitive differences that may somehow be abstracted over verb-specific meaning.

Our results have some important implications for theories of word meanings. As McKoon & Macfarland's (2000) findings, our results challenge the view that verb meanings are atomic and unstructured. On such a view, there is no principled reason to expect these differences in processing unless the lexical entry is allowed to have some sort of internal structure. In this respect, the failure of previous attempts to find lexical complexity effects could be due to the fact that indirect measures of complexity were used (e.g., Fodor et. al 1980) or very small (perhaps undetectable) differences between verb classes were investigated (e.g., Rayner & Duffy, 1986).

Our results also suggest that part of the information encoded in the verb is semantic/conceptual, and somewhat independent from number of participants and syntactic frames. Several psycholinguistic studies have shown that these syntactic variables do influence behavioral measures (e.g., Shapiro et al. 1993). Similarly, Fodor & Lepore (1998) claim that syntactic combinatorial rules can be part of lexical entries, thus increasing their complexity. Our results indicate however, that such syntactic information is not the only relevant factor for complexity effects. Purely semantic properties can also yield processing time differences.

Finally, our results have implication about the exact source of the semantic complexity effects and the overall organization of verb concepts in the lexicon. McRae et al. (1997) and Ferretti et al. (2001) have suggested that events in memory are organized in event schemas and that such schemas arise from the knowledge of their typical agents and patients acquired during learning, i.e., thematic feature knowledge. Thus, it is in principle possible that the verb classes studied here can be distinguished by such thematic features (e.g. features defining agent/patient vs. experiencer/entity structures). However, it is unclear how these features would explain the complexity effects. More importantly, it is unclear whether such features can in fact be distinguished from other aspects of the verb meaning such as the relation between participants established by the verb. Both types of information are inherently related. Our results suggest a level of abstraction or generalization of verb schemas that goes beyond verb-specific knowledge of typical participants and typical situations. Rather, as is the case for nominal concepts, verb concepts

seem to be organized in major event types, in this case distinguished by general causal properties. These types provide the domain of events with a hierarchical organizational structure from verb-specific concepts to abstract verb-general concepts.

References

Bach, E. (1986). The Algebra of Events. *Linguistics and Philosophy, 9*, 5-19.

Balota, D., Ferraro, F., and Connor, L. (1991). On the early influence of meaning in word recognition: a review of the literature. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 187-218). Hillsdale, NJ: Erlbaum.

Balota, D. (1994). Visual word recognition: The journey from features to meaning. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 303-358). San Diego, CA: Academic Press.

Blesdale, F. (1987). Concreteness-dependent associative priming: separate lexical organization for concrete and abstract words. *Journal of experimental psychology: Learning , Memory and Cognition, 13*, 582-594.

Chiarello, C., Shears, C., & Lund, K. (1999). Imageability and distributional typicality measures of nouns and verbs in contemporary English. *Behavioral Research Methods, Instruments, & Computers, 31*(4), 603-637.

Dowty, D. (1979). *Word meaning and Montague grammar*: Kluwer, Reidel, Dordrecht.

Eviatar, Z., Menn, L., & Zaidel, E. (1990). Concreteness: Nouns, verbs, and hemispheres. *Cortex, 26*(4), 611-624.

Ferretti, T. R., McRae, K. and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role. *Journal of Memory & Language, 44*(4), 516-547.

Fodor, J., Fodor, J. A., & Garrett, M. (1975). The psychological unreality of semantic representations. *Linguistic Inquiry, 6*, 515-535.

Fodor, J. A. (1978). Tom Swift and his procedural grandmother. *Cognition, 6*, 229-247.

Fodor, J. A., and Lepore, E. (1998). The emptiness of the Lexicon. *Linguistic Inquiry, 29*(2).

Fodor, J. A., Garrett, M. F., Walker, E. C. T., & Parkes, C. H. (1980). Against definitions. *Cognition, 8*, 263-367.

Gentner, D. (1981). Verb Semantic Structures in Memory for Sentences: Evidence for a Componential Representation. *Cognitive Psychology, 13*, 56-83.

Jackendoff, R. (1990). *Semantic structures*. Cambridge: MIT Press.

James, C. (1975). The role of semantic information in lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 1*(2), 130-136.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

McElree, B. (1993.). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory & Language., 32*(4).

McKoon, G., and Macfarland, T. (2000). Externally and internally caused change of state verbs. *Language, 76*, 833-858.

McRae, K., Ferretti, T., and Amyote L. (1997). Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes, 12*(2/3), 137-176.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*, 255-287.

Parsons, T. (1990). *Events in the semantics of English: a study in subatomic semantics*. Cambridge MA: MIT Press.

Rappaport Hovav, M. and Levin, B. (1998). Building Verb Meanings. In M. B. a. W. Geuder (Ed.), *The Projection of Arguments: Lexical and Compositional Factors* (pp. 97-134). Stanford, CA: CSLI Publications.

Rayner K. and Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*(3), 191-201.

Schulte im Walde, S. (1998). *Automatic Classification of Verbs according to their alternation behavior.* , Stuttgart University, Stuttgart, Germany.

Seidenberg, M., Petersen, A., MacDonald, M., and Plaut, D. . (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*(1), 48-62.

Shapiro, L., Nagel, N., and Levine, B. (1993). Preferences for a Verb's Complements and Their Use in Sentence Processing. *Journal of Memory and Language, 32*, 96-114.

Shapiro, L. Z., E., & Grimshaw J. (1987). Sentence Processing and the mental representations of verbs. *Cognition, 27*, 219-246.

Smith, C. (1991). *The Parameter of Aspect*. Dordrecht: Kluwer.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(5), 1140-1154.

Taylor, B. (1977). Tense and Continuity. *Linguistics & Philosophy, 1*, 199-220.

Trueswell J.C., T., M. K. and Garnsey S. M. (1994). Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *Journal of Memory and Language, 33*, 285-318.

Trueswell, J. C., Tanenhaus, M.K., and Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 528-553.

Vendler, Z. (1967). *Linguistics and philosophy*. Ithaca, NY: Cornell University Press.

Verkuyl, H. (1993). *A theory of aspectuality*. UK: Cambridge University Press.

Interactive Knowledge Acquisition Tools: A Tutoring Perspective

Yolanda Gil (gil@isi.edu) and Jihie Kim (jihie@isi.edu)
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292, USA

Abstract

This paper argues that interactive knowledge acquisition tools would benefit from a tighter and more thorough incorporation of tutoring and learning principles. Current systems learn from users in a very passive and disengaged manner, and could be designed to incorporate the proactive capabilities that one expects of a good student. This paper points out what tutoring and learning principles have been used to date in the acquisition literature, though unintentionally and implicitly. We discuss how a more thorough and explicit representation of these principles would help improve enormously how computers learn from users.

Introduction

Computers have long been considered an invaluable tool for education. Intelligent tutoring systems and other kinds of educational software show how people can acquire knowledge about diverse topics by interacting with a computer. Given our reliance in computers for the future of education, we need to realize that an important component of the human education revolution is the computer education revolution: anyone should be able to teach computers on topics that are of value, so that anyone can learn about those topics from computers. So an important question is: how will computers acquire knowledge? In most cases knowledge is entered by hand by software or knowledge engineers, as is often done in intelligent tutoring systems (Forbus & Feltovich, 2001). This limits severely the utility of the tools, as it would be more desirable that the people with expertise in the domain at hand would be the knowledge providers. Knowledge can also be extracted from text (Cowie & Lehnert, 1996), although given the error rates of state of the art systems and the kinds of knowledge they acquire (mostly instance-level information) it will take many years for these techniques to be of practical use to build an accurate body of knowledge about a topic domain. Another possibility is to use interactive knowledge acquisition tools that help users enter knowledge. In recent years these systems have shown that end users with no background in computer science or knowledge representation were able to enter sizeable amounts of knowledge (Kim & Gil, 1999; Eriksson et al, 1995; Clark et al, 2001).

Although interactive knowledge acquisition tools enable end users to enter knowledge, users remain largely responsible for the acquisition process, both the teaching side and the learning side. These tools are quite passive in terms of formulating or pursuing learning goals, keeping track of the flow of a lesson, and generally assess how much they are learning and how useful that knowledge is. At the same time, users are not necessarily skilled teachers by nature, so being in a position to teach a computer is already a challenge for them. Interactive acquisition tools need to be more effective and helpful to users, perhaps by incorporating some of the skills that are expected of good students. And, as good students do, they should also be able to cope with an inexperienced teacher (which their users are likely to be) and still learn from the experience by bringing to bear knowledge about how a good teacher typically goes about a lesson. An interactive acquisition tool could then be viewed as a tool to support augmented cognition, since it would supplement the user's limitations as a teacher and knowledge engineer.

The contributions of this work are twofold. First, we point out how existing knowledge acquisition tools use techniques that are related to widely used tutoring and learning principles. Second, we identify areas that the acquisition tools developed to date have neglected, and suggest promising areas of research based on our findings. This would result in a new generation of acquisition tools that are not only better students but also more helpful to the teacher (the user).

The paper begins with a short introduction and background on interactive knowledge acquisition tools. We then discuss several tutoring and learning principles that we have drawn from the educational literature and that seem useful to support the interactive acquisition process. Next, we show how some existing acquisition tools use techniques that are related to these principles in some aspects of their functionality. We finalize with a discussion of promising directions that we see in designing acquisition tools that incorporate tutoring and learning principles more thoroughly.

Acquisition Tool	*Highlights*
CHIMAERA (McGuinness et al., 2000)	To acquire concepts, relations, and instances. Diagnoses faulty definitions.
EXPECT (Blythe et al., 2001)	To acquire problem solving knowledge. Exploits dialogue scripts, knowledge interdependency models, and background knowledge.
INSTRUCTO-SOAR (Huffman & Laird 1995)	To acquire task models for Soar.
KSSn (Gaines & Shaw, 1993)	To acquire concepts, rules, and data. Based on personal construct psychology.
PROTOS (Porter et al., 1990)	Users specify cases, tool explains their classification.
SALT (Marcus, 1988)	To acquire constraints and fixes for its underlying engine for configuration design.
SEEK2 (Ginsberg, 1985)	To acquire rules. Uses verification and validation techniques.
SHAKEN (Clark et al., 2001)	To acquire process models. Loosely based on concept maps.
TAQL (Yost, 1993)	To acquire SOAR rules. Based on Problem Space Computational Model.
TEIREISIAS (Davis, 1979)	To acquire rules. Exploits context, derived rule models, and heuristics.

Table 1: Some Interactive Knowledge Acquisition Tools.

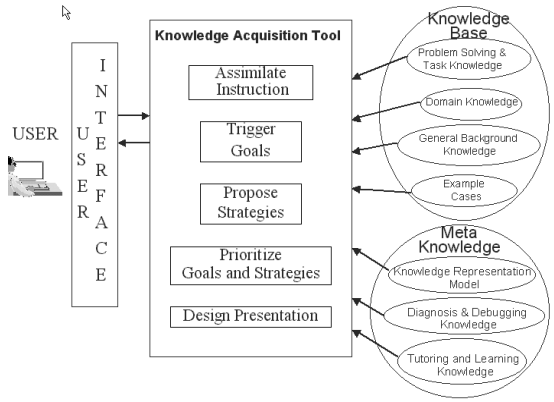


Figure 1: A Modular View of Interactive Knowledge Acquisition Tools.

Interactive Knowledge Acquisition Tools

The interactive knowledge acquisition tools summarized in Table 1 illustrate different approaches that researchers have undertaken over the years and are representative of the literature. A brief description of the tools are in Gil and Kim (2002). The techniques used range from cognitive theories of expertise and learning, case-based reasoning and analogy, non-monotonic theory revision, induction and machine learning, knowledge engineering approaches, analysis of knowledge interdependencies and buggy knowledge, and agent-based interaction.

Figure 1 shows a diagrammatic view of typical components used in various tools. The functionality of an interactive knowledge acquisition tool can be described along five dimensions, which we will use for reference later in our analysis:

- **Assimilate instruction:** Given a user's instruction, the system makes the necessary additions or changes to the knowledge base and updates any other internal structures. Instruction may be given as an example (PROTOS, SEEK2), a natural language statement (INSTRUCTO-SOAR), a descriptive piece of knowledge (CHIMAERA, EXPECT, SALT, TAQL), or a graphical rendering (KSSn, SHAKEN).

- **Trigger goals:** The system analyzes its knowledge and generates learning goals of what knowledge it still needs to acquire. Many tools focus on detecting inconsistencies or gaps in the knowledge base, which generate the goals to fix them or seek the information missing (CHIMAERA, EXPECT, SEEK2, TAQL, TEIREISIAS).

- **Propose strategies:** The tool can generate possible strategies that the user could follow in achieving the learning goals. It can also generate predictions of what strategy the user is more likely to pursue, or what answers the user is more likely to give to the user's questions (TEIREISIAS). This is often done by analyzing existing knowledge. Planning strategies are often used to make suggestions to the user in terms of what to do to achieve the active learning goals, which will make the acquisition process more efficient (EXPECT).

- **Prioritize goals and strategies:** An acquisition tool can help users further if it is able to organize and prioritize the active learning goals and candidate strategies, so that it can make more focused suggestions to the user. Sometimes these are organized by the type of knowledge sought (SALT), or by the type of goal being pursued or error being fixed (EXPECT).

- **Design presentation:** The tool can make decisions about what to bring to the attention of the user at each point in time to help him or her decide what to do next. There are many possibilities, and the tool can take into account the user's situation (user modeling), the stage of the process (initial stage versus final testing), and the content of the current knowledge base. The tool may present the user with a single question (INSTRUCTO-SOAR) or give the user a choice in the form of an agenda containing multiple items (CHIMAERA, EXPECT, SALT). The tool can suggest a specific strategy, anticipate the user's answer and ask for confirmation, or simply present the user with multiple possible strategies and suggestions (EXPECT). Other tools leave it up to the user to figure out what to do and simply make all possible options available to them (KSSn,

SHAKEN) The tool may simply ask the user to review an explanation (PROTOS), check some aspect of the knowledge (KSSn), or confirm a hypothesis (TEIREISIAS).

Another useful view of interactive acquisition tools is in terms of the kinds of knowledge and meta-knowledge that they bring to bear in order to support the user, also illustrated in Figure1. This includes:

- *General problem solving and task knowledge*: General inference structures are used to determine the role that domain-specific knowledge plays in problem solving, as is done in role-limiting approaches to knowledge acquisition (Marcus & McDermott, 1989) (e.g., SALT, TAQL).

- *Prior domain knowledge*: The initial knowledge base may contain terms that are specific to the domain at hand and that can be used to define new terms and tasks (e.g., EXPECT, INSTRUCTO-SOAR).

- *General background knowledge*: The initial knowledge base may include high level theories and ontologies that capture general knowledge, such as time, physical objects, etc. (e.g., SHAKEN).

- *Example cases*: Sample situations, test cases, and problem solving episodes can help ground abstract knowledge (e.g., INSTRUCTO-SOAR, PROTOS, SEEK2, SHAKEN, TEIREISIAS).

- *Underlying knowledge representation*: Models of the underlying knowledge representation will determine how users need to formulate new knowledge (e.g., CHIMAERA, KSSn, SEEK2, TAQL, TEIREISIAS).

- *Diagnosis and debugging knowledge*: Typical diagnosis skills are useful in order to detect errors and potential problems in the knowledge base. Effective debugging strategies can be incorporated to make suggestions to the user about how to fix the errors and problems found (e.g., CHIMAERA, EXPECT, TEIREISIAS).

One source of meta-knowledge that has not received attention is effective tutoring and learning techniques. By exploiting meta-knowledge about how to learn and how to teach, acquisition tools will become more proactive learners and will be able to help users teach them more effectively. Current tools are often too passive, and place on the user the majority of the burden of the acquisition process. Our goal is to understand whether and how knowledge acquisition tools can exploit knowledge about tutoring and learning.

Tutoring and Learning Principles in Existing Interactive Acquisition Tools

We analyzed the tutoring and educational literature to compile tutoring and learning principles that humans and computers exploit to make teaching and learning more effective. We compiled fifteen principles that could be of immediate use in our work, and that are described in detail in (Kim & Gil, 2002) including detailed references to the tutoring literature that are omitted in this paper because of space limitations.

We noticed that many of these principles could be related to the techniques used in existing acquisition tools. Yet, the tutoring literature is seldom mentioned in knowledge acquisition work. In this section, we describe our views on how acquisition techniques can be expressed in terms of these tutoring and learning principles. Table 2 summarizes our analysis, indicating the particular functionality (as outlined in Figure 1) where the principle was applied in specific acquisition tools.

Introduce lesson topics and goals

Teachers often start off by introducing the topics and goals of the lesson. There is no notion in acquisition tools that there is a lesson being started or ended, since at any point users can choose to enter knowledge about any topic. EXPECT allows users to specify the top-level tasks that the system should be able to solve with the new knowledge, which can be viewed as a statement of the goals for that acquisition session. SEEK2 has a suite of test cases that the system should be able to solve after the lesson, and that could be viewed as a statement of the goals of the lesson.

Use topics of the lesson as a guide

It is useful for students and tutors to ensure that what is being learned has some connection or relevance to the topics of the lesson. EXPECT uses the specified top-level tasks to check that any new knowledge specified solves some of their subtask, and if not it notifies the user and suggests how it could play a role in solving the tasks. SEEK2 uses the suite of test cases to detect errors, which then drive the dialogue with the user towards fixing them. SALT can be viewed as having an implicit (and very high level) topic for all sessions, namely to acquire knowledge for configuration design problems. SALT's interface asked users to specify only three kinds of knowledge (parameters, constraints, and fixes) that are relevant to those types of problems.

Subsumption to existing cognitive structure

Learning about a new topic involves relating the new knowledge to what is already known, for example by

Tutoring/Learning principle	Assimilate Instruction	Trigger Goals	Propose Strategies	Prioritize Goals and Strategies	Design Presentation
Introduce lesson topics and goals		EXPECT, SEEK2			
Use topics of the lesson as a guide	SALT	SEEK2	EXPECT		SALT
Subsumption to existing cognitive structure	PROTOS		TEIREISIAS		PROTOS, SALT
Immediate feedback	PROTOS	INSTRUCTO-SOAR	TEIREISIAS		EXPECT
Generate educated guesses		TEIREISIAS	EXPECT		
Keep on track					
Indicate lack of understanding	INSTRUCTO-SOAR				INSTRUCTO-SOAR
Detect and fix "buggy" knowledge	TAQL	EXPECT CHIMAERA			PROTOS, SEEK2 TEIREISIAS
Learn deep models					
Learn domain language					
Keep track of correct answers		SEEK2			
Prioritize learning tasks				EXPECT	
Summarize what was learned					
Assess learned knowledge		KSSn			

Table 2: Tutoring and learning principles used in acquisition tools.

checking inconsistencies, drawing analogies, or deriving generalizations. PROTOS took a new example case provided by the user, and indexes it into one of several classes (or categories) of examples. It also presented the user with an explanation of the classification of the new example to show how the new knowledge was incorporated into the existing structures. TEIREISIAS created generalized rule models from its rule base, and used them to propose to the user additional conditions to newly defined rules. The interface and presentation of SALT was always based on the kinds of knowledge needed for configuration design.

Immediate feedback

Educational systems often provide immediate feedback, as studies show that it is more effective than feedback received after a delay. PROTOS provided immediate feedback as a new case was assimilated by showing the user an explanation of its classification in the knowledge base. INSTRUCTO-SOAR generated clarification and follow-up questions for the user immediately after an instruction was given. TEIREISIAS proposed amendments to rules as soon as the user defined them. EXPECT analyzes the knowledge base after each user action and shows immediately an agenda of errors to resolve and tasks to do.

Generate educated guesses

Students often show their understanding by finishing a tutor's utterance, and tutors often invite students to guess as a way to assess and correct the student's knowledge. TEIREISIAS maps newly entered rules to rule models and proposes corrections based on how it expects a rule to follow the patterns of other rules in that model. EXPECT generates suggestions to a user about how to fix specific problems by making educated guesses about the context of the problem (related domain knowledge, past problem solving states, etc.)

Keep on track

Tutors need to keep track of the lesson and bring back issues that had to be dropped while engaging in clarifications or other side dialogues. Acquisition tools do not keep track of the history and status of the dialogue. Users have free range on what aspects of the knowledge base to extend, what parts of the tool to invoke, and what They can move freely from topic to topic and back and forth, or discontinue teaching about a topic at any point without notifying termination. Current acquisition tools would never even notice that the user is deviating from a topic in any of these situations.

Indicate lack of understanding

Students often volunteer an indication of their lack of understanding, but tutors also will point out the specific aspects introduced in a lesson that the student needs to understand. INSTRUCTO-SOAR detects missing aspects of a task description specified by a user and generates follow up questions. EXPECT and CHIMAERA detect undefined terms that will be used to guide future dialogue with the user to define them.

Detect and fix "buggy" knowledge

Many tutoring systems are aimed to diagnose and fix student's "buggy" knowledge, often by asking questions and checking the student's answers. TAQL analyzes the knowledge specified by the user and points out errors based on static analysis. CHIMAERA and EXPECT detect errors in the knowledge entered that need to be fixed by the user. PROTOS, SEEK2, and TEIREISIAS show explanations or traces to users so they can detect errors in the system's reasoning.

Learn deep models

Students should learn deep conceptual models instead of superficial ones. Knowledge acquisition tools do not have any basis to evaluate or pursue depth in their knowledge base, though this is a long

recognized shortcoming of knowledge-based systems. To date, these systems are at the mercy of the user's intention and of their implementation of any depth in the models.

Learn domain language

Students are expected to be describe their knowledge in terms that are suitable for the domain at hand. Acquisition tools do not help users specify how to describe knowledge in domain terms and how the terminology used depends on the context of the scenario at hand. Knowledge bases are annotated with some lexical information, but acquiring this kind of knowledge has not been a focus of knowledge base development.

Keep track of correct answers

Instructional tools keep track of the questions that the student is able to answer correctly as well as those answered incorrectly, which drives further interactions with the student. SEEK2 keeps track of whether the test cases are answered correctly, and alerts the user when a change to a rule causes a case to be solved incorrectly.

Prioritize learning tasks

Tutoring systems often handle multiple sub-tasks using priority rules that look at the duration and type of task, for example focusing on fixing errors before turning to omissions. EXPECT organizes errors and other problems in the knowledge base based on their type and the amount of help it can provide (e.g., if it has narrowed down the options that the user can take to resolve them).

Limit the nesting of sub-lessons

Tutoring dialogue is sometimes controlled by limiting the amount of subdialogues, which helps the student keep track of the lesson topics.

Summarize what was learned

Many educational systems will summarize to the student the main highlights at the end of the lesson, especially if the student was given hints during the lesson. Acquisition tools do not summarize what they have learned.

Assess learned knowledge

Some instructional tools isolate weaknesses in the student's knowledge and propose further lessons on those areas, some students also volunteer their assessment of how well they understand certain topics. KSSn uses clustering techniques to suggest aspects of the model that users could detail further. Other acquisition tools do not perform this kind of analysis. Users often have to put the knowledge base through a performance system that exercises it in order to be able to assess if the knowledge was learned appropriately.

Discussion

Acquisition tools have used techniques that can be cast in terms of tutoring and learning principles found in educational software research. These principles are implicit in the design of the tool, and they influence their interaction with the user to the degree that they are implemented in the underlying code. Having these principles represented explicitly and declaratively would enable acquisition tools to reason in terms of the teaching and learning process, and their interaction with the user would be dynamically generated given the situation at hand. A declarative representation of meta-knowledge about their learning state, goals, and possible strategies could turn interactive acquisition tools into more proficient and proactive learners.

The principles have only been used in some aspects of the functionality of acquisition tools, and are exhibited by some but not all the tools. The sparseness of the matrix in Table 2 points to many opportunities for future work in incorporating these principles. By having declarative representations of their learning state, goals, and possible strategies, interactive acquisition tools could more easily incorporate these principles throughout the acquisition process and the five functions shown in the table.

Acquisition interfaces should be able to structure the dialogue with the user in tutoring terms. The should organize the dialogue based on lesson topics and sub-topics, be aware of the start and the end of each and generally keep the user on track and delaying termination until the goals of the lesson are satisfied. Acquisition tools should exploit the topics of the lesson throughout the acquisition process, for example to narrow down the prior knowledge that is relevant to that portion of the dialogue and consequently narrowing down the proposed strategies and customizing the presentation of information back to the user. By keeping track of the interactions with the user, the topic of the dialogue at each point in time, and the termination of sub-topics, acquisition tools would be able to manage their participation in the dialogue better and relieve the users from having to remember and keep track of what is going on. They could exploit this information in generating goals by detecting areas where a topic is still unfinished, plan and prioritize more relevant strategies that exploit the context of the currently open topics, and help users view progress and termination.

Acquisition tools should be able to expose and assess the knowledge acquired so far, allowing the user to understand what the system has assimilated and showing the user as well what areas the system thinks need to be further improved. Currently, knowledge-based systems will answer any question they are asked, regardless of the quality of the knowledge used to answer it. It would be very useful for these systems to convey whether they are confident on the answer. This would also help users identify

further areas of improvement for future acquisition sessions.

We are pursuing these ideas in our current work, implementing a front-end dialogue management system that represents and uses tutoring and learning principles to guide knowledge acquisition.

Acknowledgments

This research was funded by the DARPA Rapid Knowledge Formation (RKF) program with award number N66001-00-C-8018. We would like to thank Ken Forbus, Lewis Johnson, Jeff Rickel, Paul Rosenbloom, David Traum, and Jim Blythe for their insightful comments on this work.

Bareiss, R., & Porter, B., & Holte, R. (1990). Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence Journal* 45(1-2):229–264.

Blythe, J., & Kim, J., & Ramachandran, S., & and Gil, Y. (2001). An integrated environment for knowledge acquisition. *Proceedings of the Intelligent User Interfaces conference (IUI-2001).*

Clark, P., & Thompson, J., & Barker, K., & Porter, B., & Chaudhri, V., & Rodriguez, A., & Thomere, J., & Mishra, S., & Gil, Y., & Hayes, P., & Reichherzer, T. (2001). Knowledge entry as the graphical assembly of components. *Proceedings of First International Conference on Knowledge Capture (K-CAP-2001).*

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM* 39(1):80–91.

Davis, R. (1979). Interactive transfer of expertise: Acquisition of new inference rules. *Artificial Intelligence* 12:121–157.

Eriksson, H., &shahar, Y., & Tu, S. W., & Puerta, A. R., & Musen, M. (1995). Task modeling with reusable problem-solving methods. *Artificial Intelligence* 79:293–326.

Forbus, K., & Feltovich, P., eds. (2001). *Smart Machines in Education.* AAAI press.

Gaines, B. R., & Shaw, M. (1993). Knowledge acquisition tools based on personal construct psychology. *The Knowledge Engineering Review* 8(1):49–85.

Gil, Y., and Kim, J. (2002). Interactive Knowledge Acquisition Tools: A Tutoring Perspective. http://www.isi.edu/expect/papers/Interactive-KA-Tools-gil-kim-02.pdf (internal project report).

Ginsberg, A., & Weiss, S., & Politakis, P. (1985). SEEK2: A generalized approach to automatic knowledge base refinement. *Proceedings of IJCAI-85.*

Huffman, S. B., & Laird, J. E. (1995). Flexibly instructable agents. *Journal of Artificial Intelligence Research* 3:271–324.

Kim, J., and Gil, Y. (1999). Deriving expectations to guide knowledge base creation. *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 235–241.

Kim, J., and Gil, Y. (2002). Deriving acquisition principles from tutoring principles. *Proceedings of the Intelligent Tutoring Systems Conference (ITS-2002)*, Biarritz, France, June 2002.

Marcus, S., and McDermott, J. (1989). SALT: A knowledge acquisition language for propose-and-revise systems. *Artificial Intelligence* 39(1):1–37.

McGuinness, D. L.; Fikes, R.; Rice, J.; and Wilde, S. 2000. An environment for merging and testing large ontologies. *Proceedings of KR-2000.*

Newell, A., & Yost, G., & Laird, J., & Rosenbloom, P., & Altmann, E. (1991). Formulating the problem-space computational model. Rashid, R. F., ed., *CMU Computer Science: A 25th Anniversary Commemorative.* ACM Press.

Novak, J., ed. 1998. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations.* Lawrence Erlbaum.

Yost, G. R. 1993. Knowledge acquisition in Soar. *IEEE Expert* 8(3):26–34.

Taking Care of the Linguistic Features of Extraversion

Alastair J. Gill (agill@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We study how Extraversion or Introversion influences people's language production. A corpus of e-mail texts was gathered from individuals categorised via Eysenck's EPQ-R personality test. One experiment analysed the corpus using existing content analysis tools, and found relatively weak effects of Extraversion. A second experiment used more sensitive bigram-based techniques from statistical natural language processing to replicate earlier findings, and uncover novel patterns of behaviour.

Introduction

Casual acquaintance with Extraverts[1] and Introverts suggests that the former talk a lot more than the latter. But apart from this intuitive difference, how does this personality dimension influence an individual's language production? Before addressing this question, we need to clarify what we mean by Extraversion, and its relevance to cognitive science.

A typical Extravert tends to be sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert is quiet, retiring, reserved, plans ahead, and dislikes excitement (Eysenck and Eysenck, 1991).

The personality trait of Extraversion—and the complementary Introversion—is one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993). Beyond it, there is greater controversy.

For instance, Eysenck's EPQ-R personality test reflects a personality model which incorporates just two further dimensions: Neuroticism, which is mainly characterised by susceptibility to anxiety; and Psychoticism, which is more complicated, but generally related to aggression and individuality. By contrast, the NEO-PI-R model incorporates five factors (Costa and McCrae, 1992). As well as Extraversion and Neuroticism, there are Conscientiousness, Agreeableness and Openness. It is generally agreed that these relate to Psychoticism, but exactly how is still the subject of debate (cf. Matthews and Deary, 1998).

Extraversion, and its linguistic consequences—if there are any—is relevant to cognitive research for at least two reasons. First, there is considerable evidence that this personality dimension is related to preferred learning styles and educational achievement, via speed of exam completion, memory retrieval and recall tasks, creativity, mathematical ability, self monitoring and communication ability (Jonassen and Grabowski, 1993). Secondly, there is evidence that computer users attribute personality to interfaces, and respond to it in robust ways (eg. Nass, Moon, Fogg, and Reeves, 1995; Isbister and Nass, 2000). Even in a text-only environment, Extraverts apparently prefer interfaces presenting information using language associated with Extravert traits; Introverts prefer Introverted interfaces. An interface with a matching personality is judged more positively, and rated as more attractive, credible and informative (Nass *et al.*, 1995).

So the personality dimension has some validity, and appears relevant to the diagnosis and projection of personality in human-computer communication, and in computer-based learning. But how does Extraversion influence an individual's language production? In addressing this question, we first outline some hypotheses from the literature, before describing our collection of a controlled corpus of language, and our analysis of it. We then report the results—some unsurprising, others unexpected—and discuss some of their implications.

Previous hypotheses

Work on textual personality within the "Computers Are Social Actors" paradigm has taken the expressive hallmarks of Extraversion or *dominance* (one facet of the dimension) to be confidence, as shown by an avoidance of hedge-expressions such as *perhaps* and *maybe* (Nass *et al.*, 1995), and is related to the empirical work of Bradac and Mulac (1984) on perceptions of powerful and powerless speech.

From an intuitive perspective, Extraverts are described as individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts mo-

[1]The spelling of *Extravert* follows Eysenck, because this paper employs his EPQ-R as the measure of personality, but this does not represent a commitment to a specifically Eysenckian theory of personality.

nopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (cf. Carment, Miles, and Cervin, 1965). With reference primarily to speech, Furnham (1990) has proposed that Extravert language is less formal, has a more restricted code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely (see also Dewaele and Furnham, 1999, for a review of speech and writing studies).

Text analysis approaches have found that transcribed texts rated as belonging to the *warm* facet of Extraversion used fewer negative emotion words and unique words, and more present tense verbs, with *dominant* texts using fewer unique words, positive emotion words and self referents (Berry, Pennebaker, Mueller, and Hiller, 1997). Finally, study of the texts *written* by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclusive words, while using more social and positive emotion words (Pennebaker and King, 1999).

Data Collection

The approach to data collection follows Pennebaker and King (1999). Written texts were collected from 105 University students or recent graduates (37 males, 68 females; mean age = 24.3 years; SD = 4.6; all native English speakers). An introductory e-mail explained the experiment, and pointed subjects to the relevant web-page. After the completion of an online demographic questionnaire and a version of the Eysenck Personality Questionnaire (Revised short form; Eysenck, Eysenck, and Barrett, 1985) (mean score for E = 7.91, SD = 3.25; normative score = 7.42 (male), 7.60 (female)), subjects were asked to compose two e-mails *to a good friend whom they hadn't seen for quite some time*, the style of which is considered to be close to oral communication (Bälter, 1998). One message concerned their activities in the past week; the other discussed their plans for the next week. Subjects were advised to spend around ten minutes per message, composed online and submitted using an HTML form. It was highlighted that responses would be treated in confidence and that subjects could remain anonymous. No payment was made for participation, and integrity of responses was monitored by reading through the transcripts. One additional submission was rejected as not being serious; the resulting corpus contained 210 texts and 65,000 words.

Experiment 1: Dictionary techniques

LIWC and MRC Methods

LIWC Each respondent's texts were individually processed using the LIWC text analysis program (Pennebaker and Francis, 1999). Items were selected for principle components analysis using the same criteria as Pennebaker and King (1999), namely reliability, topic independence, independence from other variables, and a mean minimum usage of 1%. The validity of the current data was shown using varimax rotation to derive four factors which essentially replicate their prior findings. There was minor variation in some factor loadings, which we attribute to differences in the writing tasks. See Gill and Oberlander (prep) for a fuller discussion.

Table 1: Summary of E Score and LIWC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Numbers	−.21		
	Word Count	.20	**.08**	**.0144**

By correlating their resultant LIWC factors with personality dimensions, Pennebaker and King's results suggest broad style preferences for Extraverts. But this does not identify the relative importance of their categories for identifying text as Extravert.

Thus, to identify which LIWC variables best help identify an author's personality, stepwise linear multiple regression was performed. The variables entered were those which showed at least a small correlation with the personality type—with a significance of $p < .1$—and which satisfied the criteria for inclusion in the previous principle components analysis. However, since requiring variables to have a mean usage of 1% per essay for inclusion in the analysis did not leave any LIWC variables in the regression equation for Extraversion, this criterion was ignored for the results presented below. (Interestingly, by contrast, even with the application of this criterion, Psychoticism and Neuroticism both had several strongly significant LIWC predictor variables.)

MRC In addition to the LIWC-based tests, multiple regression analysis was also performed on psycholinguistic properties of the texts, derived from the MRC Psycholinguistic Database (Coltheart, 1981). Texts were first tagged for Parts of Speech,[2] and each word-POS pair was then looked up in the database. If the word and POS tag matched a pair in the database, psycholinguistic data was returned for that word. When all the words in the text had been processed, mean scores were calculated for categories such as verbal frequency, written frequency, concreteness, age of acquisition, along with additional global information, such as the percentage of a text's words which were captured by the database. As with the LIWC regression, variables showing a correlation with the personality type with a significance of $p < .1$ were entered in to the equation.

[2]Using the MXPOST tagger (Ratnaparkhi, 1996).

Table 2: Summary of E Score and MRC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Mean Concreteness	$-.21$	**.05**	**.0278**

Results

The multiple regression analysis of the LIWC variables (Table 1) shows that a greater overall word count for a text ($\beta = .20$), and the occurrence of fewer references to numbers within that text ($\beta = -.21$), indicate Extraversion ($p < .05$). So, Extraverts *do* appear to type more than Introverts, mirroring earlier results on speech (Carment *et al.*, 1965), with the avoidance of numbers embodying a 'looser', less precise use of language (Furnham, 1990). However, the variance accounted for by these variables is relatively low at 8%. In comparable analyses, both Psychoticism and Neuroticism regression equations explain variance greater than 10%.

Similarly, with the MRC Psycholinguistic analysis (Table 2), only the novel finding of a general lowering of a text's concreteness of vocabulary ($\beta = -.21$, $p < .05$) was seen to explain 5% of variance in Extraversion. Again, equations for Psychoticism and Neuroticism explained more than 10% of variance.

Discussion

In both of the dictionary-based analyses of the texts, rather few features appeared to distinguish Extravert/Introvert texts, especially when compared to the numerous LIWC and MRC features which associated with Psychoticism and Neuroticism traits.

How could this be? At least two explanations are possible. First, the LIWC dictionary is a subjectively constructed analysis tool. It is based on judgements by health psychologists of texts written by distressed individuals for therapeutic purposes (Pennebaker and Francis, 1999). For its original purposes, this is a strength; but it also imposes a *top down* limitation on LIWC's functioning. Given this therapeutic origin, it is tempting to suggest that the linguistic features associated with the personality traits of Psychoticism and Neuroticism were more important or relevant to the distressed individuals producing the texts—and that is why these features are better represented in LIWC's dictionary.

The MRC database is also fitted to its specific purposes—for example, matching psycholinguistic stimuli—but this again imposes constraints which might prove artificial when it is applied to a different area of investigation.

Secondly, both dictionaries necessarily operate using strings corresponding to individual words, subsequently classifying them in a predefined way. Neither takes into account the context of a word. Thus it may be that for Psychoticism and Neuroticism the choice of word, or some property of the word is informative—but for Extraverts, it may be that word order or collocations are more relevant.

Experiment 2: NLP techniques

Therefore, we recruit *bottom up* statistical text analysis techniques from corpus linguistics. Specifically, bigram analysis calculates the probability of pairs of adjacent terms, or bigrams, occurring together in that order in a given text. To determine the significance of a bigram's occurrence, a statistic—log likelihood—is calculated, taking into account all the other instances of each element in the bigram pair, and the other words with which they appear.

Since bigrams can be used to calculate the probabilistic space in which language occurs, they have been put to a variety of uses (Collins, 1996; Pedersen, 2001). However, this study uses them simply as an advancement on the classified unigram (that is, single-word) analysis in Experiment 1. Because bigrams contain information about the interconnection and dependencies of words, this second analysis retains some of the contextual information of language use. Equally importantly, since bigrams are not classified subjectively, they provide a form of analysis that is bottom-up, rather than top-down.

Method

The original corpus of texts was divided by degree of Extraversion by selecting respondents whose E score was greater or less than 1 s.d. of the mean (cf. Dewaele and Pavlenko, 2002), with the 21 High Extravert authors scoring more than 11, and the 17 Low Extravert authors scoring less than 5.

Bigrams were calculated for the resulting Extravert and Introvert subcorpora; the former contained over 12,000 words; the latter around 8,000.Bigram profiles were generated for each corpus and their co-occurence significance in the current texts ranked by log-likelihood statistic ($-2 \log \lambda$),[3] since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993). Rankings for each group are based on the top 50 bigrams with frequency of $N \geq 2$, and a significance of $p < .001$. Relative frequency ratios (Damerau, 1993) were then calculated for bigrams that were common to both the subcorpora, and a Spearman Rank correlation was also performed on these bigrams.

Results

Spearman Rank Correlation

The correlation coefficient score of .53 indicates that Extravert and Introvert use of the shared bigrams is significantly correlated at the $p < .005$ (one-tailed, N=28) level, and they are therefore not distinct.

[3]Ted Pederson's bigram software is available from: http://www.d.umn.edu/~tpederse/code.html.

Table 3: Shared Extravert and Introvert bigrams.

Bigram	Extr Cnt	Intr Cnt	Extr Ratio	Intr Ratio	Rel.F Ratio
looking forward	15	4	0.0011	0.0005	2.49
it was	46	22	0.0034	0.0025	1.39
next week	24	12	0.0018	0.0013	1.33
a bit	29	15	0.0022	0.0017	1.28
up with	19	10	0.0014	0.0011	1.26
! !	45	24	0.0033	0.0027	1.24
will be	24	13	0.0018	0.0015	1.22
i was	33	18	0.0025	0.0020	1.22
at the	27	16	0.0020	0.0018	1.12
to see	32	19	0.0024	0.0021	1.12
which is	15	9	0.0011	0.0010	1.11
for a	34	21	0.0025	0.0024	1.07
i have	44	29	0.0033	0.0032	1.01
to get	34	23	0.0025	0.0026	0.98
. i	99	69	0.0074	0.0077	0.95
on friday	11	8	0.0008	0.0009	0.91
, and	48	36	0.0036	0.0040	0.88
and then	23	19	0.0017	0.0021	0.80
in the	41	34	0.0031	0.0038	0.80
apart from	6	5	0.0005	0.0006	0.80
i am	33	28	0.0025	0.0031	0.78
i think	16	14	0.0012	0.0016	0.76
, but	35	31	0.0026	0.0035	0.75
a lot	10	9	0.0007	0.0010	0.74
going to	36	33	0.0027	0.0037	0.72
a few	12	11	0.0009	0.0012	0.72
to do	23	23	0.0017	0.0026	0.66
i've been	9	12	0.0007	0.0013	0.50

However, further analysis showed Extraverts to be more distinguishable from Ambiverts or Introverts.[4]

Extraverts versus Introverts

The results of the bigram analysis include: bigrams which occurred in both the Extravert and Introvert corpora (Table 3); bigrams which were found uniquely in the Extravert corpus (Table 4); and those found only in the Introvert corpus (Table 5). The shared bigrams are ordered by their relative frequency, with the highest ratios above 1.0 showing the strongest association with Extravert authors, and the smallest ratios less than 1.0 indicating a preference on the part of more Introverted authors (the breakpoint has been indicated by a separating rule). Features which are unique to each subcorpus group can be considered the most distinctive of authorial personality. For current purposes, we divide the features into eight groupings.

Surface Realisation Features These gross features are perhaps the most intuitive in their representation of the Extraverts or Introverts. For example, [<END> *hi*], the <END> (end–of–file marker) followed by *hi*, was unique to Extravert texts; and since the <END> marker separates concatenated files in the corpus, here we have a tendency towards message-initial *hi*. By contrast the more formal [<END> *hello*] was found solely in Introvert texts. Use of punctuation also differs between the two groups, with Extraverts preferring multiple exclamation marks [*! !*], and solely using multiple full stops [. .] as in the elliptical (. . .), again a feature of informal style, and 'looser' use of language.

[4]When comparing the groups High E ($\geq$1s.d.), Mid E ($< \pm$1s.d.) and Low E ($\leq -$1s.d.) (all P and N $< \pm$1s.d.) it was found that Low E and Mid E correlate very significantly ($p < .005$; $\rho = .67; N = 19$), whilst High E and Mid E do not significantly correlate at the $p < .05$ level ($\rho = .32; N = 24$).

Table 4: Bigrams unique to Extravert corpus.

Bigram	Rank	$-2 \log \lambda$	Count	Ratio
. .	8	183.48	152	0.0113
of the	33	79.47	40	0.0030
, which	20	100.89	25	0.0019
had a	16	115.60	22	0.0016
which was	24	95.69	19	0.0014
new year	7	192.22	18	0.0013
got a	45	66.65	17	0.0013
a good	46	64.45	16	0.0012
forward to	26	94.76	15	0.0011
need to	28	89.99	15	0.0011
i'll be	22	98.70	14	0.0010
on saturday	27	90.94	13	0.0010
we went	42	67.54	11	0.0008
as well	43	67.18	11	0.0008
couple of	30	84.18	10	0.0007
want to	41	68.01	10	0.0007
the moment	44	67.09	10	0.0007
<END> hi	21	99.44	9	0.0007
able to	50	61.19	9	0.0007
take care	23	96.00	8	0.0006
catch up	39	70.50	7	0.0005
other than	49	62.84	6	0.0005

Quantification In terms of quantification, Introverts generally tend to show a preference for a greater use of quantifiers, such as [*a lot*], [*a few*] and uniquely [*all the*], [*one of*], [*lots of*] and [*loads of*], whereas Extraverts show a preference for [*a bit*] and uniquely use [*couple of*]. Not only does this demonstrate an Extravert tendency to be looser and less specific, it also apparently reveals exaggeration on the part of the Introvert.

Social Devices The Extravert use of stylistic expressions such as [*catch up*] and [*take care*] indicate a relaxed and informal style; their omission points to a more socially restrained Introvert. A surprisingly neat equivalence in expression can be found between the Extravert use of [*other than*] rather than [*apart from*], although it is not immediately clear what might give rise to this.

Self/Other Reference References to self in the texts demonstrate differences between Extraverts and Introverts: Introverts make extensive use of the first person singular pronoun ([*i don't*], [*i went*], [*i'm going*], [*i can*], [*i've got*] are all unique to the Introvert text), and also show preference for the following shared bigrams: [*i've done*], [*i think*], [*i am*], [. *i*].

Table 5: Bigrams unique to Introvert corpus.

Bigram	Rank	$-2 \log \lambda$	Count	Ratio
. <END>	17	80.13	20	0.0022
i don't	18	78.77	18	0.0020
went to	25	63.53	15	0.0017
to go	34	56.65	14	0.0016
all the	47	43.06	12	0.0013
i went	50	42.70	12	0.0013
one of	32	57.45	11	0.0012
trying to	29	60.75	10	0.0011
i'm going	36	52.84	10	0.0011
i can	46	43.90	10	0.0011
on thursday	20	72.22	9	0.0010
don't know	21	69.76	9	0.0010
i've got	35	55.19	9	0.0010
lots of	26	62.29	8	0.0009
this week	39	48.51	8	0.0009
anyway ,	45	44.79	8	0.0009
should be	40	48.10	7	0.0008
on monday	41	47.91	6	0.0007
two weeks	31	58.65	5	0.0006
loads of	49	42.72	5	0.0006
<END> hello	44	45.05	4	0.0005
exam results	42	47.26	3	0.0003

For Extraverts, the only unique first person bigram is [*i'll be*], and they also show greater use of [*i was*] and [*i will*], although relatively less preferred than Introvert forms. This underscores the increased Introvert tendency to focus on self, whereas the only bigram containing a first person plural is unique to Extraverts ([*we went*]). The Extravert preference for the bigram [*up with*] typically indicates a shared experience (prompting the question *with whom?*) and greater sociability. These results apparently contradict Furnham (1990) on pronouns, but given that the vast majority of pronouns here are first-person singular, thus focusing on self, this is unsurprising.

Valence Bigrams containing negations were used significantly only by Introverts, as in [*i don't*] and [*don't know*] (indeed [*i don't*] is the bigram with most frequent use of *i*) , whilst Extraverts used the bigram [*a good*] which is suggestive of positive affect.[5] Similarly, the Extravert preference for [*looking forward*] and [*forward to*] (presumably as in *looking forward to*) also suggests a more positive disposition.

Ability Personal views on capability are suggested by the different collocations with infinitival *to*.[6] For Extraverts, their ability to do something should they choose is confidently and assertively relayed using *want–, need–,* and *able– (to)*; which they use uniquely. Introverts more timidly and tentatively state that they are [*trying to*] or possibly—and at some point in the future—they are [*going to*].

Modality Similarly, collocations with the verb *be* show a distinction in use of modal auxiliaries which has an effect on the projection of certainty. For example, Introverts are unique in their use of the weaker and more tentative *should be*, whereas Extraverts show a greater use of the stronger predictive [*will be*], and are unique in their use of the contracted form [*i'll be*] (*i will be*) (Coates, 1983).

Message Planning/Expression Looking towards surrogates of grammatical construction, Extraverts and Introverts differ in their use of connectives: Introverts show preference for the coordinating conjunctions [*, and*] and [*, but*], whilst Extraverts uniquely show use of the subordinating [*, which*], usually deployed in an evaluative sense.

Discussion

In summary, our results support earlier findings, and suggest some new conclusions.

We found that Extraverts produce texts with more words, which supports the previous findings for speech (Carment *et al.*, 1965), whilst the reduced concreteness of Extravert language is a novel finding. It may be a direct consequence of talking or writing more, if the pressure to produce words at a high rate (in order to hold the floor, for instance) diverts resources away from more detailed lexical planning. Introverts' greater preference for numbers and quantification fits with this, and is compatible with findings concerning the use of articles (Pennebaker and King, 1999), and suggestions of a more imprecise and 'looser' Extravert style (Furnham, 1990).

Extraverts' use of other or social referents, and Introverts' preference for self referents confirms Berry *et al.* (1997)'s previous findings for Extraversion and its *dominant/submissive* facets. Another possible manifestation of the increased Extravert social ability and ease in interaction is expressed by their use of surface features and social devices. We also note in passing the tendency of Extraverts to refer to days of the weekend, where Introverts refer to weekdays.

Our results on valence are consistent with previous findings on Introverts' preference for negations and negative emotion words, and the Extravert tendency for positive affect words is consistent with results for *warmth*. However, they do suggest that care should be taken over the relation between Extraversion and *dominant* facet features (cf. Isbister and Nass, 2000).

Expressions of definite modality and ability appear to be associated with Extraversion, although they may not be the same forms as those discussed in the context of powerful/less speech. Adoption of definite modalities can also be related to avoidance of tentativity (Pennebaker and King, 1999).

Turning to connectives, we note that our Introvert

[5] Further investigation shows that *good* is not directly negated (as in [*not good*]). Compare the Introvert [*i can*], which was generally followed by *not*. Although the effect of negation was not viewed as important by Pennebaker in the functioning of LIWC, it certainly has implications for models of language generation.

[6] This confirms the appropriacy of retaining functors usually filtered out by a stop list (cf. Damerau, 1993).

preference for [, *and*] and [, *but*] is consistent with studies using LIWC which found that the dictionary categories of Inclusion and Exclusion were both inversely correlate with Extraversion. However, [*other than*] and [*apart from*] would both fall into the same LIWC category, yet appear to distinguish opposite ends of the personality dimension.

Conclusion

By combining techniques from psycholinguistics and statistical natural language processing, we have been able to replicate previous findings on the expression of Extraversion through language, and uncover some new linguistic behaviours. Where existing content analysis tools could not detect reliable differences, more sensitive linguistic tools proved their worth.

Further, more technically sophisticated analyses can be carried out on this data, and we envisage the use of machine learning techniques to identify distinctive features from the texts, along with bigram analysis exploiting Parts of Speech tags. Additionally, the role of gender could be investigated.

Our findings could be exploited within the field of automatic language generation. As they stand, stochastic techniques would be needed; however, a cognitively-based personality model would allow a deeper approach, and that is our eventual goal.

Acknowledgements

Thanks to Elizabeth Austin, James Curran and our anonymous reviewers for advice and comments. This work was supported by the Economic and Social Research Council (Award R00429934162).

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph.D. thesis, Royal Institute of Technology, Stockholm.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Bradac, J. and Mulac, A. (1984). A molecular view of powerful and powerless speech styles. *Communication Monographs*, **51**, 307–319.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. Croom Helm, London.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proc of the 34^{th} Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**, 497–505.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, **29**, 433–448.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**(3), 509–544.

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**(2), 265–324.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Eysenck, H. and Eysenck, S. (1991). *Eysenck Personality Questionnaire-Revised*. Hodder, London.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**(1), 21–29.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (in prep.). Dictionary approaches to personality language. *in prep.*

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. J Human-Computer Studies*, **53**, 251–267.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *Int J Human-Computer Studies*, **43**, 223–239.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.

Pennebaker, W. and Francis, M. (1999). *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**(6), 1296–1312.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

The Role of Roles in Translating Across Conceptual Systems

Robert L. Goldstone (rgoldsto@indiana.edu)
Brian J. Rogosky (brogosky@indiana.edu)
Department of Psychology
Indiana University
Bloomington, IN. 47405

Abstract

We explore one aspect of meaning, the identification of matching concepts across systems (e.g. people, theories, or cultures). We present a computational algorithm called ABSURDIST (Aligning Between Systems Using Relations Derived Inside Systems for Translation) that uses only within-system similarity relations to find between-system translations. While illustrating the sufficiency of within-system relations to account for translating between systems, simulations of ABSURDIST also indicate synergistic interactions between intrinsic, within-system information and extrinsic information.

Conceptual Meaning and Translation

There have been two major answers to the question of how our concepts have meaning. The first answer is that concepts' meanings depend on their connection to the external world (Harnad, 1990). By this account, the concept Dog means what it does because our perceptual apparatus can identify features that characterize, if not define, dogs. Dog is characterized by features that are either perceptually given, or can be reduced to features that are perceptually given. This will be called the "external grounding" account of conceptual meaning. The second answer is that concepts' meanings depend on their connections to each other (Markman & Stillwell, 2001; Saussure, 1915/1959). By this account, Dog's meaning depends on Cat, Domesticated, and Loyal, and in turn, these concepts depend on other concepts, including Dog. The dominating metaphor here is of a conceptual web in which concepts all mutually influence each other (Quine & Ullian, 1970). A concept can mean something within a network of other concepts but not by itself. This will be called the "conceptual web" account.

The goal of this article is to argue for the synergistic integration of conceptual web and externally grounded accounts of conceptual meaning. However, in pursuing this argument, we will first argue for the sufficiency of the conceptual web account for a particular task associated with conceptual meaning. Then, we will show how the conceptual web account can be ably supplemented by external grounding to establish meanings more successfully than either method could by itself.

Our point of departure for exploring conceptual meaning will be a highly idealized and purposefully simplified version of a conceptual translation task. Consider two individuals, Joan and John, who each possesses a number of concepts. Suppose further that we would like some way to tell that Joan and John both have a concept of, say, **Mushroom**. Joan and John may not have exactly the same concept of **Mushroom**. John may believe mushrooms grow from seeds whereas Joan believes they grow from spores. More generally, Joan and John will differ in the rest of their conceptual networks because of their different experiences and levels of expertise. Still, it seems desirable to say that Joan and John's **Mushroom** concepts correspond to one another. We will describe a network that translates between concepts in two systems, placing, for example, Joan and John's **Mushroom** concepts in correspondence with each other.

Translation across systems is generally desirable and specifically necessary in order to say things like "John's concept of mushrooms is less informed than Joan's." Fodor and Lepore have taken the existence of this kind of translation as a challenge to conceptual web accounts of meaning (Fodor & Lepore, 1992). By Fodor and Lepore's interpretation, if a concept's meaning depends on its role within the larger system, and if there are some differences between the systems, then the concept's meaning would be different in the two systems. A natural way to try to salvage the conceptual web account is to argue that determining corresponding concepts across systems does not require the systems to be identical, but only similar. However, Fodor (Fodor, 1998; Fodor & Lepore, 1992) insists that the notion of similarity is not adequate to establish that Joan and John both possess a **Mushroom** concept. Fodor argues that "saying what it is for concepts to have similar, but not identical contents presupposes a prior notion of beliefs with similar but not identical concepts" [Fodor, 1998, p. 32].

The ABSURDIST Algorithm for Cross-system Translation

We will now present a simple neural network called ABSURDIST (Aligning Between Systems Using Relations Derived Inside Systems for Translation) that finds conceptual correspondences across two systems (two people, two time slices of one person, two scientific theories, two developmental age groups, two language communities, etc.) using only inter-conceptual similarities, not conceptual identities, as input. Thus, ABSURDIST will take as input two systems of concepts in which every concept of a system is defined exclusively in terms of its dissimilarities to other concepts in the same system. Laakso and

Cottrell (2000) describe another neural network model that uses similarity relations within two systems to compare the similarity of the systems. ABSURDIST produces as output a set of correspondences indicating which concepts from System A correspond to which concepts from System B. These correspondences serve as the basis for understanding how the systems can communicate with each other without the assumption made by Fodor (1998) that the two systems have exactly the same concepts. The existence of ABSURDIST provides evidence against Fodor's argument that similarities between people's concepts are an insufficient basis for determining that two people share an equivalent concept. ABSURDIST is not a complete model of conceptual meaning or translation. Our point is that even if the only relation between concepts in a system were simply similarity, this would still suffice to find translations of the concepts in different systems.

Elements $A_{1..m}$ belong to System A, while elements $B_{1..n}$ belong to System B. $C_t(A_q,B_x)$ is the activation, at time t, of the unit that represents the correspondence between the qth element of A and the xth element of B. There will be m·n correspondence units, one for each possible pair of corresponding elements between A and B. In the current example, every element represents one concept in a system. The activation of a correspondence unit is bound between 0 and 1, with a value of 1 indicating a strong correspondence between the associated elements, and a value of 0 indicating strong evidence that the elements do not correspond. Correspondence units dynamically evolve over time by the equations:

$$\begin{aligned} &\text{if } N\left(C_t\left(A_q,B_x\right)\right) \geq 0 \text{ then } C_{t+1}\left(A_q,B_x\right) = C_t\left(A_q,B_x\right) + N\left(C_t\left(A_q,B_x\right)\right)\left(\max - C_t\left(A_q,B_x\right)\right)L \\ &\text{else } C_{t+1}\left(A_q,B_x\right) = C_t\left(A_q,B_x\right) + N\left(C_t\left(A_q,B_x\right)\right)\left(C_t\left(A_q,B_x\right) - \min\right)L \end{aligned} \quad (1).$$

If $N(C_t(A_q,B_x))$, the net input to a unit that links the qth element of A and the xth element of B, is positive, then the unit's activation will increase as a function of the net input, a squashing function that limits activation to an upper bound of max=1, and a learning rate L (set to 1). If the net input is negative, then activations are limited by a lower bound of min=0. The net input is defined as

$$N\left(C_t\left(A_q,B_x\right)\right) = \alpha E\left(A_q,B_x\right) + \beta R\left(A_q,B_x\right) - \chi I\left(A_q,B_x\right), \quad (2)$$

where the E term is the external similarity between A_q and B_x, R is their internal similarity, I is the inhibition to placing A_q and B_x into correspondence that is supplied by other developing correspondence units, and a+b+c=1. When a=0, then correspondences between A and B will be based solely on the similarities among the elements within a system, as proposed by a conceptual web account. The amount of excitation to a unit based on within-system relations is given by

$$R\left(A_q,B_x\right) = \frac{\sum_{\substack{r=1 \\ r\neq q}}^{m}\sum_{\substack{y=1 \\ y\neq x}}^{n} S\left(D\left(A_q,A_r\right),D\left(B_x,B_y\right)\right)C_t\left(A_r,A_y\right)}{Min\left(m,n\right)-1},$$

where $D(A_q,A_r)$ is the psychological distance between elements q and r in System A, and S is a negative exponential function of the absolute difference between S's two arguments. The amount of inhibition is given by

$$I\left(A_q,B_x\right) = \frac{\sum_{\substack{r=1 \\ r\neq q}}^{m} C_t\left(A_r,B_x\right) + \sum_{\substack{y=1 \\ y\neq x}}^{n} C_t\left(A_q,B_y\right)}{m+n-2}.$$

According to the equation for R, Elements q and x will tend to be placed into correspondence to the extent that they enter into similar similarity relations with other elements. For influencing alignments, the similarity between two distances is weighted by the strengths of the units that align elements that are placed in correspondence by the distances. The equation for R represents the sum of the supporting evidence (the consistent correspondences), with each piece of support weighted by its relevance (given by the S term). The inhibitory I term is based on a one-to-one mapping constraint (Falkenhainer, Forbus, & Gentner, 1989). The unit that places A_q into correspondence with B_x will tend to become deactivated if other strongly activated units place A_q into correspondence with other elements from B, or B_x into correspondence with other elements from A.

Correspondence unit activations are initialized to random values selected from a normal distribution with a mean of 0.5 and a standard deviation of 0.05. In our simulations, Equation (1) is iterated for a fixed number of cycles. It is assumed that ABSURDIST places two elements into correspondences if the activation of their correspondence unit is greater than or equal to 0.55 after the fixed number of iterations have been completed (4000 cylces in the simulations described below).

Assessing ABSURDIST's Performance

In assessing ABSURDIST's performance, we will assume that conceptual dissimilarities obey Euclidean distance metric assumptions, and are interpretable as distances between concepts lying in a geometric space. Our general method for evaluating ABSURDIST will be to generate a number of elements in a two dimensional space, with each element identified by its value on each of the two dimensions. These will be the elements of System A, and each is represented as a point in space. System B's elements are created by copying the points from System A and adding Gaussian noise to each of the dimension values of each of the points. Then, equation (1) is used to update correspondences across the two systems for a fixed number of iterations. The correspondences computed by ABSURDIST are then compared to the correct correspondences. Two elements correctly correspond to each other if the element in System B was originally copied from the element in System A.

Noise tolerance and system complexity

An initial set of simulations was conducted to determine how robust the ABSURDIST algorithm was to noise and how well the algorithm scaled to different sized systems. We ran a 7 X 6 factorial combination of simulations, with 7 levels of added noise and 6 different numbers of elements per system. Noise was infused into the

algorithm by varying the displacement between corresponding points across systems. The points in System A were set by randomly selecting dimension values from a uniform random distribution with a range from 0 to 1000. System B points were copied from System A, and Gaussian noise with standard deviations of 0, 0.1, 0.2, 0.3, 0.4, 0.5, or 0.6% was added to the points of B. The number of points per system was 3, 4, 5, 6, 10, or 15. α was set to 0, β was set to 0.4, and χ to 0.6. The values for β and χ were selected because they were the most balanced weights that produced fewer than 5% two-to-one correspondences. For each of the 42 combinations of noise and number of items, 1000 separate randomized starting configurations were tested. The results from this simulation are shown in Figure 1, which plots the percentage of simulations in which each of the proper correspondences between systems is recovered. For example, for 15-item systems, the figure plots the percentage of time that all 15 correspondences are recovered. The graph shows that performance gradually deteriorates with added noise, but that the algorithm is robust to at least modest amounts of noise.

More surprisingly, Figure 1 also shows that the algorithm's ability to recover true correspondences generally increases as a function of the number of elements in each system, at least for small levels of noise. One might have thought that as more elements were matched between systems that there would be greater confusion between elements, given that the size of the bounding region remains constant. As the number of elements in a system increases, the similarity relations between those elements provide increasingly strong constraints that serve to uniquely identify each element. If one generated random translations that were constrained to allow only one-to-one correspondences, then the probability of generating a completely correct translation would be 1/N!. Thus, with 0.6% noise, the 23% rate of recovering all 3 correspondences for a 3-item system is slightly above chance performance of 16.67%. However, with the same amount of noise, the 17% rate of recovering all of the correspondences for a 15-item system is remarkably higher than the chance rate of $7.6X10^{-13}$. Thus, at least in our highly simplified domain, we have support for the argument (Lenat & Feigenbaum, 1991) that establishing meanings on the basis of within-system relations becomes easier, not harder, as the size of the system increases.

Interactions between extrinsic and intrinsic determinants of alignments

The simulation above indicates that within-system relations are sufficient for discovering between-system translations, but this should not be interpreted as suggesting that the meaning of an element is not also dependent on relations extrinsic to the system. ABSURDIST offers a useful, idealized system for examining interactions between intrinsic (within-system) and extrinsic (external to

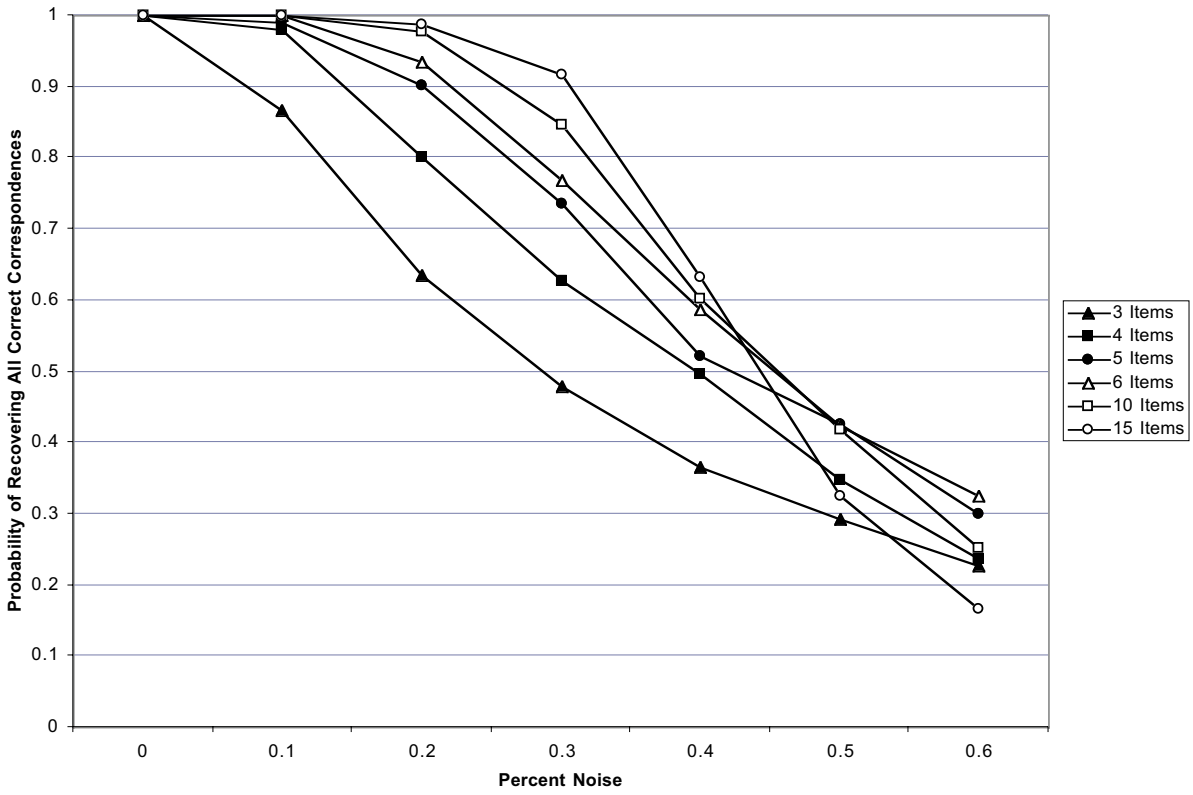


Figure 1

the system) aspects of meaning. One way to incorporate extrinsic biases into the system is by initially seeding correspondence units with values. Thus far, all correspondence units have been seeded with initial activation values tightly clustered around 0.5. However, in many situations, there may be external reason to think that two elements correspond to each other: they may receive the same label, they may have perceptual attributes in common, they may be associated with a common event, or a teacher signal may have provided a hint that the two elements correspond. In these cases, the initial seed-value may be significantly greater than 0.5.

Figure 2 shows the results of a simulation of ABSURDIST with different amounts of extrinsic support for a selected correspondence between two elements. Two systems are generated by randomly creating a set of points in two dimensions for System 1, and copying the points' coordinates to System 2 while introducing 0.6% noise to their positions. When Seed = 0.5, then no correspondence is given an extrinsically supplied bias. When Seed=0.75, then one of the true correspondences between the systems is given a larger initial activation than the other correspondences. For a system made up of 15 elements, a mapping accuracy of 31% is obtained without any extrinsic assistance (Seed=0.5). If seeding a single correct correspondence with a value of 1 rather than 0.5 allowed ABSURDIST to recover just that one correspondence with 100% probability, then accuracy would increase at most to 35.6% (((.31 * 14) + 1)/15). The reference line in Figure 2 shows these predicted increases in accuracy. For all systems tested, the observed increment in accuracy far outstretches the increase in accuracy predicted if seeding a correspondence only helped that correspondence. Moreover, the amount by which translation accuracy improves beyond the amount predicted generally increases as a function of system size. Thus, externally seeding a correspondence does more than just fix that correspondence. In a system where correspondences all mutually depend upon each other, seeding one correspondence has a ripple-effect through which other correspondences are improved.

Equation 2 provides a second way of incorporating extrinsic influences on correspondences between systems. This equation defines the net input to a correspondence unit as an additive function of the extrinsic support for the correspondence, the intrinsic support, and the competition against it. Thus far, the extrinsic support has been set to 0. The extrinsic support term can be viewed as any perceptual, linguistic, or top-down information that suggests that two objects correspond (this differs from the philosopher's use of "external meaning" to refer to the causal determinants of a concept). To study interactions between extrinsic and intrinsic support for correspondences, we conducted 1000 simulations that started with 10 randomly placed points in a two-dimensional space for System A, and then copied these points over to System B with Gaussian-distributed noise. The intrinsic, role-based support is determined by the previously described equations. The extrinsic support term of Equation 2 is given by a negative exponential function of the absolute distance between the two concepts' absolute locations. Thus, the correspondence unit connecting q and x will tend to be strengthened if q and x have similar coordinates. This is extrinsic support because the similarity of q's and x's coordinates can be determined without any reference to other elements.

In conducting this third simulation, we assigned three different sets of weights to the extrinsic and intrinsic support terms. For the "Extrinsic only" results of Figure 3, we set α=0.4, β=0, and χ=0.6. For the "Intrinsic only" results, we set α=0, β=0.4, and χ=0.6. For "Intrinsic and Extrinsic," we set α=0.2, β=0.2, and χ=0.6.

Figure 3 shows that using only information intrinsic to a system results in better correspondences than using only extrinsic information. This is because corresponding elements that have considerably different positions in their systems can often still be properly connected with intrinsic information if other proper correspondences can be

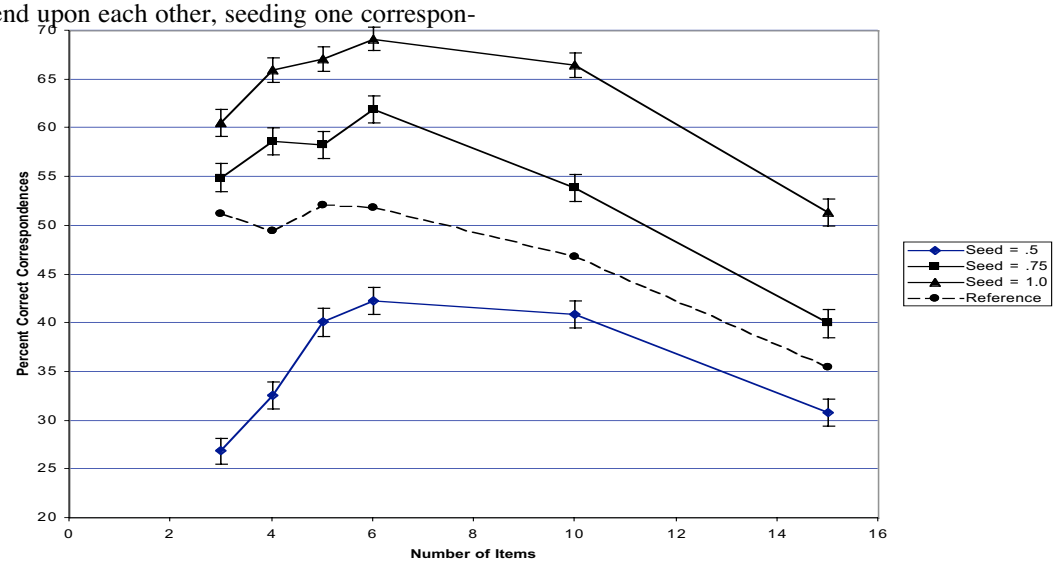


Figure 2

recovered. The intrinsic support term is more robust than the extrinsic term because it depends on the entire system of emerging correspondences. For this reason, it is surprising that the best translation performance is found when intrinsic and extrinsic information are both incorporated into Equation 2 with equal weight. The superior performance of the network that uses both intrinsic and extrinsic information derives from its robustness in the face of noise. Some distortions to points of System B adversely affect the intrinsic system more than the extrinsic system. For example, a slight distortion to a point may make its pattern of distances to other points quite similar to another point. A system that incorporates both sources of information will tend to recover well from either disruption to absolute or relative positions if the other source of information is reasonably intact.

Discussion

The ABSURDIST model makes two theoretically important points. First, translations between two systems can be found using only information about the relations between elements within a system. The claim is that the concept in Person A that matches a concept in Person B can be found considering only the relations between concepts in Person A, and the relations between concepts in Person B. ABSURDIST demonstrates how a holistic conception of meaning is compatible with the goal of determining correspondences between concepts across individuals. Two people need not have exactly the same systems, or even the same number of concepts, to create proper conceptual correspondences. Contra Fodor (Fodor, 1998; Fodor & Lepore, 1992) information in the form of inter-conceptual similarities suffices to find inter-system translations between concepts. It is often easier to find translations for large systems than small systems.

The second important theoretical contribution of ABSURDIST is to formalize some of the ways that intrinsic, within-system relations and extrinsic, perceptual information synergistically interact in determining conceptual alignments. Intrinsic relations suffice to determine cross-concept translations, but if extrinsic information is available, more robust, noise-resistant translations can be found. The synergistic benefit of combining intrinsic and extrinsic information sheds new light on the debate on accounts of conceptual meaning. It is common to think of intrinsic and extrinsic accounts of meaning as being mutually exclusive, or at least zero-sum. Seemingly, either a concept's meaning depends on information within its conceptual system or outside of its conceptual system, and to the extent that one dependency is strengthened, the other dependency is weakened. In opposition to this zero-sum perspective on intrinsic and extrinsic meaning, ABSURDIST offers a framework in which a concepts' meaning is both intrinsically and extrinsically determined (see also two-factor theories in philosophy such as Block, 1986), and the external grounding makes intrinsic information more, not less, powerful. To claim that all concepts in a system depend on all of the other concepts in a system is perfectly compatible with claiming that all of these concepts have a perceptual basis.

We have focused on the application of ABSURDIST to the problem of translating between different people's conceptual systems. However, the algorithm is applicable to a variety of situations in which elements from two systems must be placed in correspondence in an efficient and reasonable (though not necessarily optimal) manner. A

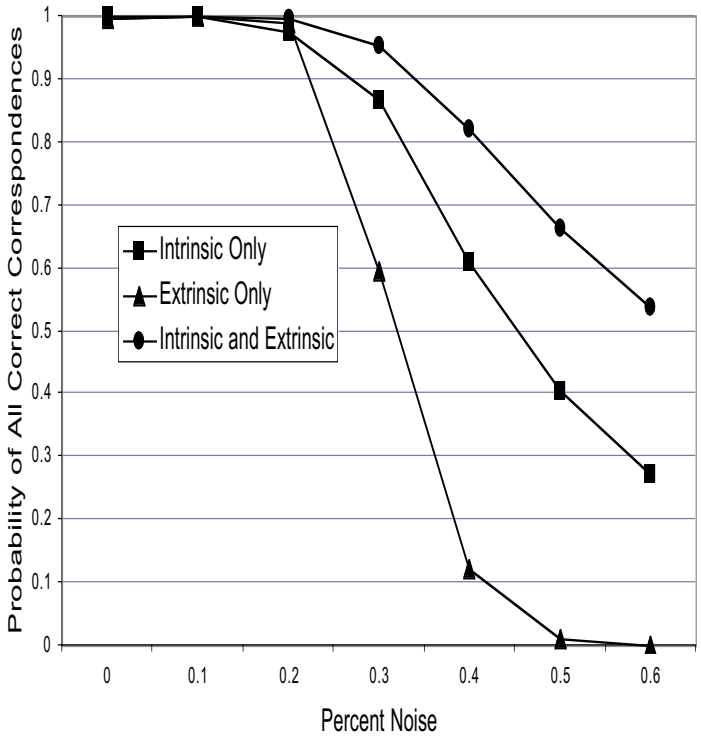


Figure 3

combination of properties makes ABSURDIST particularly useful for applications in cognitive science: 1) the algorithm can operate solely on relations within a system, 2) the within-system relations can be as simple as generic similarity relations, 3) the algorithm can combine within-system and between-systems information when each is available, 4) the algorithm has a strong bias to establish one-to-one correspondences, and 5) the algorithm does not require larger numbers of iterations for convergence as the number of elements per system increases. Some of the domains of application for ABSURDIST include object recognition, analogy, and automatic translation.

Object recognition

The ABSURDIST algorithm can be applied to the problem of object recognition that is invariant to translation, rotation, and reflection. For this application, a pictorial object is the system, and points on the object are elements of the system. A standard solution to recognizing rotated objects is to find matching landmark points that are identifiable on a known object and an input object to be recognized (Ullman, 1996). Once identified, these landmarks can reveal how the input would need to be rotated to match the known object. Even if no extrinsically aligned landmarks can be identified, ABSURDIST can still match the objects by taking advantage of the wealth of information contained in within-object proximity relations (Edelman, 1999).

Analogy

ABSURDIST offers a complementary approach to analogical reasoning between domains. Most existing models of analogical comparison represent the domains to be compared in terms of richly structured propositions (Hummel & Holyoak, 1997; Eliasmith & Thagard, 2001). In many cases, such as single words or pictures, it is difficult to come up with propositional encodings that capture an item's meaning. In such cases, ABSURDIST's unstructured similarity relations are a useful addition to existing models of analogical reasoning.

Automatic dictionary translation

The small-scale simulations conducted here leave open the promise of applying ABSURDIST to much larger translation tasks, such as dictionaries, thesauri, encyclopedias, and organizational structures. ABSURDIST could provide automatic translations between dictionaries of two different languages, using only co-occurrence relations between words within each dictionary (Burgess & Lund, 2000; Landauer & Dumais, 1997), perhaps supplemented by a small number of external hints (e.g. that French "chat" and English "cat" might correspond to each other because of their phonological similarity).

Acknowledgments

We would like to thank Gary Cottrell, Eric Dietrich, Shimon Edelman, Stevan Harnad, John Hummel, Michael Lynch, Art Markman, and Mark Steyvers for comments on an earlier version of this research. This research was funded by NIH grant MH56871 and NSF grant 0125287.

References

Block, N. (1986). Advertisement for a semantics for psychology. Midwest Studies in Philosophy, 10, 615-78.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Diettrich & A. B. Markman (Eds.) Cognitive dynamics: Conceptual change in humans and machines. (pp. 117-156). Mahwah, NJ: Lawrence Erlbaum Associates.

Edelman, S. (1999). Representation and recognition in vision. Cambridge, MA: MIT Press.

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. Cognitive Science, 25, 245-286

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41, 1-63.

Fodor, J. (1998). Concepts: Where cognitive science went wrong. Oxford: Clarendon Press.

Fodor, J., & Lepore, E. (1992). Holism. Oxford, UK: Blackwell.

Harnad, S. (1990). The symbol grounding problem. Physica D, 42, 335-346.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. Philosophical Psychology, 13, 47-76.

Lenat, D. B., & Feigenbaum, E. A. (1991). On the thresholds of knowledge, Artificial Intelligence, 47, 185-250.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. Journal of Experimental and Theoretical Artificial Intelligence, 13, 329-358.

Quine, W. V., & Ullian, J. S. (1970). The Web of Belief. New York: McGraw-Hill.

Saussure, F. (1915/1959). Course in general linguistics. New York: McGraw-Hill.

Ullman, S. (1996). High-level vision. Cambridge, MA: MIT Press.

The Theory of Mind in Strategy Representations

Andrew S. Gordon (gordon@ict.usc.edu)
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey CA USA

Abstract

Many scientific fields continue to explore cognition related to Theory of Mind abilities, where people reason about the mental states of themselves and others. Experimental and theoretical approaches to this problem have largely avoided issues concerning the contents of representations employed in this class of reasoning. In this paper, we describe a new approach to the investigation of representations related to Theory of Mind abilities that is based on the analysis of commonsense strategies. We argue that because the mental representations of strategies must include concepts of mental states and processes, the large-scale analysis of strategies can be informative of the representational scope of Theory of Mind abilities. The results of an analysis of this sort are presented as a description of thirty representational areas that organize the breadth of Theory of Mind concepts. Implications for Theory Theories and Simulation Theories of Theory of Mind reasoning are discussed.

Investigating the Theory of Mind

One of the most challenging areas of research in the cognitive sciences has concerned the Theory of Mind, in reference to the abilities humans have to perceive and reason about their own mental states and the mental states of other people. Along with the inherent difficulties in investigating behavior that is largely unobservable, researchers in this area are required to be extremely interdisciplinary. Many research fields contribute evidence that influences our understanding of these human abilities, although the methods used to gather this evidence are diverse.

Researchers in developmental psychology largely choose to investigate the Theory of Mind as a set of abilities that progressively emerge in normal child development (Wellman & Lagattuta, 2000). By the last half of their second year, toddlers demonstrate an understanding of the role of intentionality in action, and that other people have subjective experiences. By the age of four and five, children comprehend and use vocabulary to refer to mental states such as thoughts, imaginations, and knowledge. As children advance into grade-school years and adulthood, there is a growing appreciation of people as active constructors and interpreters of knowledge, and awareness that others have ongoing thoughts. There is evidence that Theory of Mind capabilities continue to improve into the later adult years, even while non-social reasoning abilities begin to degrade (Happé et al., 1998).

In the research area of abnormal psychology, compelling cases have been made relating illnesses such as autism (Baron-Cohen, 2000) and schizophrenia (Corcoran, 2001) to deficits in Theory of Mind abilities. Neuropathology studies of stroke patients have provided evidence that Theory of Mind mechanisms may be localized in the brain (Happé et al., 1999), and ongoing functional neuroimaging studies continue to provide further evidence for localization (Frith & Frith, 2000).

In search of a more process-oriented understanding of Theory of Mind abilities, it is the philosophy community that has made the most contributions, proposing two classes of process theories that have been extensively debated. First, the Theory Theory hypothesizes that Theory of Mind abilities are computed by prediction and explanation mechanisms by employing representation-level knowledge about mental attitudes (Gopnik & Meltzoff, 1997; Nichols & Stich, forthcoming). The opposing view is that of Simulation Theory (Goldman, 2000), which argues that Theory of Mind abilities are computed by imagining that you are in the place of the other person, then inferring their mental states by monitoring the processing that is done by your own cognitive mechanisms. While some high-level process-oriented cognitive models have been proposed (e.g. Nichols and Stich, 2000), there are many unanswered questions that prohibit the creation of detailed, computational models of Theory of Mind abilities.

Most lacking in our theoretical understanding of Theory of Mind abilities is a description of the specific *contents* of the mental representations that are employed in this reasoning. There is general agreement that these representational elements must include concepts such as *beliefs* and *desires* (e.g. Harris, 1996), and these two concepts in particular have taken a privileged role in the cognitive models that have been proposed. A potential benefit of the focus on these concepts is that this representational area (beliefs, desires, intentionality) is among the very few where established axiomatic theories have been developed in the artificial intelligence community (Cohen & Levesque, 1990). Continued artificial intelligence progress in developing axioms for inference concerning

mental states (e.g. Ortiz, 1998) will greatly support the plausibility of the Theory Theory approach.

However, there is a general sense throughout the fields investigating Theory of Mind abilities that the contents of these representations go far beyond simple notions of beliefs and desires, particularly among developmental psychologists investigating the role that language plays in acquiring mental state concepts. Several studies have been conducted that investigate the linguistic environment of children for the presence of Theory of Mind related terms, where the conceptual scope is much more broadly construed. Dyer et al. (2000) best exemplifies the broad conceptual scope of this line of work, which compared the frequency of 455 mental state terms that appear in young children's storybooks. This list included 102 cognitive state terms (e.g. notice, wonder), 152 emotional state terms (e.g. nervous, boring), 84 desire and volition terms (e.g. hope, wish), and 117 moral evaluation and obligation terms (e.g. ought, terrible), where the complete list was compiled from previous language studies.

While these linguistic approaches help to broaden our conception of the scope of representational elements in Theory of Mind reasoning, many of the traditional concerns about the relationship between language and mental representation may apply. Particularly, there is no reason to believe that any full enumeration of mental state terms must parallel the breadth of concepts that are represented and manipulated by reasoning processes. The inherent subjectivity of these concepts may serve to restrict the introduction of new vocabulary in the lexicon as compared with other topics of discourse. Likewise, the remarkable creativity that is evident in human language use may mislead us to believe that there are representational distinctions between concepts that are in fact functionally synonymous. While these linguistic approaches have been persuasive in arguing for a broader scope of Theory of Mind representations in our cognitive models, a new investigative methodology for concept enumeration would be useful.

Analogy as an Investigative Tool

In previous work (Gordon, 2001a), we argued that progress in a different area of cognition – that of analogical reasoning – could be a basis for a novel methodology for the investigation of mental representations. As a cognitive process, analogical reasoning has received an enormous amount of attention, both theoretical and experimental, with the aim of understanding how people draw analogies between two different cases in working memory. The prevailing explanation is based on the notion of structural alignment of the mental representations that people have of these cases (Gentner, 1983). That is, two different cases are judged as strongly analogous when portions of the structured mental representation of one case can be mapped onto structurally identical portions of the other. Strong empirical support for the structure mapping theory of analogical reasoning (see Gentner & Markman, 1997, for a review) presents an opportunity: if structural alignment of representations are necessary to process analogies, then an analysis of the analogies that people naturally make can reveal the sorts of representations that they must employ.

In this previous work, two main claims were put forth. First it was noted that there is something particularly interesting about the commonsense notion of a strategy as it relates to analogies between planning cases. People readily see analogies between planning behaviors exhibited in vastly different goal-driven domains. For example, a retreating military force that destroys the supplies that they can't take with them may be viewed as analogous to the company that publicly releases its closely guarded industrial secrets in the face of a hostile corporate takeover. In both cases we would say the actors were using the same strategy, one that is so commonly recognized that it has been given a name, *scorched earth policy*. In accordance with the structure mapping theory of analogy, it was argued that strategies like this one are structured mental representations that are shared between the analogous planning cases where they are employed.

The second claim was that mental representations of strategies necessarily include references to the mental states and processes of people. For example, to be considered as an example of scorched earth policy, it must be the case that the actor foresees he will loose possession of a valuable resource to an advancing enemy, he foresees that after the enemy gains possession of it he will use the resource to further advance against him, and that the actor imagines that what he does to these resources will make them useless to the enemy. Strategic analogies show us that concepts such as these that specifically refer to mental processes must be explicitly represented in cognition. As the mental state concepts in these statements are exactly the sort relevant to Theory of Mind abilities, our claim is that the analysis of strategies provides a means of identifying the breadth of reified mental state concepts that are available in support of this class of reasoning.

In order to explore the representational scope of strategies, we undertook a large-scale strategy representation effort (Gordon, 2001b). First, 372 commonsense strategies were collected from 10 different planning domains using directed expert interviews, the analysis of texts that are encyclopedic of strategies in a particular domain, and the introspective elaboration of strategies in our own areas of expertise. To identify the representational requirements of this catalog of strategies, we developed a notational form called a *pre-formal representation* that would allow us

to commit to the specific semantic elements in the representation of a strategy without adhering to the syntactic constraints that would be necessary in more formal, logic-based representations. After authoring pre-formal representations of each of the 372 strategies, the component concepts were grouped into sets of synonyms to form a controlled vocabulary consisting of 989 unique concepts. This list was then organized into 48 representational areas that parallel both those that are traditionally the subject of formal commonsense knowledge representation (e.g. time and events) and those that are viewed as component cognitive processes in previous cognitive modeling work (planning and memory retrieval).

Eighteen of the representational areas that were identified in this previous work did not concern the mental states and processes of people. A large portion of these areas related more generally to the physical world, including concepts of time, space, events, states, objects, numbers, sets, and taxonomies. The remaining portion of these eighteen areas concerned people directly, but not their mental states in particular, and included terms for the relationships they hold, the organizations they participate in, their abilities, activities, and non-mental actions.

The other thirty representational areas that were identified deal specifically with the mental life of people. What is interesting about this collection of representational terms is that its scope is significantly larger than what has been suggested in cognitive models of Theory of Mind abilities or even in the contents of the lexicons used in the analysis of language for Theory of Mind concepts.

The primary direction in which these representational areas expand the scope of previous work is with respect to folk psychological conceptions of mental *processes*, whereas previous work has focused mostly on mental *states*. While the terms revealed in our investigation certainly include mental state concepts such as beliefs and desires, these are coupled with concepts describing the mental processes that affect these states, such as the mental processes of removing the justification for a belief and the process of abandoning of a goal to achieve some desired state. In short, the representations that appear to be necessary to account for strategic analogies outline a set of processes that constitute a cognitive architecture.

Theory of Mind Representations

In order to elaborate on the mental state and mental process components that are evident in the organization of strategy representation terms, this section briefly describes each of the thirty representational areas (of the 48 total) specifically related to Theory of Mind reasoning. Each area is listed with a short area title, the number of unique representational terms (out of 989) in the area found in strategy representations, a short definition of the scope of the area, and a few examples of the specific terms in the area.

1. Managing knowledge (30 terms): The knowledge that agents have is a set of beliefs that may be true or false based on certain justifications, and can be actively assumed true, affirmed, or disregarded entirely. Examples: *Assumption*, *Justification*, *Revealed false belief*.

2. Similarity comparison (16 terms): Agents can reason about the similarity of different things using different similarity metrics, where analogies are similar only at an abstract level. Examples: *Class similarity*, *Similarity metric*, *Make analogy*.

3. Memory retrieval (3 terms): Agents have a memory that they use to store information through a process of memorization, and may use memory aids and cues to facilitate retrieval. Examples: *Memory cue*, *Memory retrieval*, *Memorize*.

4. Emotions (8 terms): Agents may experience a wide range of emotional responses based on their appraisal of situations, which defines their emotional state. Examples: *Anxiety emotion*, *Pride emotion*, *Emotional state*.

5. Explanations (17 terms): Agents generate candidate explanations for causes in the world that are unknown, and may have preferences for certain classes of explanations. Examples: *Candidate explanation*, *Explanation preference*, *Explanation failure*.

6. World envisionment (48 terms): Agents have the capacity to imagine states other than the current state, to predict what will happen next or what has happened in the past, and to determine the feasibility of certain state transitions. Examples: *Causal chain*, *Envisioned likelihoood*, *Possible envisioned state*.

7. Execution envisionment (23 terms): One mode of envisionment is that of imagining the execution of a plan for the purpose of predicting possible conflicts, execution failures, side effects, and the likelihood of successful execution. Examples: *Envisioned failure*, *Side effect*, *Imagine possible execution*.

8. Causes of failure (31 terms): In attempting to explain failures of plans and reasoning, agents may employ a number of explanation patterns, such as explaining a scheduling failure by the lack of time, or a planning failure by a lack of resources. Examples: *False triggered monitor*, *Lack of ability*, *Successful execution of opposing competitive plan*.

9. Managing expectations (8 terms): Envisionments about what will happen next constitute expectations, which can be validated or violated based on what actually occurs. Examples: *Expectation violation*, *Unexpected event*, *Remove expectation*.

10. Other agent reasoning (8 terms): Envisionments about the planning and reasoning processes of other

agents allow an agent to imagine what they would be thinking about if they were them. Examples: *Guess expectation*, *Guess goal*, *Deduce other agent plan*.

11. Threat detection (15 terms): By monitoring their own envisionments for states that violate goals, an agent can detect threats and track their realization. Examples: *Envisioned threat*, *Realized threat*, *Threat condition*.

12. Goals (27 terms): Goals of agents describe world states and events that are desired, and include both states and events that are external to the planner as well as those that characterize desired internal mental states and processes. Examples: *Auxiliary goal*, *Knowledge goal*, *Shared goal*.

13. Goal themes (6 terms): A potential reason that an agent may have a goal could be based on the roles that agents have in relationships and organizations, or because of a value that they hold. Examples: *Generous theme*, *Good person theme*, *Retaliation theme*.

14. Goal management (28 terms): Agents actively manage the goals that they have, deciding when to add new goals, commence or suspend the pursuit of goals, modify or specify their goals in some way, or abandon them altogether. Examples: *Currently pursued goal*, *Goal prioritization*, *Suspend goal*.

15. Plans (32 terms): The plans of agents are descriptions of behaviors that are imagined to achieve goals, and can be distinguished by the types of goals that they achieve or by how they are executed, and may be composed of other plans or only partially specified. Examples: *Adversarial plan*, *Repetitive plan, Shared plan*.

16. Plan elements (28 terms): Plans are composed of subplans, including branches that are contingent on factors only known at the time of execution. They may have iterative or repetitive components, or include components that are absolutely required for a plan to succeed. Examples: *If then*, *Iteration termination condition*, *Triggered start time*.

17. Planning modalities (17 terms): The selection of plans can be done in a variety of different ways, such as adapting old plans to current situations, collaboratively planning with other agents, and counterplanning against the envisioned plans of adversaries. Examples: *Adversarial planning*, *Auxiliary goal pursuit*, *Imagined world planning*.

18. Planning goals (27 terms): The planning process is directed by abstract planning goals of an agent, which include goals of blocking threats, delaying events, enabling an action, preserving a precondition, or satisfying the goals of others. Examples: *Avoid action*, *Delay duration end*, *Maximize value*.

19. Plan construction (30 terms): Agents construct new plans by specializing partial plans, adding and ordering subplans, and resolving planning problems when they arise. Examples: *Candidate plan*, *Planning failure*, *Planning preference*.

20. Plan adaptation (18 terms): Existing plans can be adapted and modified by substituting values or agency, and by adding or removing subplans to achieve goals given the current situation. Examples: *Adaptation cost*, *Adaptation failure*, *Substitution adaptation*.

21. Design (8 terms): One modality of planning is design, where the constructed plan is a description of a thing in the world within certain design constraints, and where the resulting things have a degree of adherence to this design. Examples: *Design adherence*, *Design failure*, *Designed use*.

22. Decisions (38 terms): Agents are faced with choices that may have an effect on their goals, and must decide among options based on some selection criteria or by evaluating the envisioned consequences. Examples: *Best candidate*, *Decision justification*, *Preference*.

23. Scheduling (23 terms): As agents select plans, they must be scheduled so that they are performed before deadlines and abide by other scheduling constraints. Plans may have scheduled start times and durations, or may be pending as the planner waits for the next opportunity for execution. Examples: *Deadline*, *Pending plan*, *Scheduling constraint*.

24. Monitoring (18 terms): Agents monitor both states and events in the world and in their own reasoning processes for certain trigger conditions which may prompt the execution of a triggered action. Examples: *First monitor triggering*, *Monitoring duration*, *Monitor envisionment*.

25. Execution modalities (11 terms): Plans can be executed in a variety of ways, including consecutively along with other plans, in a repetitive manner, and collaboratively along with other agents. Examples: *Concurrent execution*, *Continuous execution*, *Periodic execution*.

26. Execution control (28 terms): A planner actively decides to begin the execution of a plan, and may then decide to suspend or terminate its execution. A suspended plan can later be resumed from the point the agent left off. Examples: *Execution delay*, *Suspend execution*, *Terminate activity*.

27. Repetitive execution (16 terms): Some plans and subplans are executed iteratively for some number of times, or repetitively until some termination condition is achieved. Examples: *Current iteration*, *Iteration completion*, *Remaining repetition*.

28. Plan following (29 terms): Agents track the progress of their plans as they execute them in order to recognize when deadlines are missed, preconditions are satisfied, and when they have successfully achieved the goal. Examples: *Achieve precondition*, *Miss deadline*, *Successful execution*.

29. Observation of execution (29 terms): Agents can track the execution of plans by other agents, evaluating the degree to which these executions adhere to performance descriptions known to the observing agent. Examples: *Observed execution*, *Assessment criteria*, *Performance encoding*.

30. Body interaction (15 terms): The physical body of an agent translates intended actions into physical movements, and sometimes behaves in unintended ways. The body modifies the planner's knowledge through perception of the world around it, and by causing a sensation of execution. Examples: *Impaired agency*, *Nonconscious execution*, *Attend*.

Discussion

There exists no infallible technique for identifying the contents of the mental representations used in reasoning. The approach described here, where our theoretical understanding of analogical reasoning is used as an investigative tool, relies heavily on our analytic abilities in describing the shared relational structure of analogous cases as much as on the validity of the structure-mapping theory itself. We feel that while the specific concepts chosen in the course of authoring pre-formal representations of 372 strategies can rightly be questioned, the scope of these concepts as a whole cannot. The evidence provided by terms used in strategy representations suggests that the scope of mental representations that may support Theory of Mind abilities includes concepts for both the mental states of people and of the cognitive processes that they employ.

This evidence of process-oriented mental representations does not by itself provide support for either of the two prominent Theory of Mind theories (Theory Theory and Simulation Theory). However, it does have relevance to how proponents of these theories proceed to produce more detailed, even computational, process models of Theory of Mind abilities.

Proponents of the Theory Theory should view these representation areas as a catalog of the component theories that will be necessary to specify a complete folk psychology, in much the same way that artificial intelligence researchers have attempted to define the component theories of naïve physics (Hayes, 1985). If the two endeavors are indeed similar, then terms like those that comprise the representational areas listed here will appear as notations (predicates or otherwise) in formal axiomatic theories that could drive deductive reasoning. Because breadth of component theories for a full folk psychology appears to be at least as rich as those in naïve physics, we would expect that the same methodological problems in specifying these theories would prohibit progress (see Davis, 1998). As folk psychology has received little attention within the artificial intelligence community as compared with naïve physics, few axiomatic theories exist today for the majority of the representational areas that are listed in this paper.

While axiomatic theories have not been forthcoming, most of these representational areas have been extensively studied as cognitive processes. Cognitive science and artificial intelligence researchers have constructed an enormous number of computational models in support of our theoretical understandings, with one notable exception of Theory of Mind reasoning itself (representational area 10, Other Agent Reasoning). For proponents of the Simulation Theory it is this set of computational models that will have to be employed in the off-line reasoning that allows a person to perform Theory of Mind tasks. Evidence of mental representations that correspond to these processes could suggest that there is a representational interface to support off-line reasoning. That is, terms like those in each of the representational areas could be viewed as a vocabulary for expressing inputs (e.g. commands and arguments) to these processes as well as their outputs (e.g. inferences). Further agreement within the cognitive modeling community concerning inputs and outputs could potentially promote the development of more modular computational theories, facilitating the integration of models that will be necessary in providing a process account of Simulation Theory, among others.

While the investigation of Theory of Mind representations may affect the theoretical debate only in the long run, its utility in linguistic studies of Theory of Mind language use may be more direct. Specifically, the identification of these representation areas – as well as the specific terms that appear in each – may be valuable in identifying a broader lexicon for use in the analysis of language data. For example, a process-oriented term such as *suspend goal* (from area 14, Goal Management) is expressible in a wide variety of ways in English, as in "Let's *put it off for now*" or "I'll *come back to it later*" where the direct object in both statements is the suspended goal. Compiling word and phrase lexicons for each of the terms in these representational areas could provide enough coverage over a language to facilitate more automated text analysis approaches, which in turn could greatly scale up the amount of linguistic data that could be analyzed.

Conclusions

While there has been great interest in understanding the Theory of Mind abilities of people, the experimental and theoretical approaches to this problem have largely avoided issues concerning the *contents* of representations employed in this class of reasoning. In

this paper, we have argued that progress in our understanding of a different cognitive process – that of analogical reasoning – provides us with a tool that can be used to investigate these representations in a new way. The curious nature of commonsense strategies, in accounting for analogies in planning domains and including references to mental states and processes, makes them a particularly important subject of analysis. By conducting a large-scale analysis of strategies from many planning domains, authoring pre-formal representations for each, we have improved the understanding of the scope of representations that would be available in support of Theory of Mind reasoning abilities. In addition to the mental state concepts that have traditionally been discussed in Theory of Mind research, this investigation suggests that rich representations of mental processes are also part of our representations.

References

Baron-Cohen, S. (2000) Theory of mind and autism: a fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Cohen, P. and Levesque, H. (1990) Intention is Choice with Commitment. *Artificial Intelligence* 42, 213-261.

Corcoran, R (2001) Theory of Mind in Schizophrenia. In: D. Penn and P. Corrigan (Eds.) *Social Cognition in Schizophrenia*. APA.

Davis, E. (1998) The Naïve Physics Perplex, *AI Magazine*, Winter 1998.

Dyer, J., Shatz, M., & Wellman, H. (2000) Young children's storybooks as a source of mental state information. *Cognitive Development* 15, 17-37.

Frith, C. & Frith, U. (2000) The physiological basis of theory of mind: functional neuroimaging studies. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Gentner, D. (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7, 155-170.

Gentner, D. & Markman, A. (1997) Structure mapping in analogy and similarity. *American Psychologist* 52, 45-56.

Goldman, A. (2000) Folk Psychology and Mental Concepts. *Protosociology* 14, 4-25.

Gopnik, A. & Meltzoff, A. (1997). Words, thoughts, and theories. Cambridge, Mass.: Bradford, MIT Press.

Gordon, A. (2001a) Strategies in Analogous Planning Cases. In J. Moore & K. Stenning (eds.) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gordon, A. (2001b) The Representational Requirements of Strategic Planning. Fifth symposium on Logical Formalizations of Commonsense Reasoning. (http://www.cs.nyu.edu/faculty/davise/commonsense01/).

Happé, F., Brownell, H., & Winner, E. (1998) The getting of wisdom: Theory of mind in old age. *Developmental Psychology*, 34 (2), 358-362.

Happé, F., Brownell, H., & Winner, E. (1999) Acquired 'theory of mind' impairments following stroke. *Cognition* 70, 211-240.

Harris, P. (1996) Desires, beliefs, and language. In P. Carruthers and P. Smith (Eds.) *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Hayes, P. (1985) The second naïve physics manifesto. In J. Hobbs & B. Moore, Formal Theories of the Commonsense World. Ablex Publishing.

Nichols, S. & Stich, S. (2000) A cognitive theory of pretense. *Cognition* 74, 115-147.

Nichols, S. & Stich, S. (forthcoming) How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In Q. Smith and A. Jokic (Eds.) *Consciousness: New Philosophical Essays*, Oxford University Press.

Ortiz, C. (1999) Introspective and elaborative processes in rational agents. *Annals of Mathematics and Artificial Intelligence* 25, 1-34.

Wellman, H.M., & Lagattuta, K. H. (2000). Developing understandings of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

A probabilistic approach to semantic representation

Thomas L. Griffiths & Mark Steyvers
{gruffydd,msteyver}@psych.stanford.edu
Department of Psychology
Stanford University
Stanford, CA 94305-2130 USA

Abstract

Semantic networks produced from human data have statistical properties that cannot be easily captured by spatial representations. We explore a probabilistic approach to semantic representation that explicitly models the probability with which words occur in different contexts, and hence captures the probabilistic relationships between words. We show that this representation has statistical properties consistent with the large-scale structure of semantic networks constructed by humans, and trace the origins of these properties.

Contemporary accounts of semantic representation suggest that we should consider words to be either points in a high-dimensional space (eg. Landauer & Dumais, 1997), or interconnected nodes in a semantic network (eg. Collins & Loftus, 1975). Both of these ways of representing semantic information provide important insights, but also have shortcomings. Spatial approaches illustrate the importance of dimensionality reduction and employ simple algorithms, but are limited by Euclidean geometry. Semantic networks are less constrained, but their graphical structure lacks a clear interpretation.

In this paper, we view the function of associative semantic memory to be efficient prediction of the concepts likely to occur in a given context. We take a probabilistic approach to this problem, modeling documents as expressing information related to a small number of topics (cf. Blei, Ng, & Jordan, 2002). The topics of a language can then be learned from the words that occur in different documents. We illustrate that the large-scale structure of this representation has statistical properties that correspond well with those of semantic networks produced by humans, and trace this to the fidelity with which it reproduces the natural statistics of language.

Approaches to semantic representation

Spatial approaches Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) is a procedure for finding a high-dimensional spatial representation for words. LSA uses singular value decomposition to factorize a word-document co-occurrence matrix. An approximation to the original matrix can be obtained by choosing to use less singular values than its rank. One component of this approximation is a matrix that gives each word a location in a high dimensional space. Distances in this space are predictive in many tasks that require the use of semantic information. Performance is best for approximations that used less singular values than the rank of the matrix, illustrating that reducing the dimensionality of the representation can reduce the effects of statistical noise and increase efficiency.

While the methods behind LSA were novel in scale and subject, the suggestion that similarity relates to distance in psychological space has a long history (Shepard, 1957). Critics have argued that human similarity judgments do not satisfy the properties of Euclidean distances, such as symmetry or the triangle inequality. Tversky and Hutchinson (1986) pointed out that Euclidean geometry places strong constraints on the number of points to which a particular point can be the nearest neighbor, and that many sets of stimuli violate these constraints. The number of nearest neighbors in similarity judgments has an analogue in semantic representation. Nelson, McEvoy and Schreiber (1999) had people perform a word association task in which they named an associated word in response to a set of target words. Steyvers and Tenenbaum (submitted) noted that the number of unique words produced for each target follows a power law distribution: if k is the number of words, $P(k) \propto k^{\gamma}$. For reasons similar to those of Tversky and Hutchinson, it is difficult to produce a power law distribution by thresholding cosine or distance in Euclidean space. This is shown in Figure 1. Power law distributions appear linear in log-log coordinates. LSA produces curved log-log plots, more consistent with an exponential distribution.

Semantic networks Semantic networks were proposed by Collins and Quillian (1969) as a means of storing semantic knowledge. The original networks were inheritance hierarchies, but Collins and Loftus (1975) generalized the notion to cover arbitrary graphical structures. The interpretation of this graphical structure is vague, being based on connecting nodes that "activate" one another. Steyvers and Tenenbaum (submitted) constructed a semantic network from the word association norms of Nelson et

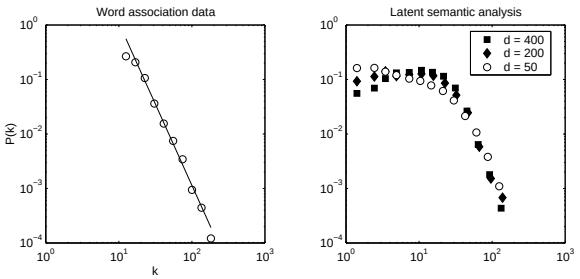


Figure 1: The left panel shows the distribution of the number of associates named for each target in a word association task. The right shows the distribution of the number of words above a cosine threshold for each target in LSA spaces of dimension d, where the threshold was chosen to match the empirical mean.

al. (1999), connecting words that were produced as responses to one another. In such a semantic network, the number of associates of a word becomes the number of edges of a node, termed its "degree". Steyvers and Tenenbaum found that the resulting graph had the statistical properties of "small world" graphs, of which a power law degree distribution is a feature (Barabasi & Albert, 1999).

The fact that semantic networks can display these properties reflects their flexibility, but there is no indication that the same properties would emerge if such a representation were learned rather than constructed by hand. In the remainder of the paper, we present a probabilistic method for learning a representation from word-document co-occurences that reproduces some of the large-scale statistical properties of semantic networks constructed by humans.

A probabilistic approach

Anderson's (1990) rational analysis of memory and categorization takes prediction as the goal of the learner. Analogously, we can view the function of associative semantic memory to be the prediction of which words are likely to arise in a given context, ensuring that relevant semantic information is available when needed. Simply tracking how often words occur in different contexts is insufficient for this task, as it gives no grounds for generalization. If we assume that the words that occur in different contexts are drawn from T topics, and each topic can be characterized by a probability distribution over words, then we can model the distribution over words in any one context as a mixture of those topics

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and $P(w_i|z_i = j)$ is the probability of the ith word under the jth topic. The words likely to be used in a new context can be determined by estimating the distribution over topics for that context, corresponding to $P(z_i)$.

Intuitively, $P(w|z = j)$ indicates which words are important to a topic, while $P(z)$ is the prevalence of those topics within a document. For example, imagine a world where the only topics of conversation are love and research. In such a world we could capture the probability distribution over words with two topics, one relating to love and the other to research. The difference between the topics would be reflected in $P(w|z = j)$: the love topic would give high probability to words like joy, pleasure, or heart, while the research topic would give high probability to words like science, mathematics, or experiment. Whether a particular conversation concerns love, research, or the love of research would depend upon the distribution over topics, $P(z)$, for that particular context.

Formally, our data consist of words $\mathbf{w} = \{w_1, \ldots, w_n\}$, where each w_i belongs to some document d_i, as in a word-document co-occurrence matrix. For each document we have a multinomial distribution over the T topics, with parameters $\theta^{(d_i)}$, so for a word in document d_i, $P(z_i = j) = \theta_j^{(d_i)}$. The jth topic is represented by a multinomial distribution over the W words in the vocabulary, with parameters $\phi^{(j)}$, so $P(w_i|z_i = j) = \phi_{w_i}^{(j)}$. To make predictions about new documents, we need to assume a prior distribution on the parameters $\theta^{(d_i)}$. The Dirichlet distribution is conjugate to the multinomial, so we take a Dirichlet prior on $\theta^{(d_i)}$.

This probability model is a generative model: it gives a procedure by which documents can be generated. First we pick a distribution over topics from the prior on θ, which determines $P(z_i)$ for words in that document. Each time we want to add a word to the document, we pick a topic according to this distribution, and then pick a word from that topic according to $P(w_i|z_i = j)$, which is determined by $\phi^{(j)}$. This generative model was introduced by Blei et al. (2002), improving upon Hofmann's (1999) probabilistic Latent Semantic Indexing (pLSI). Using few topics to represent the probability distributions over words in many documents is a form of dimensionality reduction, and has an elegant geometric interpretation (see Hofmann, 1999).

This approach models the frequencies in a word-document co-occurrence matrix as arising from a simple statistical process, and explores the parameters of this process. The result is not an explicit representation of words, but a representation that captures the probabilistic relationships among words. This representation is exactly what is required for predicting when words are likely to be used. Because we treat the entries in a word-document co-occurrence matrix as frequencies, the representation developed from this information is sensitive to the natural statistics of language. Using a generative model, in which we articulate the assumptions about how the data were generated, ensures that we are

able to form predictions about which words might be seen in a new document.

Blei et al. (2002) gave an algorithm for finding estimates of $\phi^{(j)}$ and the hyperparameters of the prior on $\theta^{(d_i)}$ that correspond to local maxima of the likelihood, terming this procedure Latent Dirichlet Allocation (LDA). Here, we use a symmetric Dirichlet(α) prior on $\theta^{(d_i)}$ for all documents, a symmetric Dirichlet(β) prior on $\phi^{(j)}$ for all topics, and Markov chain Monte Carlo for inference. An advantage of this approach is that we do not need to explicitly represent the model parameters: we can integrate out θ and ϕ, defining model simply in terms of the assignments of words to topics indicated by the z_i. [1]

Markov chain Monte Carlo is a procedure for obtaining samples from complicated probability distributions, allowing a Markov chain to converge to the target distribution and then drawing samples from the Markov chain (see Gilks, Richardson & Spiegelhalter, 1996). Each state of the chain is an assignment of values to the variables being sampled, and transitions between states follow a simple rule. We use Gibbs sampling, where the next state is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data. We will sample only the assignments of words to topics, z_i. The conditional posterior distribution for z_i is given by

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha} \quad (1)$$

where $\mathbf{z}_{-i}$ is the assignment of all z_k such that $k \neq i$, and $n_{-i,j}^{(w_i)}$ is the number of words assigned to topic j that are the same as w_i, $n_{-i,j}^{(\cdot)}$ is the total number of words assigned to topic j, $n_{-i,j}^{(d_i)}$ is the number of words from document d_i assigned to topic j, and $n_{-i,\cdot}^{(d_i)}$ is the total number of words in document d_i, all not counting the assignment of the current word w_i. α, β are free parameters that determine how heavily these empirical distributions are smoothed.

The Monte Carlo algorithm is then straightforward. The z_i are initialized to values between 1 and T, determining the initial state of the Markov chain. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Equation 1. After enough iterations for the chain to approach the target distribution, the current values of the z_i are recorded. Subsequent samples are taken after an appropriate lag, to ensure that their autocorrelation is low. Gibbs sampling is used in each of the following simulations in order to explore the consequences of this probabilistic approach.

[1] A detailed derivation of the conditional probabilities used here is given in a technical report available at http://www-psych.stanford.edu/~gruffydd/cogsci02/lda.ps

Simulation 1: Learning topics with Gibbs sampling

The aim of this simulation was to establish the statistical properties of the sampling procedure and to qualitatively assess its results, as well as to demonstrate that complexities of language like polysemy and behavioral asymmetries are naturally captured by our approach. We took a subset of the TASA corpus (Landauer, Foltz, & Laham, 1998), using the 4544 words that occurred both in the word association norm data and at least 10 times in the complete corpus, together with a random set of 5000 documents. The total number of words occurring in this subset of the corpus, and hence the number of z_i to be sampled, was $n = 395853$. We set the parameters of the model so that 150 topics would be found ($T = 150$), with $\alpha = 0.1$, $\beta = 0.01$.

The initial state of the Markov chain was established with an online learning procedure. Initially, none of the w_i were assigned to topics. The z_i were then sequentially drawn according to Equation 1 where each of the frequencies involved, as well as W, reflected only the words that had already been assigned to topics.[2] This initialization procedure was used because it was hoped that it would start the chain at a point close to the true posterior distribution, speeding convergence.

Ten runs of the Markov chain were conducted, each lasting for 2000 iterations. On each iteration we computed the average number of topics to which a word was assigned, $\langle k \rangle$, which was used to evaluate the sampling procedure for large scale properties of the representation. Specifically, we were concerned about convergence and the autocorrelation between samples. The rate of convergence was assessed using the Gelman-Rubin statistic $\hat{R}$, which remained below 1.2 after 25 iterations. The autocorrelation was less than 0.1 after a lag of 50 iterations.

A single sample was drawn from the first run of the Markov chain after 2000 iterations. A subset of the 150 topics found by the model are displayed in Table 1, with words in each column corresponding to one topic, and ordered by the frequency with which they were assigned to that topic. The topics displayed are not necessarily the most interpretable found by the model, having been selected only to highlight the way in which polysemy is naturally dealt with by this representation. More than 90 of the 150 topics appeared to have coherent interpretations.[3]

The word association data of Nelson et al. (1999) contain a number of asymmetries – cases where people were more likely to produce one word in response to the other. Such asymmetries are hard to ac-

[2] Random numbers used in all simulations were generated with the Mersenne Twister, which has an extremely deep period (Matsumoto & Nishimura, 1998).

[3] The 20 most frequent words in these topics are listed at http://www-psych.stanford.edu/~gruffydd/cogsci02/topics.txt

COLD	TREES	COLOR	**FIELD**	GAME	ART	BODY	KING	LAW
WINTER	TREE	BLUE	CURRENT	**PLAY**	MUSIC	BLOOD	GREAT	RIGHTS
WEATHER	FOREST	RED	ELECTRIC	BALL	**PLAY**	**HEART**	SON	**COURT**
WARM	LEAVES	**GREEN**	ELECTRICITY	TEAM	**PART**	MUSCLE	LORDS	LAWS
SUMMER	GROUND	LIKE	TWO	PLAYING	SING	FOOD	QUEEN	ACT
SUN	PINE	WHITE	FLOW	GAMES	LIKE	OTHER	EMPEROR	LEGAL
WIND	GRASS	BROWN	WIRE	FOOTBALL	POETRY	BONE	OWN	STATE
SNOW	LONG	BLACK	SWITCH	BASEBALL	BAND	MADE	PALACE	PERSON
HOT	LEAF	YELLOW	TURN	**FIELD**	WORLD	SKIN	DAY	CASE
CLIMATE	CUT	**LIGHT**	BULB	SPORTS	RHYTHM	TISSUE	PRINCE	DECISION
YEAR	WALK	BRIGHT	BATTERY	PLAYER	POEM	MOVE	LADY	CRIME
RAIN	SHORT	DARK	PATH	COACH	SONG	STOMACH	CASTLE	IMPORTANT
DAY	OAK	GRAY	CAN	LIKE	LITERATURE	**PART**	ROYAL	JUSTICE
SPRING	**FALL**	MADE	LOAD	HIT	SAY	OXYGEN	MAN	FREEDOM
LONG	**GREEN**	LITTLE	**LIGHT**	TENNIS	CHARACTER	THIN	MAGIC	ACTION
FALL	FEET	TURN	RADIO	SPORT	AUDIENCE	SYSTEM	**COURT**	OWN
HEAT	TALL	WIDE	MOVE	BASKETBALL	THEATER	CHEST	**HEART**	SET
ICE	GROW	SUN	LOOP	LEAGUE	OWN	TINY	GOLDEN	LAWYER
FEW	WOODS	PURPLE	DEVICE	FUN	KNOWN	FORM	KNIGHT	YEARS
GREAT	WOOD	PINK	DIAGRAM	BAT	TRAGEDY	BEAT	GRACE	FREE

Table 1: Nine topics from the single sample in Simulation 1. Each column shows 20 words from one topic, ordered by the number of times that word was assigned to the topic. Adjacent columns share at least one word. Shared words are shown in boldface, providing some clear examples of polysemy

count for in spatial representations because distance is symmetric. The generative structure of our model allows us to calculate $P(w_2|w_1)$, the probability that the next word seen in a novel context will be w_2, given that the first word was w_1. Since this is a conditional probability, it is inherently asymmetric. The asymmetries in $P(w_2|w_1)$ predict 77.47% of the asymmetries in the word association norms of Nelson et al. (1999), restricted to the 4544 words used in the simulation. These results are driven by word frequency: $P(w_2)$ should be close to $P(w_2|w_1)$, and 77.32% of the asymmetries could be predicted by the frequency of words in this subset of the TASA corpus. The slight improvement in performance came from cases where word frequencies were very similar or polysemy made overall frequency a poor indicator of the frequency of a particular sense of a word.

Bipartite semantic networks

The standard conception of a semantic network is a graph with edges between word nodes. Such a graph is unipartite: there is only one type of node, and those nodes can be interconnected freely. In contrast, bipartite graphs consist of nodes of two types, and only nodes of different types can be connected. We can form a bipartite semantic network by introducing a second class of nodes that mediate the connections between words. One example of such a network is a thesaurus: words are organized topically, and a bipartite graph can be formed by connecting words to the topics in which they occur, as illustrated in the left panel of Figure 2.

Steyvers and Tenenbaum (submitted) discovered that bipartite semantic networks constructed by humans, such as that corresponding to Roget's (1911) Thesaurus, share the statistical properties of unipartite semantic networks. In particular, the number of topics in which a word occurs, or the degree of that word in the graph, follows a power law distribution as shown in the right panel of Figure 2. This result is reminiscent of Zipf's (1965) "law of meaning": the number of meanings of a word follows a power law distribution. Zipf's law was established by analyzing dictionary entries, but appears to describe the same property of language.

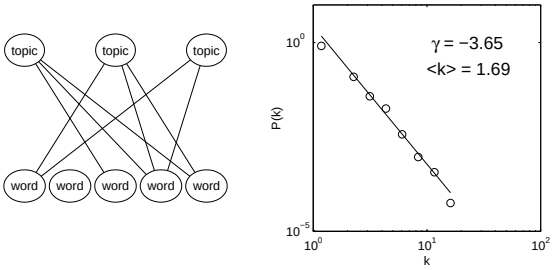


Figure 2: The left panel shows a bipartite semantic network. The right shows the degree distribution a network constructed from Roget's Thesaurus.

Our probabilistic approach specifies a probability distribution over the allocation of words to topics. If we form a bipartite graph by connecting words to the topics in which they occur, we obtain a probability distribution over such graphs. The existence of an edge between a word and a topic indicates that the word has some significant probability of occurring in that topic. In the following simulations, we explore whether the distribution over bipartite graphs resulting from our approach is consistent with the statistical properties of Roget's Thesaurus and Zipf's law of meaning. In particular, we examine whether we obtain structures that have a power law degree distribution.

Simulation 2: Power law degree distributions

We used Gibbs sampling to obtain samples from the posterior distribution of the z_i for two word-document co-occurrence matrices: the matrix with the 4544 words from the word association norms used in Simulation 1, and a second matrix using

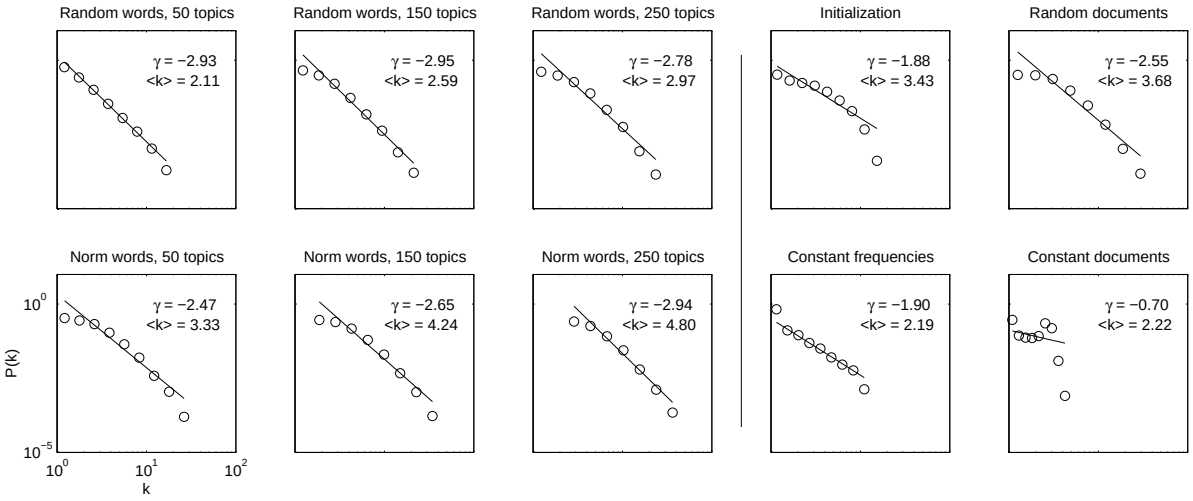


Figure 3: Degree distributions for networks constructed in Simulations 2 and 3. All are on the same axes.

4544 words drawn at random from those occurring at least 10 times in the TASA corpus ($n = 164401$). Both matrices used the same 5000 random documents. For each matrix, 100 samples were taken with $T = 50, 100, 150, 200$ and 250. Since the results seemed unaffected by the number of topics, we will focus on $T = 50, 150, 250$. Ten samples were obtained in each of 10 separate runs with a burn-in of 1000 iterations in which no samples were drawn, and a between-sample lag of 100 iterations.

For each sample, a bipartite semantic network was constructed by connecting words to the topics to which they were assigned. For each network, the degree of each word node was averaged over the 100 samples.[4] The resulting distributions were clearly power-law, as shown in Figure 3. The γ coefficients remained within a small range and were all close to $\gamma = -3.65$ for Roget's Thesaurus. As is to be expected, the average degree increased as more topics were made available, and was generally higher than Roget's. Semantic networks in which edges are added for each assignment tend to be quite densely connected. Sparser networks can be produced by setting a more conservative threshold for the inclusion of an edge, such as multiple assignments of a word to a topic, or exceeding some baseline probability in the distribution represented by that topic.

Our probabilistic approach produces power law degree distributions, in this case indicating that the number of topics to which a word is assigned follows a power law. This result is very similar to the properties of Roget's Thesaurus and Zipf's observations about dictionary definitions. This provides an opportunity to establish the origin of this distribution, to see whether it is a consequence of the modeling approach or a basic property of language.

Simulation 3: Origins of the power law

To investigate the origins of the power law, we first established that our initialization procedure was not responsible for our results. Using $T = 150$ and the matrix with random words, we obtained 100 samples of the degree distribution immediately following initialization. As can be seen in Figure 3, this produced a curved log-log plot and higher values of γ and $\langle k \rangle$ than in Simulation 2.

The remaining analyses employed variants of this co-occurrence matrix, and their results are also presented in Figure 3. The first variant kept word frequency constant, but assigned instances of words to documents at random, disrupting the co-occurrence structure. Interestingly, this appeared to have only a weak effect on the results, although the curvature of the resulting plot did increase. The second variant forced the frequencies of all words to be as close as possible to the median frequency. This was done by dividing all entries in the matrix by the frequency of that word, multiplying by the median frequency, and rounding to the nearest integer. The total number of instances in the resulting matrix was $n = 156891$. This manipulation reduced the average density in the resulting graph considerably, but the distribution still appeared to follow a power law. The third variant held the number of documents in which a word participated constant. Word frequencies were only weakly affected by this manipulation, which spread the instances of each word uniformly over the top five documents in which it occurred

[4] Since power law distributions can be produced by averaging exponentials, we also inspected individual samples to confirm that they had the same characteristics.

and then rounded up to the nearest integer, giving $n = 174615$. Five was the median number of documents in which words occurred, and documents were chosen at random for words below the median. This manipulation had a strong effect on the degree distribution, which was no longer power law, or even monotonically decreasing.

The distribution of the number of topics in which a word participates was strongly affected by the distribution of the number of documents in which a word occurs. Examination of the latter distribution in the TASA corpus revealed that it follows a power law. Our approach produces a power law degree distribution because it accurately captures the natural statistics of these data, even as it constructs a lower-dimensional representation.

General Discussion

We have taken a probabilistic approach to the problem of semantic representation, motivated by considering the function of associative semantic memory. We assume a generative model where the words that occur in each context are chosen from a small number of topics. This approach produces a lower-dimensional representation of a word-document co-occurrence matrix, and explicitly models the frequencies in that matrix as probability distributions. Simulation 1 showed that our approach could extract coherent topics, and naturally deal with issues like polysemy and asymmetries that are hard to account for in spatial representations. In Simulation 2, we showed that this probabilistic approach was also capable of producing representations with a large-scale structure consistent with semantic networks constructed from human data. In particular, the number of topics to which a word was assigned followed a power law distribution, as in Roget's (1911) Thesaurus and Zipf's (1965) law of meaning. In Simulation 3, we discovered that the only manipulation that would remove the power law was altering the number of documents in which words participate, which follows a power law distribution itself.

Steyvers and Tenenbaum (submitted) suggested that power law distributions in language might be traced to some kind of growth process. Our results indicate that this growth process need not be a part of the learning algorithm, if the algorithm is faithful to the statistics of the data. While we were able to establish the origins of the power law distribution in our model, the growth processes described by Steyvers and Tenenbaum might contribute to understanding the origins of the power law distribution in dictionary meanings, thesaurus topics, and the number of documents in which words participate.

The representation learned by our probabilistic approach is not explicitly a representation of words, in which each word might be described by some set of features. Instead, it is a representation of the probabilistic relationships between words, as expressed by their probabilities of arising in different contexts. We can easily compute important statistical quantities from this representation, such as $P(w_2|w_1)$, the probability of w_2 arising in a particular context given that w_1 was observed, and more complicated conditional probabilities. One advantage of an explicitly probabilistic representation is that we gain the opportunity to incorporate this representation into other probabilistic models. In particular, we see great potential for using this kind of representation in understanding the broader phenomena of human memory.

Acknowledgments The authors were supported by a Hackett Studentship and a grant from NTT Communications Sciences laboratory. We thank Tania Lombrozo, Penny Smith and Josh Tenenbaum for comments, and Tom Landauer and Darrell Laham for the TASA corpus. Shawn Cokus wrote the Mersenne Twister code.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ.

Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82, 407-428.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior*, 8, 240-248.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, Suffolk.

Hofmann, T. (1999). Probablistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Matsumoto, M. & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling & Computer Simulation*, 8, 3-30.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1999). The University of South Florida word association norms. *http://www.usf.edu/FreeAssociation*.

Roget, P. (1911). *Roget's Thesaurus of English Words and Phrases*. Available from Project Gutenberg.

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model, relating generalization to distance in psychological space. *Psychometrika*, 22, 325-345.

Steyvers, M. & Tenenbaum, J. B. (submitted). *The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth.*

Tversky, A. & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93, 3-22.

Zipf, G. K. (1965). *Human behavior and the principle of least effort*. Hafner, New York.

Strategic Differences in the Coordination of Different Views of Space

Glenn Gunzelmann (glenng@andrew.cmu.edu)
John R. Anderson (ja+@cmu.edu)
Department of Psychology, Baker Hall 342-C
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Participants were trained to use one of two different strategies in an orientation task, which were based on verbal reports from participants in another experiment. The data provide support for the conclusion that participants in the two training conditions searched the screen differently to complete the task, but that neither group used mental transformations like image rotation. These results have implications for research in this area as well as for conceptualizing how individuals perform such tasks. A comparison of the results from the two strategy conditions is made based on an ACT-R model of one of them. Small differences in how information on the screen is scanned can produce the observed differences in performance.

Introduction

The coordination of different views of space is a fundamental task in human functioning. An everyday example of it involves determining which way to turn at an intersection by using a map. The visual scene presents one view of the space (egocentric), while the map presents an alternative representation (allocentric). In order to accurately decide which way to go, it is necessary to bring these two views of the space into correspondence. Of course, with a physical map it may be possible to actually rotate it to align it with your own orientation. In other situations, mental transformations may need to be done in order to coordinate these views to make accurate decisions.

On a continuum of reasoning about orientation within a space, deciding whether the correct turn is left or right is a fairly straightforward task. Still, research on this issue has shown that it becomes increasingly difficult to perform as a function of the difference in orientation between the two views of space (Shepard and Hurwitz, 1984). The phenomenon bears a strong resemblance to findings in the mental rotation literature (Shepard and Metzler, 1971) where the time needed to determine that two objects are identical increases linearly as a function of the angular disparity between them. These findings have been used to support the conclusion that performance in orientation tasks involves analog mental rotation of mental images. Note, however, that the task of coordinating views of space adds a layer of complexity to the traditional mental rotation task. In a spatial orientation task, the information is presented in two different formats. Thus, deciding whether the visual scene matches the information on the map requires additional reasoning beyond the image transformation.

In an important series of experiments, Hintzman, O'Dell, and Arndt (1981) had participants perform orientation tasks in a variety of ways. In the basic task, participants had to indicate the direction of a target relative to a given orientation. Figure 1 shows the orientation task used in the experiment presented here. In this figure, the left side represents the target field as viewed from a camera (on a plane above the field) and the darkened circle indicates the target. The right side represents a map-view with the target field at the center. The arrow on that side shows the camera's orientation for viewing the target field. Participants are asked to indicate in which cardinal direction the target is located relative to the center of the target field. In the sample trial in Figure 1, the correct response is South. The general finding is that decisions for targets in line with the assumed orientation are made more rapidly, and response times for other targets increase as they depart from the nearest point immediately in front of the viewpoint. Although not explicitly addressed by Hintzman, et al., this increase in response time is not strictly linear. In addition, no evidence was presented in their study about how participants claimed to be performing the task.

In order to investigate what factors influence performance on this task, we asked participants to complete the task and then questioned them as to the manner in which they solved it. While we will not go into detail about this experiment, the data are presented below and bear a strong resemblance to results from similar studies, including Hintzman, et al. (1981). However, by questioning participants after they had completed the experiment, we discovered that participants were using at least two distinct strategies to do the task. Some participants claimed to be implementing a strategy that incorporated imagery and mental rotation to determine correct responses.

However, other participants indicated that they used a different strategy altogether, one that did not depend on mental imagery or mental rotation at all.

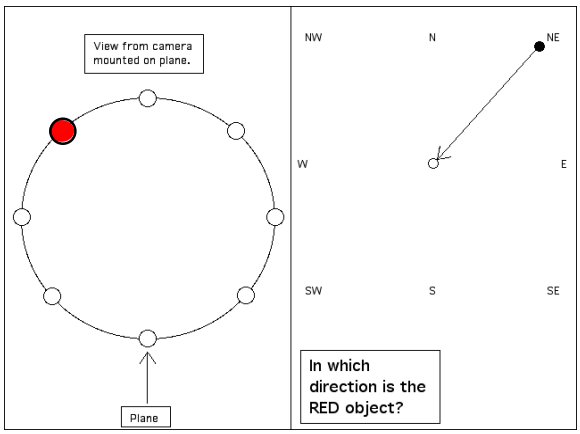


Figure 1: Sample trial for the orientation task.

In the imagery-based strategy, participants reported forming an angle connecting the camera to the target in the camera view with the vertex at the center of the target field (a 135 degree angle in Figure 1). This angle was then mentally transformed to line it up with the position of the camera on the map view. A second group of the participants simply counted around the target field to the target on the camera view (3 in Figure 1), and then counted that number of steps from the camera's position in the appropriate direction on the map view.

Both strategies are equally effective and valid for doing the task, but one depends on mental imagery while the other does not. In addition, the verbal reports indicate that few of the participants treated the task as an orientation task. Rather, the strategies they reported suggest that they treated the task more like a traditional mental rotation experiment. That is, participants effectively eliminated the added level of complexity introduced by having different representations of the information by selecting strategies that bypassed the need to consider them. This finding casts some doubt on some previous explanations for performance on similar tasks. The experiment presented in detail below was conducted to further investigate the implications of these strategies, by training participants to perform the task using either the imagery-based "angle" strategy or the more analytic "counting" strategy.

Experiment

Previous research aimed at addressing performance on tasks similar to the one presented here have based explanations largely on mental imagery and mental rotation (Shepard and Hurwitz, 1984; Hintzman, et al., 1981), though Hintzman et al. do consider a sequential scanning explanation. However, all these explanations fail to account for some of the more subtle aspects of the data and ignore the potential for different approaches to the task. The experiment presented here examines strategic differences in an orientation task that is similar to those used by Hintzman, et al (1981).

If the strategies were to be implemented according to the descriptions provided to participants, there should be clear differences in performance between the two strategy conditions. For the counting strategy, the position of the target relative to the camera should greatly influence response times. That is, response time should increase linearly as a function of the amount of the counting that needs to be done. However, the location of the camera in the map view should have no impact on performance, since the strategy can be implemented identically regardless of the camera's position in the map view. The angle strategy makes the opposite set of predictions. Response time should be unaffected by the target's location, since the angle to be formed is similar in complexity regardless of the target's position in the camera view. However, the degree of rotation that needs to be done depends on the camera's position in the map view, suggesting that response times should increase linearly as a function of the camera's position relative to the bottom of the screen.

Method

The experimental task was based both on the experimental task used by Hintzman, et al (1981) as well as on an unmanned air vehicle (UAV) flight simulator used by the Air Force for training UAV pilots (see Gugerty, et al, 2000; Figure 1). The display consisted of two static views, an egocentric "camera" view of a target field, and an allocentric "map" view. The target field was in the center of the map view, and the perspective of the camera was identified with an arrow (the right half of Figure 1). The target field was a circle, containing eight objects equally spaced at 45 degree intervals on the circle (the left portion of Figure 1), with one of them highlighted in red to identify it as the target. Participants were asked to indicate in which cardinal direction the target was located relative to the target field's center. Responses were made using the number pad on the keyboard.

After being introduced to the experimental task, participants were trained to complete the task using either the angle or counting strategy (n=16 per condition). They first read a brief description of the strategy, and then were shown how the strategy applied to a sample trial. After that, participants completed 16 paper-based practice trials in random order. In these practice trials, participants were asked to explicitly demonstrate use of the strategy they had been taught by labeling them appropriately based on the strategy they

had been taught. In the counting strategy condition, participants were taught to use positive numbers for targets on the left (clockwise from the camera), and negative numbers for targets on the right (counterclockwise). Participants in angle strategy condition were instructed to note the direction in which the angle "opened". Feedback was given on each of the practice trials by the experimenter.

After training, participants completed 4 blocks of trials on the computer. Each block included all 64 possible trials in random order. A dropout procedure was used such that if an error was made on one of the trials it was presented again later in the block. During this phase of the experiment feedback was still given after each trial, including what the correct answer was in cases where participants made an error.

Results

The results for the original experiment and the two training conditions in this experiment are presented in Figures 2 and 3. In Figure 2, response time is plotted as a function of the target's clockwise deviation from the camera. The numbers correspond to the measure of the clockwise arc from the camera position to the target on the target field in the camera view. In Figure 3, the data are presented as a function of the location of the camera relative to the target field in the map view. In the sample trial shown in Figure 1, the target angle is 135 and the camera's location is NE. One aspect of the data that should be immediately apparent from these graphs is that performance was symmetrical in terms of left and right positions of both the camera and the target. In addition, response times were somewhat faster in this experiment than in the first one. This may be a result of the training given in this experiment, which participants in the first experiment did not receive.

Finally, the training conditions used in this experiment seem to separate out two components of the data from the first experiment in terms of the effect of the target's position. Specifically, data produced by participants using the counting strategy increase linearly with the target's angular deviation from the camera. In contrast, the data produced by participants using the angle strategy show a scalloped effect, with no difference between 45 and 90 degrees (or 315 and 270 degrees). The data from the original experiment show evidence of a combination of both trends. This suggests that averaging data over all participants may not provide a complete story of the effects in this task.

At the highest level of abstraction, there was no main effect of strategy condition in average response time, $F(1,210)=0.233, p=.63$, suggesting that at a global level both strategies were equally effective for completing the task. One has to be struck by the overall similarity of the results between the two strategy conditions and their close relation to the results from the first study, given that participants were taught quite different ways of doing the task. Despite the overall similarity, there was a significant interaction between the strategies and the particular target angle, $F(7,210)=3.534, p<.02$, as well as between the strategies and the camera angle, $F(7,210)=3.810, p=.01$. Looking at Figure 2, response times were higher for participants using the angle strategy when the target was directly in front of the camera or when it was 45 degrees to the right or left. In terms of camera angle, Figure 3 shows that participants trained to use the angle strategy exhibit relatively longer latencies when the camera is located in a northerly position.

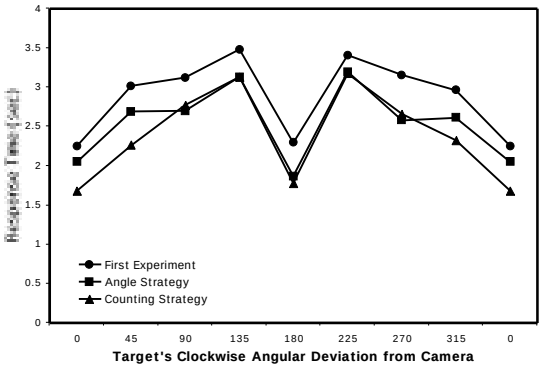


Figure 2: Response time (sec) as a function of the target's position.

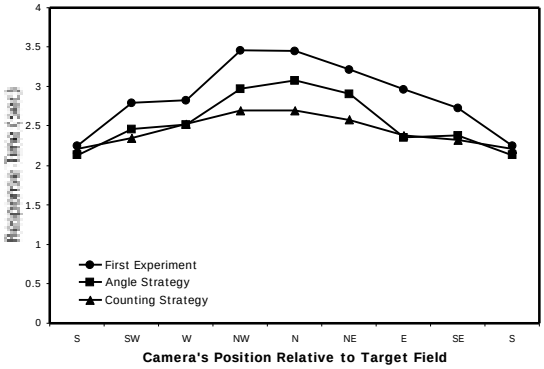


Figure 3: Response time (sec) as a function of the camera's position.

Discussion

Based upon the data, it is clear that participants were not executing the strategies precisely according to the instructions provided. In fact, only one of the predictions is clearly borne out in the data. Specifically, response times increased linearly as a function of the extent of counting for participants trained to use the counting strategy. However, these participants still

showed a small effect of the camera's location. In addition, data from participants trained to use the angle strategy showed a discontinuous effect of both the camera's position and the target's relative position.

The most curious result is the effect of the target's position relative to the camera in the angle strategy. That is, the description of the angle strategy predicts no increase in response time as a function of the target's location. However, an increase does occur, and it is complicated by the discontinuity at 90 and 270 degrees. This finding, in particular, casts doubt on the claim that participants were using mental rotation at all in performing this task. In particular, it is hard to imagine how an imagery-based strategy can account for this particular effect without resorting to specialized mechanisms relating to imagining and/or manipulating 90-degree angles. Research does suggest that cognitive representations of space tend to distort angles to be closer to 90 degrees (Glicksohn, 1984), and also indicates that horizontal and vertical lines are preferred in visual perception (45 and 135 degree angles involve oblique angles; Cecala and Garner, 1986). Still, it is not clear how this should have such a large impact on the ability to manipulate or create mental images of angles of various sizes. A more likely explanation is that the differences in performance between the two strategy conditions arise from small differences in how the screen was scanned by participants as a result of their training, rather than because of differences in higher-level cognitive operations on the information.

In the counting strategy, the linearity of the target-position effect suggests that participants were indeed counting from the camera's position to find the target. The small effect of the camera's position, however, indicates that the strategy was not being implemented exactly according to the instructions. We believe that participants encoded the location of the target as being to the "left" or "right" of the camera, rather than as "clockwise" or "counterclockwise". While this is a small difference in encoding, it does have implications for locating the target on the map view. If a target location is encoded as clockwise, the map view can be scanned in a clockwise direction regardless of where the camera is located. However, if the location of the target is represented as "left" instead, the correct scanning direction is "right" when scanning from NW, N, or NE. So, whenever participants search the screen from one of these locations, extra cognitive steps are needed to make sure that the screen is scanned in the appropriate direction.

An example should clarify how we believe the counting strategy was implemented by participants. For this purpose, consider the trial presented in Figure 1. We believe that counting participants would begin this trial by locating the target on the camera view and encoding it as "3-left". At that point, they would find the camera's location on the map view. Since the camera is located at NE, the correct search direction is actually "right", so an extra operation is needed to convert the direction of scanning. Then, the screen can be scanned to locate East, and the count can then be incremented. Then, Southeast can be found, and the count incremented again, followed by South and the final increment in the count sequence. At this point, participants have located the answer and can issue their response by pressing the "2" key on the number pad (keys were assigned to correspond to the layout of cardinal locations on the screen).

Given that participants using this strategy produced data that were largely in line with predictions and the results were similar to the other condition, we decided to develop a model for the counting strategy. This is a first step to an overall model for the task, which will involve some mixture of strategies.

ACT-R Model

The ACT-R theory (Anderson and Lebiere, 1998) provides an architecture in which the proposed mechanisms can be implemented to determine how well they fit with the data. In addition, ACT-R now incorporates a theory of perceptual-motor action, allowing it to interact directly with the experimental software (Byrne and Anderson, 1998). In this way, an ACT-R model can participate in the experiment exactly as though it were a participant by gathering information from the screen using visual perception, operating on that information within its cognitive system, and issuing a response by sending commands to its motor module. This integration means that all aspects of performance are considered in the model's performance.

Model Design

There is certainly a large degree of overlap between the two strategy conditions. In particular, the details of gathering information and issuing responses in the task are assumed to be largely the same for both strategies. Thus, by understanding how participants executed one of the strategies it will be easier to understand how participants in the other condition may have performed the task. Toward that end, a model of the counting strategy has been implemented and is described next. In the conclusion, we will describe how we believe the behavior of participants trained to use the angle strategy may have differed to produce the observed results.

When a new trial is presented to the model, its first action is to search for the location of a red object on the left side of the screen. Its location is encoded as being left or right of the camera and as an integer value from 0 to 4 to define its distance from the camera. Then, the model finds the location of the camera on the map view and shifts its attention to that location.

Since it is hypothesized that the location of the target is encoded as left or right rather than clockwise or counterclockwise, the model needs to alter its scanning direction when the camera is in the NW, N, or NE positions. Once the appropriate scanning direction is selected, the model finds the nearest cardinal direction to the camera and increments its count. This process is repeated until it has incremented the count the prescribed number of times. At that point, the current cardinal location is encoded and mapped to a response on the number pad. Finally, the model issues a response by sending a command to press the correct key.

Based on verbal reports from participants, there were a couple of exceptions to this operation. First, when the target was located in line with the camera, participants reported that they did not bother to count. Rather, for target positions of 0 degrees they simply responded with whatever position the camera was in, and for target positions of 180 degrees they responded with the cardinal direction directly opposite the camera's position. The other instance where the strategies were not used was when the position of the camera was S. In this case, participants reported that they went directly from the target's location on the camera view to a response. In response to these verbal reports, these special cases were implemented in the model. These reports also correspond to data presented in previous studies (e.g., Hintzman, et al., 1981).

Model Performance

The model's performance using the counting strategy compared to the data in the two conditions is presented in Figure 4. As can be seen, it makes accurate predictions for response times for both conditions in both aggregations of the data (correlation = .98, mean deviation = .11 seconds). The model performs the task in exactly the way we believe participants are doing the task. That is, the model incorporates all of the perceptual, motor, and cognitive steps that humans would need to go through to do the task. Based on this completeness, we feel that the model captures all of the relevant aspects of participant performance.

The linear increase in response time as a function of target location is produced in the model by the simple act of counting and scanning cardinal locations sequentially. For target angles of 45 and 315 degrees, one step is counted, for 90 and 270 degrees this is done twice, and for 135 and 225 degrees there are three cycles. The small effect of the camera's position results from the left/right encoding of the target position. As described above, this creates the need to perform extra cognitive operations to switch the scanning direction at any point when searching from NW, N, and NE, thus increasing response times in trials where those situations arise. This evaluation occurs at each step in the search process. So, each time the model searches for the next cardinal location, it determines whether or not the encoded direction of the target is the correct search direction, and then alters the search direction when necessary.

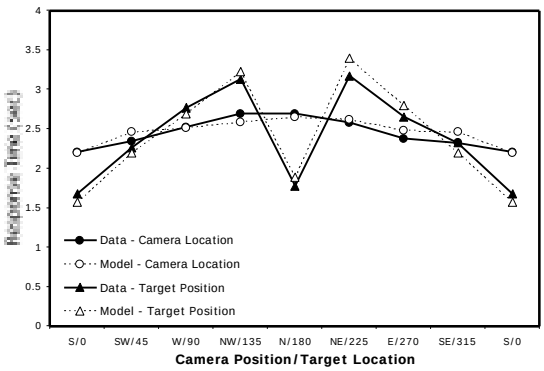


Figure 4. performance of the model of the counting strategy compared to participants' data.

In terms of the overall qualitative pattern of data, the performance of the model is parameter-free. By constructing a model that really does the task, its performance is highly constrained at this level. The parameters serve only to adjust the magnitude of the effects. First, retrievals from memory are an important aspect of the model's operation. The model retrieves various facts from memory as it performs the task, including counting sequences for the counting process, associations between cardinal directions and number keys for making responses, and information about cardinal directions for guiding the search and problem solving process. In this model, the time to perform these retrievals was set to .05 (seconds). The only other parameter that was explicitly set in this model is the execution time for the production that encodes the target's location on the camera view. This value was set to .7 (seconds) and impacts all conditions similarly. The remaining parameters all reflect default perceptual-motor parameters in ACT-R/PM (Byrne and Anderson, 1998). The model's source code is available online at http://act.psy.cmu.edu/.

Conclusions

The experiment and model presented here provide an alternative view of findings in the area of spatial cognition concerning how participants perform orientation tasks. There are two basic questions to answer. First, are participants actually performing an orientation task in these studies? The participants in this experiment were clearly not treating this task as a traditional orientation task where two distinct representations of spatial information are brought into correspondence. Rather, much of the complexity was

eliminated by implementing strategies that avoided this aspect of the task. It is unclear whether similar strategic choices can achieve the same effect in more realistic orientation tasks (e.g., Gugerty, et al., 2000).

The other basic question to ask based upon these results is whether participants use mental imagery in performing the task. If they do, it is important to investigate how such cognitive abilities are applied in these tasks. If not, the question becomes what mechanisms are responsible for participant performance on these sorts of tasks. Based on the data presented here, it appears that participants assumed a more analytic approach to the task, simply scanning the screen in a systematic way to determine the correct answer. These findings also illustrate that there is variability in how participants approach virtually any task, and these variations have implications for performance.

The model shows that we can reproduce much of the qualitative form of the results in this task by implementing a strategy that involves systematically scanning the information on the screen. Moreover, this strategy corresponds to what some participants spontaneously report. However, what about the other participants who spontaneously report an angle strategy? We believe that they may be just engaging in a variant of the implemented scanning strategy, which explains why their behavior is so similar to the participants who were counting. More specifically, we believe that implementing the angle strategy involves such differences as looking at more of the information on the camera view but not systematically looking at the intermediate points between the camera and target on the map view. Both of these differences could be produced by the different training conditions in the experiment. We are currently implementing a model which incorporates such a variant of the scanning strategy and doing an eye movement study to see if we can find evidence for the hypothesized scanning patterns.

Basically our proposal is that participants prefer to process the information given on the screen rather than transform an internal image of this information. This aversion for mental transformations is consistent with the results of Kirsh & Maglio (1994) who found that people prefer to rotate objects on the Tetris screen rather than rotate them in their head. We suggest that some results attributed to mental rotation like those in this task may reflect the operation of some other process like the scanning in the counting strategy that we have implemented. While Hintzman, et al. (1981) considered sequential scanning as an alternative explanation to mental rotation, they did not consider the possibility of strategic differences in the scanning process. The results presented here demonstrate that such strategic differences exist and that some scanning strategies can result in data that approximately match predictions based on imagery and mental rotation. In addition, participants trained to use mental imagery produced data that does not fit with the imagery account. An evaluation of the model for the counting strategy suggests that small differences in encoding and visual scanning can account for the differences found in the angle strategy. These findings suggest that mental rotation may not provide a full account of human performance in orientation tasks.

Acknowledgements

The research reported here was supported by grant number F49620-99-1-0086 from AFOSR.

References

Anderson, J. R., & Lebiere, C. L. (1998). *The atomic components of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Byrne, M. D. & Anderson, J. R. (1998). Perception and action. In J. R. Anderson, & C. Lebiere (Eds.). *The atomic components of thought* (167-200). Mahwah, NJ: Lawrence Erlbaum.

Cecala, A. J., & Garner, W. R. (1986). Internal frame of reference as a determinant of the oblique effect. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 314-323.

Glicksohn, J. (1994). Rotation, orientation, and cognitive mapping. *American Journal of Psychology, 107,* 39-51.

Gugerty, L., deBoom, D., Jenkins, J. C., & Morley, R. (2000). Keeping north in mind: How navigators reason about cardinal directions. In *Proceedings of the Human Factors and Ergonomics Society 2000 Congress* (pp. I148-I151). Santa Monica, CA: Human Factors and Ergonomics Society.

Hintzman, D. L., O'Dell, C. S., & Arndt, D. R. (1981). Orientation in cognitive maps. *Cognitive Psychology, 13,* 149-206.

Huttenlocher, J., & Presson, C. C. (1979). The coding and transformation of spatial information. *Cognitive Psychology, 11,* 375-394.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 513-549.

Presson, C. C. (1982). Strategies in spatial reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 243-251.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three dimensional objects. *Science, 171,* 701-703.

Shepard, R. N., & Hurwitz, S. (1984). Upward direction, mental rotation, and discrimination of left and right turns in maps. *Cognition, 18,* 161-193.

Understanding Similarity in Choice Behavior: A Connectionist Model

Frank Y. Guo (fyguo@ucla.edu)
UCLA, Department of Psychology, 405 Hilgard Ave.
Los Angeles, CA 90095-1563, USA

Keith J. Holyoak (holyoak@psych.ucla.edu)
UCLA, Department of Psychology, 405 Hilgard Ave.
Los Angeles, CA 90095-1563, USA

Abstract

Classical choice theories assume choice behavior is based on value maximization computed over the entire choice set. However, empirical evidence has revealed violations of axioms of rational choice that cannot be explained by value maximization. We argue that choice behavior can be reconceptualized as value maximization constrained by categorization processes, and describe a neural network model developed to account for key empirical findings. The model simulates two important phenomena that have been construed as irrational choice behavior, namely, the similarity effect and the attraction effect. We argue that there are important commonalities among choice behavior, categorization and perception.

Introduction

Many axiomatic theories of choice behavior are based on the assumption that decision making is based on a process of value maximization performed over all attributes (c. f., Tversky & Simonson, 1993). However, empirical evidence has demonstrated that axioms of rational decision making are often violated in choice behavior, and value maximization alone is unable to explain these violations. Recently, an alternative perspective that is concerned with the relations between similarity processes and decision processes has been proposed to conceptualize choice behavior and to understand violations of rational decision making (Medin, Goldstone, & Markman, 1995). That view has been embodied in a comprehensive computational model of choice behavior (Roe, Busemeyer, & Townsend, 2001).

In the spirit of this alternative perspective, we have developed a connectionist model to account for two key violations of rational choice, namely, the similarity effect and the attraction effect. Both of these phenomena involve adding a third alternative (decoy) to a choice set of two options, thereby leading to inconsistency of choice. If the decoy is similar and competitive (two alternatives are competitive when their additive utilities are almost identical to each other) to one of the original options, then the addition of the decoy decreases the choice probability of that option. This phenomenon is called the similarity effect (Tversky, 1972). If the decoy is similar to and dominated by one of the two original alternatives but not the other, then the addition of the decoy increases the choice probability of the dominant option more than the other alternative. This phenomenon is referred to as the attraction effect (Huber, Payne, & Puto, 1982). Both phenomena can potentially lead to violations of rational choice. Few theories were able to provide an integrated explanation of both phenomena prior to the model proposed by Roe et al. (2001), which is a neural network instantiation of the decision field theory (Busemeyer & Townsend, 1993). That model explains the two effects (in addition to several other important choice phenomena) by taking into consideration similarity relations among options and the dynamic nature of decision processes. The model described here is similar to that of Roe et al. in that it also takes into account similarity among alternatives; however, the manner in which similarity is represented and processed differs between the two models. We will briefly discuss the relationship between the two models after we present our proposal.

Neural network models have been one of the major modeling tools in cognitive science (Rumelhart, McClelland, & PDP Research Group, 1986). However, such models have had only limited applications to decision behavior (Holyoak & Simon, 1999; Roe et al., 2001; Thagard & Millgram, 1995). The model we describe here, like that of Roe et al. (2001), uses a neural network approach to provide an account of the similarity and attraction effects.

Operation of the Model

Decision Scenario and Model Architecture

The decision scenario used here is adapted from that used by Roe et al. (2001). The decision maker has to choose one car from a set of two or three alternatives by evaluating their ratings on two attributes: gas mileage and performance (see Figure 2). A simple neural network is constructed for this scenario. Figure 2 shows the architecture of the model, adapted from ECHO (Thagard, 1989), a neural network

model of how people achieve coherence in making explanations. Two nodes represent the attributes, gas mileage and performance, and three others represent the three alternatives. One special node, labeled as External Driver in Figure 2, represents the motivational and attentional sources that drive the decision process. The lines between nodes represent node connections. Each attribute or alternative is thus represented by one node in the network, with relations among attributes and alternatives represented by connection weights.

Bidirectional excitatory links (represented by dark arrowheads in Figure 2) connect attribute nodes to their respective alternatives. The alternative nodes send out inhibitory influences (represented by empty arrowheads in Figure 2) to one another. Node activation ranges from 0.0 to 1.0. The special node, which drives the decision-making process, always feeds excitatory influence to the attribute nodes, thereby initiating and maintaining activation throughout the entire network. The special node has a constant activation of 1.0, and the weight of its connections to the attribute nodes is 0.05 (there are no reciprocal connections to the special node from the attribute nodes, as the former is intended to be the source of activation). Because the three alternative nodes compete via inhibitory connections with one another, one winning node generally achieves a much higher activation than the rest.

Setting Connection Weights and Initial Activations

Initially, the connection weight between an attribute and an alternative node (called *attribute-alternative weight* from now on) is set to the rating of the alternative on the corresponding attribute. For example, in Figure 2, the option Target is rated 8 and 2 on performance and gas mileage, respectively, so its initial weights are set to 8.0 and 2.0 for the performance-target and gas-mileage-target connections, respectively.

Next, each initial weight is normalized:

$$w_{ij} = \eta + \frac{(w_{ij} - \min(w)).(\kappa - \eta)}{\max(w) - \min(w)}. \qquad (1)$$

Here, w_{ij} is the weight of the connection to node i from j. Weights are normalized according to their range; κ and η are maximum (set to 0.8) and minimum (set to 0.2) values for that range, respectively. Accordingly, the normalized weight

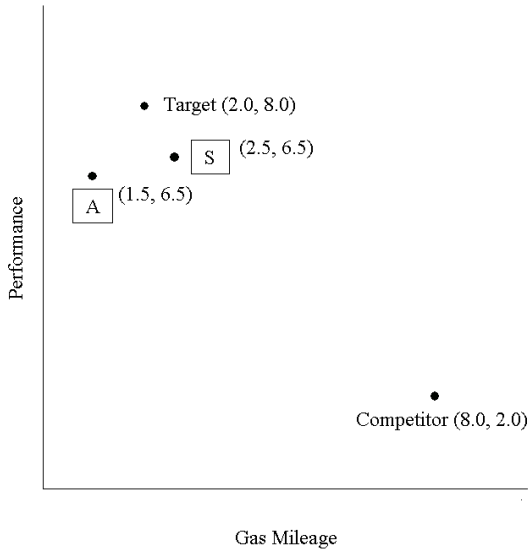


Figure 1. A summary of the phenomena simulated. The letters S and A stand for where the decoy is positioned: Decoy S yields the similarity effect; decoy A yields the attraction effect. The numbers in parentheses are the attribute ratings of the nearby alternative: The first number is the rating of that alternative on gas mileage and the second number is its rating on performance.

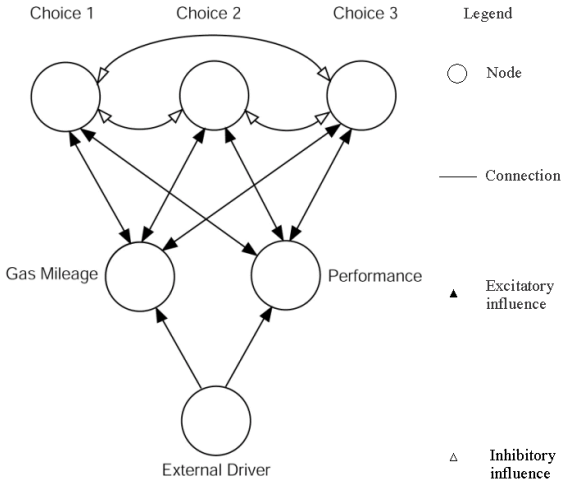


Figure 2: The architecture of the model. Choice 1, Choice 2 and Choice 3 are the alternatives, and Gas Mileage and Performance are the attributes. External Driver represents the motivational and attentional sources that drive the decision process.

should always be within the range of 0.2 to 0.8. The choice of this range is arbitrary, but it reflects the

assumption that the perception of an attribute value should never actually reach 0, which can be viewed as reflecting no value at all, nor should it reach 1, which can be viewed as reflecting sublime satisfaction. The range of actual attribute value is computed by max(w) - min(w), where max(w) and min(w) are the largest and smallest attribute values obtained for all attributes.

The attribute-alternative weights as defined in Equation 1 are linearly related to the actual attribute ratings. In choice behavior, we are concerned with a *subjective* measure of utility in which the impact of a given increase in rating declines with the absolute magnitude of the rating. Accordingly, attribute-alternative weights are transformed:

$$w_{ij} = \frac{w_{ij} l}{w_{ij} + \lambda}. \qquad (2)$$

Here, both l and λ are constants. After exploration of the parameter space, l was set to 1.4 and λ was set to 0.5 to achieve good simulation results. Equation 2 describes a basic psychophysical function in which sensitivity to an increase of stimulus strength declines as the stimulus strength increases. Finally, weights undergo a linear transformation specified by

$$w_{ij} = w_{ij}\tau / 10.0. \qquad (3)$$

Here, τ (set to 4.0) is a parameter intended to amplify the attribute-alternative weights so that the same difference between attribute values now has a larger impact on node activations (see Equations 4 and 5). Finally, these weights are divided by 10.0 so that they are kept reasonably small in relation to node activations. Although the model has several parameters, and specific values for them were selected after extensive search of parameter space, the choices of parameter values do not affect the underlying conceptual framework of the model. Moreover, it is very likely that other sets of parameter values exist that would allow the model to exhibit desired behavior.

The inhibitory connections among the alternative nodes are all set to -0.60. The initial activations are set to 1.0 for the special node and 0.5 for all other nodes (0.5 is the middle point of the activation range, 0.0 - 1.0). To increase psychological realism, some randomness is introduced: The initial activation of an alternative node is a random number within the range of 0.5 $\pm$ 0.01. The generation of random numbers conforms to a uniform distribution. There is no randomness for the activations of the special node and the attribute nodes.

Running the Model

The model runs in an iterative fashion. In each iteration the activation of a node is updated by a commonly-used activation function,

$$a_i(t+1) = \qquad (4)$$
$$\begin{cases} input_i(MAX - a_i(t))\gamma + a_i(t)(1-\theta) & if\ \ input_i > 0 \\ input_i(a_i(t) - MIN)\gamma + a_i(t)(1-\theta) & otherwise \end{cases}$$

$a_i(t+1)$ is the activation of node i at iteration $t + 1$; it is a function of $a_i(t)$, the activation of the same node at the previous iteration. *MAX* and *MIN* are the upper (1.0) and lower (0.0) limits of node activation. θ (set to 0.015) is a decay parameter specifying how much the activation decays in each iteration, and γ (set to 0.12) is a growth rate specifying the increment of activation as a function of the input. The parameter $input_i$ is the total influence received by node i from other nodes connected to it, specified by

$$input_i(t) = \sum_j w_{ij} a_j(t). \qquad (5)$$

The model runs iteratively according to Equations 4 and 5 until the activation of each node no longer changes from the previous iteration by more than a settling criterion (set to 0.001 here). According to Equation 4, a major determinant of node activation is the total input a node receives from other nodes; and according to Equation 5, this input depends on the attribute-alternative weights. It follows that an alternative with a high additive attribute rating tends to have a higher node activation than those with low additive attribute ratings; this is an instantiation of the value maximization principle, which implies that the winning choice should have the highest additive utility summed across all attributes.

The choice probability of an alternative depends on the activation of the corresponding node. Luce's (1959) choice model is used to convert the activation into choice probability for alternative i:

$$probability(i) = \frac{activation(i)}{\sum_j activation(j)} \qquad (6)$$

Simulations and Results

The two phenomena simulated are schematized in Figure 1. For each phenomenon, 100 simulations were run and the results were averaged for each attribute and alternative. The averaged results are presented both as node activations, which are the final activation values of the nodes (see Table 1), and

choice probabilities, which are converted from activations using Equation 6 (see Table 2).

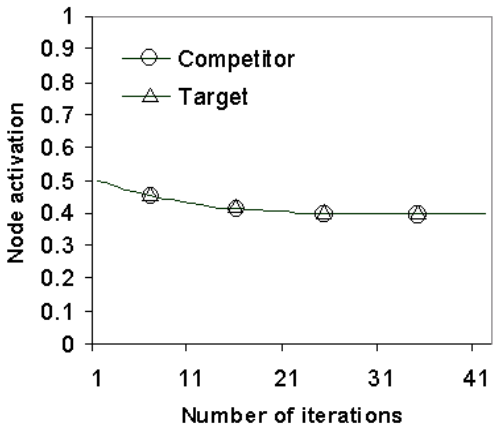


Figure 3: Decision process of binary choice. The activation of alternative nodes is plotted as a function of number of iterations.

Binary Choice

The original choice set contains two alternatives, one of which is arbitrarily selected as the target, and the other the competitor (see Figure 1). Both cars receive ratings on a 10-point (1 - 10) scale for gas mileage and performance. To simplify the choice scenario, the two options are made equal in terms of additive attribute rating: The competitor is rated 8 on gas mileage and 2 on performance, whereas the target is rated 2 on gas mileage and 8 on performance. It is a trivial prediction that (assuming the two attributes are equally important) the two alternatives should be equally likely to be chosen. The model makes this prediction: when these two alternatives are equally attractive, both have a 50% chance of being chosen (see Table 2).

Similarity Effect

If the decoy is similar and competitive compared to one of the two original choices, the target, the introduction of the decoy reduces the probability of the target being chosen relative to that of the other choice in the original set, the competitor. This similarity effect (Tversky, 1972) can lead to a violation of an axiom of rational choice, independence of irrelevant alternatives, which implies that adding an alternative to a choice set will not alter the rank order of the original options. To produce a similarity effect, the decoy should be roughly as good as the target in terms of additive attribute rating. In the simulation, the decoy is chosen to have attribute values of 2.5 and 6.5 for gas mileage and performance, respectively (see Figure 1).

To model the similarity effect, we first run the model on a choice set that includes only the target and the decoy. After the network settles for that comparison, we run it on the entire set of three alternatives. The psychological rationale is that because the target and the decoy are similar to each other, they are grouped together in a manner similar to a perceptual grouping (e.g., in visual perception, when two shapes are close to each other, they are perceived as belonging to the same cluster). Our assumption is that the two similar alternatives are perceived as belonging to the same category, and therefore are compared to each other before all three alternatives are compared.

The simulation was thus divided into two stages: a binary comparison in which only the target and the decoy were compared, and a trinary comparison in which all three alternatives were compared. The activations are carried over from the first to the second stage; accordingly, any activation differences from the first stage will have an effect on the second stage. At the end of the binary-comparison stage, the target has an activation lower than 0.5, the baseline activation, due to its competition with the decoy. This low activation is carried over to the trinary-comparison stage, where the competitor joins the comparison with the default initial activation of 0.5. Thus in the trinary-comparison stage the target starts with a lower activation as compared to the competitor; as a result, the target attains a lower activation and choice probability as compared to the

Table 1: Simulation results as node activations.

Choice scenarios	average node activations				
	gas mileage	performance	competitor	target	decoy
Binary choice	0.647	0.647	0.398	0.398	-----
Similarity effect	0.695	0.729	0.424	0.343	0.317
Attraction effect	0.708	0.741	0.465	0.627	0.019

Note. Each node activation displayed here is the average of activations for the corresponding node calculated over 100 simulation runs.

competitor at the end of simulation. The dynamic process of the two-stage comparison is shown in Figure 4, where the sudden change in activation indicates the transition from the first to the second stage. The final choice probabilities of the target and the competitor are 0.317 and 0.391 respectively (see Table 2), indicating that the competitor ranks higher in terms of preference. Since in the binary choice the choice probabilities of the two alternatives are equal, the altered rank order is a violation of the principle of independence of irrelevant alternatives.

Table 2 Simulation results as choice probabilities.

Choice scenarios	average choice probabilities		
	competitor	target	decoy
Binary choice	0.500	0.500	-----
Similarity effect	0.391	0.317	0.292
Attraction effect	0.419	0.564	0.017

Note. Each choice probability displayed here is the average of choice probabilities for the corresponding node calculated over 100 simulation runs.

Attraction Effect

Huber et al. (1982) showed that when the additional alternative (a dominated decoy) is similar to and obviously inferior to one of the alternatives (the target) of the original choice set, the introduction of this decoy will increase the probability of the target being chosen more than that of the competitor. This effect can potentially increase the probability that the target is chosen, thereby leading to violation of an axiom of rational choice, the regularity principle, which states that adding additional alternatives into the choice set would not increase the choice probabilities of options in the original choice set (cf. Huber et al., 1982). The violation of the regularity principle is a stronger form of preference reversal than the violation of independence of irrelevant alternatives.

The same two-stage comparison is employed to model the attraction effect, because the target and the decoy are similar to each other and therefore form a natural grouping. At the end of the binary comparison, the target has an activation higher than 0.5, the baseline activation, due to its superiority as compared to the decoy. This advantage in activation is carried over to the trinary comparison, and as a result the target has a relatively high activation and choice probability at the end of the simulation run. The dynamic process of the two-stage comparison is shown in Figure 5, where the sudden change in activation indicates the transition between the two stages of comparison. The final choice probability of the target is 0.564 (see Table 2). In the original binary choice set, the target has a choice probability of 0.5 (see Table 2); thus adding the decoy leads to a violation of regularity principle.

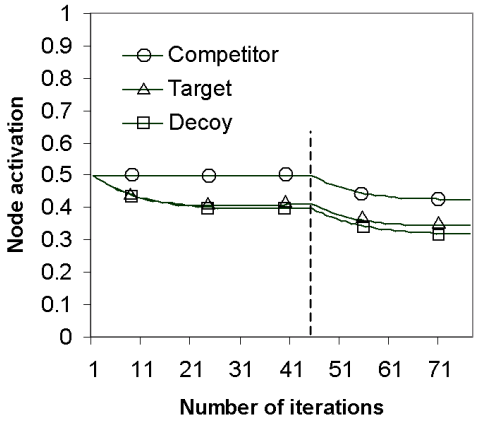


Figure 4: Decision process of similarity effect. Axes are the same as Figure 3. The vertical dashed line indicates the transition from binary comparison to trinary comparison.

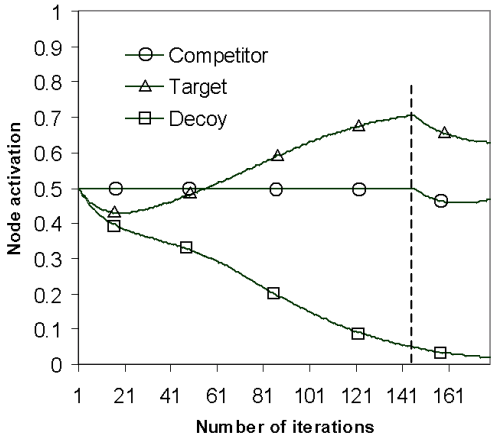


Figure 5: Decision process of attraction effect. Axes are the same as Figure 3. The vertical dashed line indicates the transition from binary comparison to trinary comparison.

In simulating both effects, the model still computes a form of value maximization; however, the computation is carried out in a local instead of global manner during the first stage of comparison, due to the categorization process in which two similar alternatives are grouped and processed together independently of the third alternative.

Conclusions

The connectionist model presented here explains two perplexing empirical findings in choice behavior using a straightforward neural network algorithm and simple psychological principles. It has been argued that the principle of value maximization underlying rational choice is in conflict with some apparently irrational choice behaviors (Simonson & Tversky, 1992). However, the present model shows that choice behavior can be viewed as value maximization constrained by categorization processes.

Roe et al. (2001) also used similarity relations to account for the similarity and attraction effects. In their neural network model, lateral inhibition among alternatives is set in such a way that the more similar two options are, the stronger is the lateral inhibition between them. This differential inhibition provides a foundation for modeling similarity-related findings. In contrast, in the present model similarity is assumed to lead to a grouping effect, which in turn leads to the two-stage comparison process. Thus while both models emphasize the role of similarity in choice behavior, Roe et al.'s algorithm models the impact of similarity by variations in a continuous parameter for inhibition; whereas the present algorithm hold inhibition constant and instead assumes that similarity alters the grouping of options, leading to a multi-stage comparison process. Further empirical investigations will be required to distinguish between these two possible mechanisms by which similarity may modulate choice behavior.

The present model has several limitations that will need to be addressed in future work. For example, the choice scenario is constructed in a highly schematic way, and more complex and realistic choice scenarios need to be used in future studies. Also, the way the connection weights are set by explicit equations is rather artificial; future efforts need to address how the weights may be acquired using a connectionist learning mechanism. Perhaps most importantly, the critical assumption that similar choices are grouped together and therefore processed together in choice behavior requires further empirical investigation.

The present model may have implications for applied work. Expert systems based on the current model can be developed to analyze and predict choice behavior. In contrast to more traditional axiom-based systems, such systems may make it possible to analyze apparently irrational choice and decision processes, thereby leading to more accurate predictions of human decisions.

Acknowledgments

This research was supported by NSF Grant SES-0080375. We are grateful for comments from Jerry R. Busemeyer, Patricia W. Cheng, Aimee Drolet and Shi Zhang.

References

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432-459.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3-31.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9*, 90-98.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review, 2*, 1-19.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*, 370-392.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing : explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research, 29*, 281-295.

Thagard, P. (1989). Explanatory coherence. *Behavioral & Brain Sciences, 12*, 435-502.

Thagard, P., & Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram & D. B. Leake (Eds.), *Goal-driven learning* (pp. 439-454). Cambridge, MA: MIT Press.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*, 281-299.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science, 39*, 1179-1189.

Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions

Todd M. Gureckis (gureckis@love.psy.utexas.edu)
Department of Psychology; The University of Texas at Austin
Austin, TX 78712 USA

Bradley C. Love (love@psy.utexas.edu)
Department of Psychology; The University of Texas at Austin
Austin, TX 78712 USA

Abstract

How do people learn and organize examples in the absence of a teacher? This paper explores this question through a examination of human data and computational modeling results. The SUSTAIN (Supervised and Unsupervised STratified Incremental Network) model successfully fits human learning data drawn from two published studies. The first study examines how correlations between features can facilitate unsupervised learning. The second set of studies examines the role that similarity and attention play in unsupervised category construction (i.e., sorting) tasks. Importantly, SUSTAIN suggests two novel behavioral predictions that are confirmed.

Introduction

The study of human category learning has focused on supervised learning. Researchers typically utilize a experimental procedure in which the participant must learn to classify a set of stimuli while receiving corrective feedback on every trial. Certainly, there are many other ways to learn about the world. Our environment does not always provide us with explicit feedback and thus, some learning is better characterized as unsupervised. For example, we routinely categorize incoming email as "junk mail" in the absence of a teacher. A great deal of human learning may be unsupervised. The goal of this paper is to expand our understanding of how humans learn from examples without supervision.

To achieve this goal, we fit the SUSTAIN model of category learning to Billman and Knutson's (1996) studies concerning how humans learn correlations through observation and to Medin, Wattenmaker, and Hampson's (1987) data on unsupervised category construction (i.e., sorting) behavior. SUSTAIN successfully accounts for human performance in both of these studies with one set of parameters. Importantly, SUSTAIN's account of these studies suggests novel predictions which are subsequently tested (and confirmed) with human subjects.

The Modeling Approach

SUSTAIN has been successfully applied to an array of challenging human data sets spanning a variety of category learning paradigms including supervised classification (Love & Medin, 1998), inference learning (Love, Markman, & Yamauchi, 2000), and unsupervised learning (Gureckis & Love, 2002). One primary goal of our modeling approach is to address multiple forms of category learning (both supervised and unsupervised) with one consistent set of principles. After a brief introduction to the operation of SUSTAIN, these core principles will be discussed.

Introduction to SUSTAIN

SUSTAIN is a clustering model of human category learning. The internal representation of the model consists of a set of clusters. Category representations consist of one or more associated clusters. At the start of learning, the network has a single cluster that is centered in this representational space upon the first input pattern.

When a new stimulus item is presented, SUSTAIN attempts to assign the item to the most similar existing cluster. This assignment is unsupervised since it is based only on the similarity between item and cluster. If a *surprising* event occurs, such as a misprediction in supervised learning or a stimulus is encountered in unsupervised learning that is not similar to any existing cluster, SUSTAIN creates a new cluster to encode the current stimulus. This new cluster is centered in the representational space on the misclassified item.

When a stimulus is not surprising, the item is assigned to the most similar existing cluster and this cluster updates its internal representation to become more similar to the current item (a process somewhat analogous to prototype formation). Classification decisions are based on the cluster to which a stimulus instance is assigned. Like other models of category learning (e.g., Kruschke, 1992), SUSTAIN's selective attention mechanism learns to selectively weight stimulus feature dimensions that are most useful for categorization.

The Principles of SUSTAIN

With this general understanding of the operation of the model, we now examine the six key principles that underly SUSTAIN.

Principle 1, SUSTAIN is directed towards simple solutions SUSTAIN is initially directed towards simple solutions. At the start of learning, SUSTAIN has only one cluster which is centered on the first input item. It then adds clusters (i.e., complexity) only as needed to accurately describe the category structure of the learning task. Its selective attention mechanism further serves to bias SUSTAIN towards simple solutions by focusing the model on the stimulus dimensions that provide consistent information.

Principle 2, similar stimulus items tend to cluster together In learning to classify stimuli as members of two distinct categories, SUSTAIN will cluster similar items together. For example, different instances of a bird subtype (e.g., sparrows) could cluster together and form a sparrow cluster instead of leaving separate traces in memory for each instance. Clustering is an unsupervised process because cluster assignment is done on the basis of similarity, not feedback.

Principle 3, SUSTAIN relies on both unsupervised and supervised learning processes As discussed above, SUSTAIN can cluster based on similarity (an unsupervised process). SUSTAIN's operation is also affected by supervision (when available). Consider the example of SUSTAIN learning to classify stimuli as members of the category mammals or birds. Let's assume that a cluster representing four-legged, hairy, land creatures is already acquired, as well as another cluster representing small, winged, creatures that fly. The first time SUSTAIN is asked to classify a bat, the model will predict that a bat is a bird because the bat stimulus will be more similar to the existing bird cluster than to the existing mammal cluster. Upon receiving corrective feedback (supervision), SUSTAIN will note its error and create a new cluster to store the anomalous bat stimulus. Now, when this bat or one similar to it is presented to SUSTAIN, SUSTAIN will correctly predict that the bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary through cluster recruitment (see Principle 1).

Principle 4, Clusters are recruited in response to surprising events As the previous example illustrates, surprising events lead to new clusters being recruited. In unsupervised learning, a surprising event is simply exposure to a stimulus that is not sufficiently similar to any existing cluster (i.e., a very novel stimulus).

Principle 5, the pattern of feedback matters As the bird-mammal example above illustrates, feedback affects the inferred category structure. Prediction failures result in a cluster being recruited, thus different patterns of feedback can lead to different representations being acquired. This principle allows SUSTAIN to predict different acquisition patterns for different learning modes (e.g., inference versus classification learning) that are informationally equivalent but differ in their pattern of feedback.

Principle 6, cluster competition Clusters can be seen as competing explanations of the input. The strength of the response from the winning cluster (the cluster the current stimulus is most similar to) is attenuated in the presence of other clusters that are somewhat similar to the current stimulus (compare to Sloman's, 1997, account of competing explanations in reasoning).

Model Fits and Predictions

In the following sections, Billman and Knutson's (1996) results are described, fit, and SUSTAIN's novel predictions are tested. Following Billman and Knutson, Medin et al.'s (1987) work is given similar consideration.

Modeling Billman and Knutson's (1996)

Billman and Knutson's experiments tested the prediction that category learning is easier when certain stimulus feature dimensions are predictive of other feature dimensions (e.g., "has wings", "can fly", "has feathers" are all inter-correlated features of birds) than when correlations are unrelated or are not numerous. Their studies evaluate how relations among stimulus feature dimensions affect learning in an unsupervised task. SUSTAIN has successfully fit Billman and Knutson's (1996) Experiment 2 and 3 (Gureckis & Love, 2002). Here, we focus on Experiment 3.

Fitting Billman and Knutson's (1996) data Subjects studied stimulus items that depicted imaginary animals made up of of seven feature dimensions: type of head, body, texture, tail, legs, habitat, and time of day pictured. Each dimension could take on one of three values. For example, the time of day could be "sunrise", "nighttime", or "midday". The correlational structure of the feature dimensions varied according to which of two conditions (either the Structured or the Orthogonal condition) the subject was randomly assigned. The abstract structure of the two conditions is shown in Table 1. In the Structured condition, the first three stimulus dimensions are intercorrelated (for a total of three correlations), while the remaining four dimensions vary freely. The Orthogonal condition's structure also contains three correlations (the first and second dimensions are correlated, as are the third and fourth, and the fifth and the sixth), but the correlations are isolated (e.g., the first and third dimension are not correlated).

In the learning phase for both conditions, subjects were told that they were participating in a visual memory experiment and viewed 27 stimulus items for four blocks (a block is a single pass through all training items). Each of the 27 items appeared once

Table 1: The logical structure of the stimulus items for the Orthogonal and Structured conditions in Experiment 3 of Billman and Knutson (1996). The seven columns denote the seven stimulus dimensions. Each dimension can display one of three different values, indicated by a 1, 2, or 3. An x indicates that the dimension was free to assume any of the three possible values.

Structured Condition		
1 1 1 x x x x	2 2 2 x x x x	3 3 3 x x x x
Orthogonal Condition		
1 1 1 1 1 1 x	2 2 1 1 1 1 x	3 3 1 1 1 1 x
1 1 1 1 2 2 x	2 2 1 1 2 2 x	3 3 1 1 2 2 x
1 1 1 1 3 3 x	2 2 1 1 3 3 x	3 3 1 1 3 3 x
1 1 2 2 1 1 x	2 2 2 2 1 1 x	3 3 2 2 1 1 x
1 1 2 2 2 2 x	2 2 2 2 2 2 x	3 3 2 2 2 2 x
1 1 2 2 3 3 x	2 2 2 2 3 3 x	3 3 2 2 3 3 x
1 1 3 3 1 1 x	2 2 3 3 1 1 x	3 3 3 3 1 1 x
1 1 3 3 2 2 x	2 2 3 3 2 2 x	3 3 3 3 2 2 x
1 1 3 3 3 3 x	2 2 3 3 3 3 x	3 3 3 3 3 3 x

per block in a random order. The only difference between the Structured and Orthogonal conditions was the abstract structure of the stimuli that were shown during the learning phase.

In the test phase of the experiment, subjects viewed a novel set of 54 stimulus pairs. Each member of the pair had two of the seven feature dimensions obscured (e.g., the locations where the tail and head should have been were blacked out) so that information about only one correlation was available for each item in test pair. One item in the pair preserved the studied correlation, while the other item violated the correlation. Subjects were asked to choose the stimulus item in the pair that seemed most similar to the items studied in the learning phase (a forced choice procedure). The item that preserved the studied correlation was considered the correct choice. For example, in the isolating condition the correct item of the pair might have the abstract structure [1 1 m 1 m 1 2] because it preserves the correlation between the first and second dimensions (the 'm' represents a dimension that was blocked). The incorrect item of the pair might then be [1 2 m 1 m 1 2] which breaks the correlation present in the training items between the first and second dimension.

The basic result from Experiment 3 was that the "correct" item was chosen more often in the Structured condition than in the Orthogonal condition (77% vs. 66% from Table 2). This finding supports the hypothesis that extracting a category's structure is facilitated by intercorrelated dimensions.

Table 2: The mean accuracy for humans and SUSTAIN in Billman and Knutson's (1996) Experiment 3.

	Orthogonal	Structured
Human	.66	.77
SUSTAIN	.60	.77

Table 3: SUSTAIN's best fitting parameters for the studies considered. SUSTAIN's parameters are not discussed in this paper, but this table is included for readers who wish to replicate our results.

function/adjusts	symbol	value
learning rate	η	0.0966
cluster competition	β	6.40
decision consistency	d	1.98
attentional focus	r	10.0
threshold	τ	0.5

Modeling Results SUSTAIN was trained in a manner analogous to how subjects were trained by using four randomly ordered learning blocks. No feedback was provided as all stimulus items were encoded as being members of the same category. New clusters were recruited according to the unsupervised notion of surprise. In order for SUSTAIN to mimic the forced choice nature of the test phase, a response probability was calculated for each of the two items. The ultimate response of the network was biased towards the item in the forced choice that had the strongest response probability.

SUSTAIN was run numerous times on both conditions in both experiments and the results were averaged. The best fitting parameters are shown in Table 3. SUSTAIN correctly predicts greater accuracy in the Structured condition than in the Orthogonal condition (see Table 2).

In Experiment 3, SUSTAIN's most common solution in the Orthogonal condition was to partition the studied items into three clusters. However, the nature of the three partitions varied across runs. SUSTAIN tended to focus on one of three correlations present in the Isolated condition and ignored the other two. For instance, during training SUSTAIN might create three clusters organized around the first two input dimensions (one cluster for each correlated value across the two dimensions) and ignore the correlation between the third and fourth dimensions and the fifth and sixth dimensions.

SUSTAIN also recruited three clusters in the Structured condition. The same dynamics that lead SUSTAIN to focus on only one correlation in the Orthogonal condition leads SUSTAIN to focus on all of the interrelated correlations in the Structured condition. When SUSTAIN learns one correlation in the

Structured condition, SUSTAIN necessarily learns all of the pairwise correlations because of the way clusters are updated (i.e., three clusters are formed that capture the three basic subtypes of stimuli). This type of learning in the Structured condition is what lead to the higher accuracy levels.

SUSTAIN's solution to Experiment 3 suggests some novel predictions: (a) When correlations are not interrelated, learning one correlation should block the learning of other correlations, and (b) When correlations are interrelated, either all of the correlations are learned or none of the correlations are learned. These predictions are explored in the following section.

Testing the Predictions In the original Billman and Knutson article, accuracy was considered in aggregate for all three correlations. Here, we reanalyze Billman and Knutson's data by considering each subjects' performance on each correlation (i.e., each subject contributes three scores to the analysis instead of one). SUSTAIN predicts that human subjects will learn only one of the three correlations in the Orthogonal condition, but will learn either all or none of the correlations in the Structured condition. If this is true, the mean variance of subjects' accuracies for the three correlations should be higher in the Orthogonal condition than in the Structured condition. This was indeed the case. The mean variance of each subject's three accuracy scores was 0.030 for the Orthogonal condition, but only 0.010 in the Structured condition ($t(46) = 2.76$, $p < .001$).

Discussion Due to the way SUSTAIN organizes its clusters, it predicts that learning one correlation in the Orthogonal condition blocks the learning of other correlations (which should result in a high within subject variance), whereas in the Structured condition learning one correlation is tied to learning all three correlations (which should result in a low within subject variance). These predictions were made prior to obtaining access to Billman and Knutson's data. The combined results of the original Billman study and the subsequent analysis, suggest that people find categories that are organized around highly correlated features to be easier to learn because correlations enable the transfer of knowledge across features. The mechanism that supports this operation may bare a strong resemblance to SUSTAIN.

Modeling Sorting Behavior with SUSTAIN

Billman and Knutson's (1996) studies suggest that subjects prefer stimulus organizations in which the perceptual dimensions are intercorrelated. However, studies in category construction reveal a contrasting pattern — subjects tend to sort stimuli along a single dimension. This behavior persists despite the fact that alternate organizations exist that respect the intercorrelated nature of the stimuli, such as an intercorrelated family resemblance structure (Medin, Wattenmaker, & Hampson, 1987).

Table 4: The logical structure of the perceptual dimensions in Medin et al. (1987) sorted in two ways. In the family resemblance table, the stimuli with a preponderance of 1's can be seen as forming one family, while the stimuli with a preponderance of 2's can be seen as forming a second family or covert category. In the one-dimensional sort table, the same stimuli items are grouped on the basis of a single dimension (the first dimension).

Family Resemblance		One-dimensional Sort	
1 1 1 1	2 2 2 2	1 1 1 1	2 2 2 2
1 1 1 2	2 2 2 1	1 1 1 2	2 2 2 1
1 1 2 1	2 2 1 2	1 1 2 1	2 2 1 2
1 2 1 1	2 1 2 2	1 2 1 1	2 1 2 2
2 1 1 1	1 2 2 2	1 2 2 2	2 1 1 1

SUSTAIN was applied to the sorting data from Medin et al.'s (1987) Experiment 1 in hopes of reconciling the apparently contradictory findings. In Experiment 1, subjects were instructed to sort ten stimuli into two equal sized piles. Stimuli were cartoon-like animals that varied on four binary-valued perceptual dimensions (head shape, number of legs, body markings, and tail length). The logical structure of the items is shown in Table 4. The basic finding is that subjects sort along a single dimension (the one-dimensional sort in Table 4) as opposed to sorting stimuli according to their intercorrelated structure (i.e., the family resemblance structure shown in Table 4).

In these simulations, SUSTAIN was constrained to create only two piles (i.e., clusters) like Medin at al.'s subjects. This was accomplished by preventing SUSTAIN from recruiting a third cluster. SUSTAIN was presented with the items from Table 4 for 10 random training blocks to mirror subjects' examination of the stimulus set and their ruminations as to how to organize the stimuli. To evaluate the performance of the model, we looked at how SUSTAIN's two clusters were organized. Using the same parameters that were used in the Billman and Knutson (1996) studies listed in Table 3, SUSTAIN correctly predicted that the majority of sorts (99%) are organized along one stimulus dimension.

SUSTAIN's natural bias to focus on a subset of stimulus dimensions (which is further stressed by the selective attention mechanism) led it to predict the predominance of one-dimensional sorts. Attention is directed towards stimulus dimensions that consistently match at the cluster level. This leads

to certain dimensions becoming more salient over the course of learning (i.e., the model's attention value along that dimension becomes larger). The dimension that develops the greatest salience over the course of learning becomes the basis for the one-dimensional sort.

Which dimension provides consistent information during the course of learning will, in part, be determined by the order in which the stimulus items are presented to the model. Thus, SUSTAIN predicts that the order of card consideration in a sorting task might constrain which dimension human subjects focus their sort on. If card ordering has no effect and subjects randomly choose a dimension to sort on or choose due to individual differences in the salience of a particular dimension, then SUSTAIN's account should be insufficient.

Testing the Prediction

The following study tests this prediction by creating a modified version of the Medin, et al. sorting experiment in which the order that subject may consider cards is manipulated. Our interest was to test if the dynamics that led SUSTAIN to choose a particular dimension to sort on were the same dynamics that constrained subjects' sorting strategies.

Procedure Stimuli in our experiment were geometric shapes, printed on laminated cards, that varied on four of five binary valued dimensions (one dimensions value was held constant and thus had no influence on subjects sorting decisions). The dimensions were size (big or small), color of border (white or yellow), main color (blue or purple), a slash across the shape (present or absent), and texture (smooth or rough). Each dimension is independent and equally salient (as verified by multi-dimensional scaling of subjects' pairwise similarity ratings).

Participants were given a large board that was divided in half with a dark line (see Figure 1). Each side of the board had five positions in which to place cards. Before the start of an experiment trial, two "guide" cards were placed on the board that had opposing values along each dimension. Figure 1 shows an empty board with the abstract structure of these two guide cards. The particular values and meaning of each stimulus dimension was random for each subject (i.e., the values of the stimulus dimensions such as size and color were randomly assigned to one column of the abstract structure shown).

During the experiment, participants were given one new card at a time by the experimenter and were asked to place the card in an empty position on one side of the board according to what seemed most natural or sensible given the other cards on that side. The first two cards actually handed to subjects were constrained so that they mismatched on one dimension from the guide cards already on the board. For example, given the two cards in Figure 1 the abstract structure of the first two cards actually handed to the subjects might be [0 0 0 1] and [1 1 1 0].

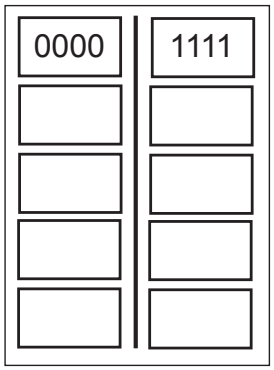


Figure 1: The layout and initial configuration of the board given to subjects is shown.

The final 6 cards given to subjects were drawn from the remaining possible. Cards were randomly chosen but came in pairs of opposing values. For example, if the fourth card had the abstract structure [0 0 1 0], the fifth might be [1 0 1 1]. This manipulation also helped to encourage subjects to fill the board up in a more or less even fashion rather than filling up one side completely, then having no choice as where to place the remaining cards.

Our hypothesis was that subjects would, like SUSTAIN, place the first two cards on the board on the basis of overall similarity to the guide cards as opposed to randomly choosing a single dimension on which to focus their sorting strategy. Thus in our example, [0 0 0 1] would be placed under the [0 0 0 0] prototype and [1 1 1 0] would be placed under the [1 1 1 1] prototype. If subjects allocated attention to dimensions that provide consistent information like SUSTAIN, then attention would be increased on only the dimensions that matched the guide cards (all but the fourth dimension in this case). This initial attentional disadvantage on the fourth, mismatched dimension would prevent subjects from sorting on that dimension.

Results Twenty-eight psychology undergraduate students participated in the study for course credit. The results collected for this study are show in Table 5. Of the 28 subjects, 23 subjects performed a one-dimensional sort while 5 used an alternate sorting strategy. Of the 23 subjects that performed a one dimensional sort, only 2 of these 23 subjects sorted the cards using the mismatched dimension as their basis for organization. If subjects had no particular preference for any dimension and the manipulation of the cards had no effect, then the probability of getting 21 out of 23 subjects to sort on a dimension other than the one mismatching dimension is

Table 5: The results of the sorting study.

	Number of Subjects
Subjects using a 1D sort	23
—*Mismatched Dimension*	2
—*Other Dimensions*	21
Subjects using a non 1D sort	5
—*Family Resemblance*	3
—*Unknown Strategy*	2
Total Subjects	28

less than .05 as given by a two-tailed binomial trial (n=23, p = .25). Of the five subjects that did not perform a one-dimensional sort, three performed a family resemblance sort and two performed a sort using an undecipherable sorting strategy.

SUSTAIN was simulated using the same parameters used for the Billman and Knutson studies (Table 3) and using the same conditions from the Medin, et al. sorting simulation, but with the specific card orderings that subjects were given in our experiment. In 100% of the trials, the model used a dimension other than the mismatched dimension as the basis for a one-dimensional sort.

Discussion

The dimension that subjects choose to sort in this task cannot be explained as random choice. The results presented in our experiment provide evidence that the order of card presentation plays a role in influencing subjects to sort on a particular dimension.

Specifically, sorting behavior is influenced by the way we perceive similarity between stimuli. In this unsupervised task, attention is allocated such that the similarity space changes during the course of learning. At the start of learning, each dimensions is more or less equally important, but as learning proceeds, certain dimensions become more salient (because they are more informative) while others become less. This warping of the similarity space is what ultimately causes judgments in this type of task to become based on a single dimension, rather than on the overall similarity between items. The fact that SUSTAIN predicted this behavior gives additional support to the notion that it's principles reflect some of the true operational principles of human learning.

Conclusions and Implications

SUSTAIN's combined account of Billman and Knutson's (1996) studies and Medin et al. (1987) suggest that the salience of stimulus dimensions change as a result of unsupervised learning and that the correlated structure of the world is more likely to be respected when there are numerous intercorrelated dimensions that are strong. In cases where the total number of correlations is modest, and the correlations are weak and not interrelated (such as in the Medin et al. stimuli), SUSTAIN predicts that stimuli will be organized along a single dimension.

The ability of SUSTAIN to account for two diverse unsupervised learning data sets with a single set of parameters demonstrate how it's formulation positions it as a robust model of category learning. In addition to the studies reported here, SUSTAIN's principles has been shown to generalize across a number of other forms of category learning (such as supervised learning and inference learning). It is these well-defined principles and the transparent operation of SUSTAIN that allow it to make the two predictions which have been successfully confirmed here.

Acknowledgments

We would like to offer our sincere thanks to Dorrit Billman for providing access to her data and to Rob Goldstone for motivating the sorting study. This work was supported by AFOSR Grant F49620-01-1-0295.

References

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*(2), 458-475.

Gureckis, T. M., & Love, B. C. (2002). *Modeling unsupervised learning with sustain.* (In Press, *FLAIRS 2002 Special Track: Categorization and Concept Representation: Models and Implications*)

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Love, B. C., Markman, A. B., & Yamauchi, T. (2000). Modeling classification and inference learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 136-141.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. In *Proceedings of the fifteenth national conference on artificial intelligence* (p. 671-676). Cambridge, MA: MIT Press.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242-279.

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking & Reasoning*, *3*, 81-110.

A Constraint Satisfaction Model of Causal Learning and Reasoning

York Hagmayer (york.hagmayer@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Following up on previous work by Thagard (1989, 2000) we have developed a connectionist constraint satisfaction model which aims at capturing a wide variety of tasks involving causal cognitions, including causal reasoning, learning, hypothesis testing, and prediction. We will show that this model predicts a number of recent findings, including asymmetries of blocking, and asymmetries of sensitivity to structural implications of causal models in explicit versus implicit tasks.

Introduction

Causal reasoning has been widely investigated during the last decade, which has led to a number of interesting novel findings (see Shanks, Holyoak, & Medin, 1996; Hagmayer & Waldmann, 2001, for overviews). For example, it has been shown that participants' causal judgments are sensitive to the contingency between the cause and the effect, and that people's judgments reflect the causal models underlying the observed learning events (see Hagmayer & Waldmann, 2001; Waldmann, 1996). Moreover, causal reasoning has been studied in the context of a number of different tasks, such as learning, reasoning, categorization, or hypothesis testing.

Most psychological theories and computational models of causal learning and reasoning are rooted in two traditions. They are either based on associationistic or on probabilistic or Bayesian models (see Shanks et al., 1996; Thagard, 2000). Both kinds of models have been criticized. Associationistic learning networks have proven unable to capture the fundamental semantics of causal models because they are insensitive to the differences between learning events that represent causes versus effects (see Waldmann, 1996). By contrast, Bayesian networks are perfectly capable of representing causal models with links directed from causes to effects (see Pearl, 2000). However, although the goal of these networks is to reduce the complexity of purely probabilistic reasoning, realistic Bayesian models still require fairly complex computations, and they presuppose competencies in reasoning with numerical probabilities which seem unrealistic for untutored people (see Thagard, 2000, for a detailed critique of these models).

The aim of this paper is to introduce a more qualitatively oriented, connectionist constraint satisfaction model of causal reasoning and learning. Our model is inspired by Thagard's (2000) suggestion that constraint satisfaction models may qualitatively capture many insights underlying normative Bayesian network models in spite of the fact that constraint satisfaction model use computationally far simpler, and therefore psychologically more realistic processes. The model differs from standard associationist learning models (e.g., Rescorla & Wagner, 1972) in that it is capable of expressing basic differences between causal models. Our model embodies a uniform mechanism of learning and reasoning, which assesses the fit between data and causal models. This architecture allows us to model a wide range of different tasks within a unified model, which in the literature have so far been treated as separate, such as learning and hypothesis testing.

Constraint Satisfaction Models

Constraint satisfaction models (Thagard, 1989, 2000) aim at capturing qualitative aspects of reasoning. Their basic assumption is that people hold a set of interconnected beliefs. The beliefs pose constraints on each other, they either support each other, contradict each other, or are unrelated. Coherence between the beliefs can be achieved by processes which attempt to honor these constraints.

Within a constraint satisfaction model beliefs are represented as nodes which represent propositions (e.g., "A causes B"). The nodes are connected by symmetric relations. The numerical activation of the nodes indicates the strength of the belief in the proposition. A belief that is highly activated is held strongly, a belief that is negatively activated is rejected. The activation of a node depends on the activation of all other nodes with which it is connected. More precisely, the net input to a single node *j* from all other nodes *i* is defined as the weighted sum of the activation *a* of all related nodes (following Thagard, 1989, p.466, eq.5):

$$\mathrm{Net}_j = \sum_i w_{ij} a_i(t) \qquad (1)$$

The weights *w* represent the strength of the connection of the beliefs. In our simulations, they are generally pre-set to default values which are either positive or negative and remain constant throughout the simulation. At the beginning of the simulations, the activation of the nodes representing hypotheses are set to a low default value. However, nodes representing empirical evidence are connected to a special activation node whose activation remains constant at 1.0. This architecture allows us to capture the intuition that more faith is put into empirical evidence than into theoretical hypotheses (see Thagard, 1989). To update the activation in each

cycle of the simulation, first the net input net_j to each node is computed using Equation 1. Second the activation of all nodes is updated using the following equation (Thagard, 1989, p.446, eq.4):

$$a_j(t+1) = a_j(t)(1-\theta)+net_j(max-a_j(t)) \text{ if } net_j>0$$
$$= a_j(t)(1-\theta)+net_j(a_j(t)-min) \text{ otherwise.} \quad (2)$$

In Equation 2, θ is a decay parameter that decrements the activity of each node in every cycle, min represents the minimum activation (-1) and max the maximum activation (+1). The activations of all nodes are updated until a stable equilibrium is reached, which means that the activation of all nodes do no longer substantially change. To derive quantitative predictions it would be necessary to specify rules that map the final activations to different types of responses. This is an important goal which should be addressed in future research. In the present article we only derive ordinal, qualitative predictions from the model.

The Model

Following causal-model theory (Waldmann, 1996) we assume that people typically enter causal tasks with initial assumptions about the causal structure they are going to observe. Even though specific knowledge about causal relations may not always be available, people often bring to bear knowledge about abstract features of the models, such as the distinction between events that refer to potential causes and events that refer to potential effects. In virtually all psychological studies this information can be gleaned from the initial instructions and the materials (see Waldmann, 1996).

Figure 1 displays an example of how the model represents a causal model. The nodes represent either causal hypotheses or observable events. The causal hypothesis node at the top represents a structural causal hypothesis (H1), in this case the hypothesis that the three events e_1, e_2, x form a common-effect structure with e_1 and e_2 as the two alternative causes and x as the common effect. The two nodes on the middle level refer to the two causal relations H2 and H3 that are part of the common-effect model with two causes and a single effect. The nodes on the lowest level refer to all patterns of events that can be observed with three events (a dot represents "and"). On the left side, the nodes represent patterns of three events, in the middle pairs, and on the right side single events. Not only the present but also the corresponding absent events are represented within this model (for example ~x). The links connecting the nodes represent belief relations. Thus, they do not represent probabilities or causal relations as in Bayesian models. There are two different kinds of connections between the nodes. Solid lines indicate excitatory links, dashed lines inhibitory links. How are the connections defined? A connection is positive if the propositions support each other. For example, if all three events are present, the observation is in accordance with both hypotheses H2 and H3. This pattern might be observed if both e1 and e2 cause x. Therefore the evidence node e1.e2.x is positively connected to H2 and H3. In general, a hypothesis is positively connected to an evidence node if the events mentioned in the hypothesis are either all present or all absent. If this is not the case, that is if one of the relevant events specified in the hypothesis is absent, the link is assigned the negative default value. Exploratory studies have shown, that participants share a common intuition whether a certain pattern of events supports or contradicts a hypothesis (Hagmayer & Waldmann, 2001). The assigned weights mirror these general intuitions. The weights of the links remain the same throughout the simulations. Figure 1 does not display the special activation node. This node was pre-set to 1.0 and attached to event nodes describing present events in the respective experiment.

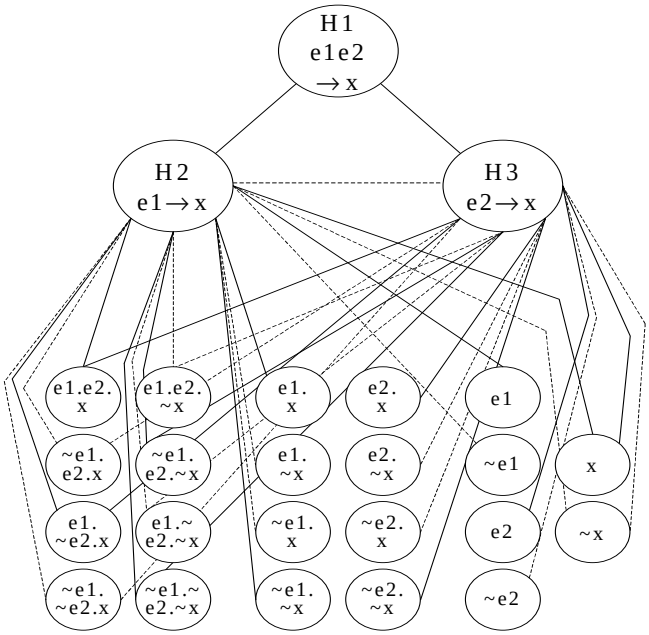


Figure 1: Constraint satisfaction model of causal learning and reasoning. See text for further explanations.

In Figure 1, the dashed line between the hypotheses H1 and H2, which signifies an inhibitory link, is of special interest. The network represents a common-effect structure. This means that there are two causes e1 and e2 which compete in explaining the occurrence of effect x. Therefore the two hypotheses referring to the individual causal relations have to be connected by a inhibitory link (see also Thagard, 2000). However, both hypotheses H2 and H3 are positively connected to the structural hypothesis H1. By contrast, a common-cause structure is represented slightly differently. In such a structure, event x would be the common cause of the two effects e1 and e2 (i.e., H1: x→e1.e2). A model of this structure looks almost identical to the one for the common-effect structure in Figure 1. There is only one very important difference. Because there is no competition between the effects of a common cause, a common-cause model has no inhibitory link between H2 and H3. All other nodes and links in the two models are identical.

Both the common-effect and the common-cause model were implemented using Microsoft Excel. Default values were adopted from the literature if not indicated otherwise (Thagard, 1989). Initial activations were set to 0.01, inhibitory links between nodes to –0.05, and excitatory links to +0.05. The inhibitory link between H1 and H2 within the common-effect model was pre-set to a value of –0.20. The

special activation node was attached to all evidence nodes. The additional activation was divided among the evidence nodes according to the relative frequency of the evidence in the learning input. This principle captures the intuition that more faith is put into evidence that is observed more frequently.

Evaluation

In order to evaluate the proposed constraint satisfaction model different tasks and paradigms from the literature on causal learning and reasoning were modeled. One of our main goals was to show that the same architecture can be used to simulate different types of tasks. However, different tasks required different sections of the model depicted in Figure 1. We used two principles for the construction of task specific networks. The first principle is that we only included the event nodes that corresponded to the event patterns observed in the learning phase or that corresponded to events that have to be evaluated or predicted in the test phase. For example, to model a task in which only event triples were shown, only the event nodes on the left side of the event layer in Figure 1 would be incorporated in the model. However, if the task following the learning phase required the prediction of single events, the corresponding nodes for single events would have to be added to the event layer. The second principle is that only the hypothesis nodes were included that represent hypotheses that are given or suggested to participants. These two principles ensure that for each paradigm a minimally sufficient sub-model of the complete model is instantiated.

Test 1: Asymmetries of Blocking

Blocking belongs to the central phenomena observed in associative learning which, among other findings, have motivated learning rules that embody cue competition (e.g., Rescorla & Wagner, 1972). A typical blocking experiment consists of two learning phases. In Phase 1 participants learn that two events e1 and x are either both present or absent. In Phase 2 a third event e2 is introduced. Now all three events are either present or absent. In both phases, events e1 and e2 represent cues and x the outcome to be predicted. Associative theories generally predict a blocking effect which means that participants should be reluctant about the causal status of the redundant event e2 that has been constantly paired with the predictive event e1 from Phase 1. This prediction has come under attack by recent findings that have shown that the blocking effect depends on the causal model learners bring to bear on the task (see Waldmann, 1996, 2000). If participants assume that e1 and e2 are the causes of x (common-effect structure) a blocking effect can be seen. In contrast, if participants assume that e1 and e2 are the collateral effects of the common cause x (common-cause structure), no blocking of e2 is observed. In this condition, learners tend to view both e1 and e2 as equally valid diagnostic cues of x.

To model blocking, we used a network that was extended after Phase 1. In Phase 1, the net consisted of a hypothesis node (H2) and the nodes for patterns of two events (e1, x). After Phase 1, the final activation of the hypothesis node was transferred to Phase 2. In Phase 2, the network consisted of two nodes for the two causal hypotheses (H2 and H3), and nodes that represented patterns of three events, the patterns participants observed within the learning phase. Furthermore, the node H1 was included, which, depending on the condition, either coded a common-cause or a common-effect hypothesis. The nodes for the event pairs from Phase 1 were removed.

Figure 2 shows the activation of the two hypotheses referring to the causal relations in Phase 1 and 2. Figure 2A depicts the activation for the common-cause model and Figure 2B for the common-effect model.

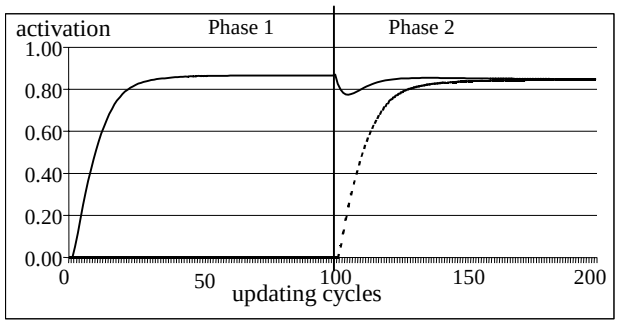


Figure 2A: Simulation of a blocking paradigm (Test 1). Activation of hypothesis nodes for a common-cause model. The solid line represents the activation of H2:x→e1, the dotted line of H3:x→e2. Phase 2 started at the 101st. cycle.

The model shows no blocking for event e2 in the context of the common-cause model. It quickly acquires the belief that there is a causal connection between x and e2.

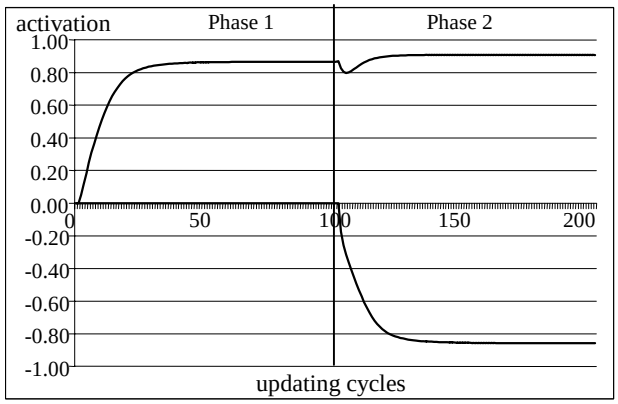


Figure 2B: Simulation of a blocking paradigm (Test 1). Activation of hypothesis nodes for a common-effect structure. The upper line represents the activation of H2:e1→x, the lower line of H3:e2→x. Phase 2 started at the 101st cycle.

For the common-effect model the simulation shows blocking of the second cause, that is the second hypothesis is believed to be wrong. Thus, the simulations closely correspond to the empirical finding that blocking interacts with the structure of the causal model used to interpret the learning data.

Test 2: Testing Complex Causal Hypotheses

The first test of the model used a phenomenon from the literature on causal learning. We now want to turn to a completely different paradigm, hypothesis testing. In experiments on causal learning participants are typically instructed about a causal structure, and the task is to learn about the causal relations within the structure. They are not asked whether they believe that the structure is supported by the learning data or not. In recent experiments (Hagmayer, 2001; Hagmayer & Waldmann, 2001) we gave participants the task to test a complex causal model hypothesis. For example, we asked them whether three observed events support a common-cause hypothesis or not. Normatively this task should be solved by testing the implications of the given structural hypothesis. For example, a common-cause model implies a (spurious) correlation of the effects of the single common cause. In contrast, a common-effect structure does not imply a correlation of the different causes of the joint effect. Unless there is an additional hidden event that causes a correlation among the causes, they should be uncorrelated. In the experiment, participants were given data which either displayed a correlation between all three events (data set 1) or correlations between e1-x and e2-x only, that is e1 and e2 were marginally independent in this data (data set 2). Data set 1 was consistent with a common-cause hypothesis which implies correlations between all three events. In contrast, data set 2 favors the common-effect hypothesis with x as the effect and e1 and e2 as independent causes. However, in a series of experiments we found that participants were not aware of these differential structural implications when testing the two hypotheses. Instead they checked whether the individual causal relations within the complex structures held (e.g., e1-x). Thus, participants dismissed a hypothesis if one of the assumed causal links was missing. However, they proved unable to distinguish between the common-cause and the common-effect structure when both structures specified causal connections between the same events (regardless of the direction).

To model this task we used the model without the nodes for event pairs and individual events. The special activation node was connected to the patterns of three events. As before the activation of the individual event patterns was proportional to the frequency of the respective pattern in the data. To test the model, we used three sets of data. Either all three events were correlated (data set 1), e1 and x, and e2 and x were correlated and e1 and e2 were marginally independent (data set 2), or e1 and x, and e1 and e2 were correlated, and e2 and x were uncorrelated (data set 3). As competing hypotheses we either used a common-cause model with x as the common cause, or a common-effect model with x as the common effect. Figure 3 shows the activation of the node H1 which represents the hypothesis that the respective causal model underlies the observed data.

Figure 3A shows the results for the common-cause hypothesis, Figure 3B for the common-effect hypothesis. The results clearly mirror the judgments of our participants. Whenever the two assumed causal relations within either causal model were represented in the data, the structural hypothesis was accepted (solid lines), if one link was missing the hypothesis was rejected (dotted line).

One slight deviation from our empirical findings was observed. In early cycles there seems to be an effect favoring the common-effect hypothesis with data consistent with this hypothesis. However, the difference between the hypotheses is relatively small and further decreases after 100 updating cycles. Thus, the results are consistent with participants' insensitivity to structural implications of causal models in hypothesis testing tasks.

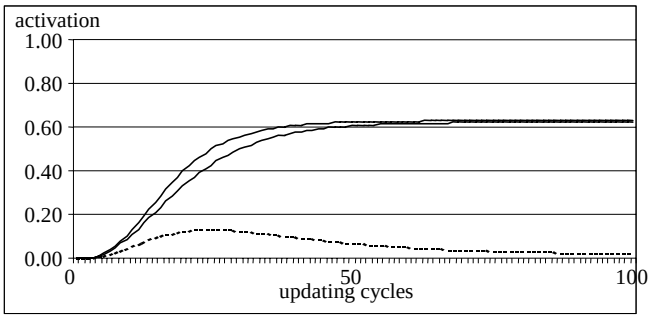


Figure 3A: Activation of hypothesis node H1 for a common-cause model (Test 2). The solid lines represent the activations for data set 1 and 2, the dotted line the activations for data set 3.

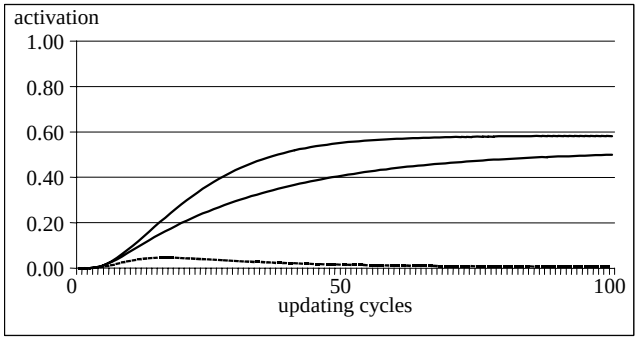


Figure 3B: Activation of hypothesis node H1 for a common-effect model (Test 2). The solid lines represent the activations for data set 1 and 2, the dashed line at the bottom the activations for data set 3

Why does the model not differentiate between the two causal structures? The reason is that it is assumed that complex structural hypotheses are not directly linked to empirical evidence. In our model empirical evidence is connected to the hypotheses that represent individual causal links which in turn are linked to more complex model-related hypotheses. This architecture allows it to model learning and hypothesis testing within the same model. It also seems to capture the empirical finding that participants can easily decide whether a certain pattern of events supports a simple causal hypothesis, but have a hard time to relate event patterns to complex causal hypotheses.

Test 3: Causal Inferences

In the previous section we have mentioned studies showing insensitivity to spurious relations implied by causal models. A last test for our model is a task in which participants have to predict other events under the assumption that a certain causal model holds. Interestingly we have empirically demonstrated sensitivity to structural implications of causal models in this more implicit task (Hagmayer & Waldmann, 2000). In this task participants do not have to evaluate the validity of a causal model in light of observed evidence but rather are instructed to use causal models when predicting individual events. In our experiments we presented participants with two learning phases in which they learned about two causal relations one at a time. Thus, in each phase participants only received information about the presence and absence of two events (x and e1, or x and e2). They never saw patterns of all three events during the experiment. The initial instructions described the two causal relations, which were identically presented across conditions, either as parts of a common-cause model with x as the cause or as part of a common-effect model with x as the effect. After participants had learned about the two causal relations we asked them to predict whether e1 and e2 were present given that x was present. We found that participants were more likely to predict that both e1 and e2 would co-occur when x was viewed as the common cause than when it was seen as a common effect. Thus, in this more implicit task the predictions expressed knowledge about structural implications of causal models. In particular, the patterns the participants predicted embodied a spurious correlation among the effects of a common cause, whereas the causes of a common effect tended to be marginally uncorrelated in the predicted patterns. By contrast, in a more direct task which required explicit judgments about correlations, no such sensitivity was observed, which is consistent with the results reported in the previous section.

To model this experiment we eventually used the complete network depicted in Figure 1 which was successively augmented according to our two principles. In Phase 1, the learning phase, patterns of two events were connected to the hypotheses H2 and H3. Depending on the learning condition, these two hypotheses were either linked to a common-cause or a common-effect hypothesis (H1). The activations of the hypothesis nodes at the end of Phase 1 were used as initial activation values in Phase 2. In Phase 2 the model consisted of the three hypothesis nodes, the nodes for patterns of three events and the nodes representing single events. The single event nodes were included because the task required the prediction of individual events. The special activation node was now attached to event x. The model then predicted the other two individual events and patterns of all three events.

The model quickly learned the causal relations during Phase 1 of the experiment. Figure 4 depicts the results of Phase 2. Figure 4A shows the predictions of the model for the condition in which participants assumed a common-cause model, Figure 4B shows the results for the common-effect condition. The results of the simulations are consistent with the behavior we have observed in our participants. When the model assumes a common-cause model the presence of x leads to a high positive activation of the two effects e1 and e2. This means that the model tends to prefer the prediction that the two effects of a common cause co-occur. In contrast, for the common-effect structure the model does not show such a preference. In this condition, both causes or either one of them equally qualify as possible explanations of the observed effect. This means that our model, similar to the one Thagard (2000) has proposed, tends to "explain away" the second cause when one of the competing causes is present. This is a consequence of the competition between the two causal hypothesis H2 and H3.

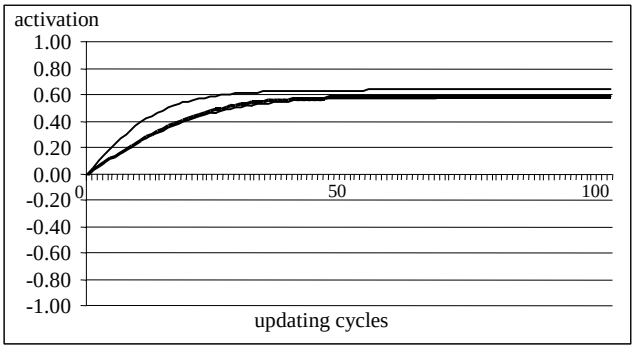


Figure 4A: Implicit causal inferences (Test 3). Activation of single event nodes for the common-cause model: Event x (top), events e1 and e2 (bottom)

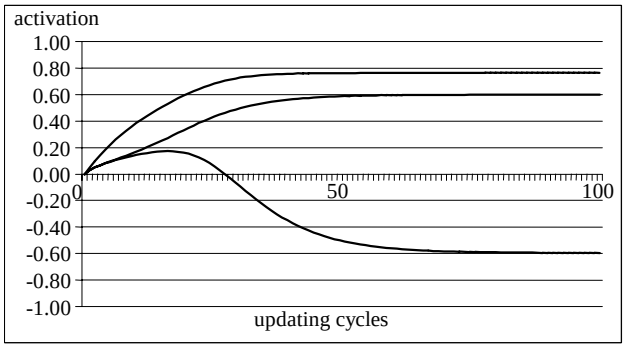


Figure 4B: Implicit causal inferences (Test 3). Activation of single event nodes for the common-effect model: Event x (top), event e1 (middle), event e2 (bottom)

Discussion

A constraint satisfaction model of causal learning and reasoning was presented in this paper that extends the architecture and scope of the model proposed by Thagard (2000). Thagard's model focuses upon causal explanations of singular events and belief updating. Our aim was to create a model that allows it to model both learning and reasoning within causal models. The model was successfully applied to three different tasks. It modeled people's sensitivity to structural implications of causal models in tasks involving learning and predictions whereas the same model also predicted that people would fail in tasks which required explicit knowledge of the statistical implications of causal models.

One question that might be raised is whether the proposed model really captures learning or just models causal judgment. In our view, the concept of learning does not necessarily imply incremental updating of associative weights Our model embodies a hypothesis testing approach to learning which assumes that learners modify the strength of belief in deterministic causal hypotheses based on probabilistic learning input. This view also underlies recent Bayesian models of causality (Pearl, 2000). In the model the activation (i.e., degree of belief) of the hypothesis nodes is modified based on the learning input. This way the model is capable of modeling trial-by-trial learning as well as learning based on summary data within the same architecture.

Thus far we have pre-set the weights connecting evidence and hypotheses. In our view, the assigned values reflect everyday qualitative intuitions about whether an event pattern supports or contradicts a hypothesized causal hypothesis. These weights remained constant throughout the simulations. Despite this restriction the model successfully predicted empirical phenomena in learning and reasoning. However, pre-setting these weights is not a necessary feature of the model. It is possible to add a learning component that acquires knowledge about the relation between event patterns and hypotheses based on feedback in a prior learning phase (see Wang et al., 1998, for a model adding associative learning to Echo).

In summary, our constraint satisfaction model seems to offer a promising new way to model causal learning and reasoning. It is capable of modeling phenomena in a wide range of different tasks, which thus far have been treated as separate in the literature. Relative to normative Bayesian models, our connectionist model allows it to simulate a large number of different tasks and different phenomena while using fairly simple computational routines. It proved capable of capturing a number of recent phenomena that have presented problems to extant models of causal cognition. More tests of the model clearly seem warranted.

References

Hagmayer, Y. (2001). *Denken mit und über Kausalmodelle.* Unpublished Doctoral Dissertation, University of Göttingen.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214-219). Mahwah, NJ: Erlbaum.

Hagmayer, Y., & Waldmann, M. R. (2001). Testing complex causal hypotheses. In M. May & U. Oestermeier (Eds.), *Interdisciplinary perspectives on causation* (pp. 59-80). Bern: Books on Demand.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II. Current research and theory* (pp. 64-99) New York: Appleton-Century-Crofts.

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (Eds.)(1996*). The psychology of learning and motivation, Vol. 34: Causal learning*. San Diego: Academic Press.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Thagard, P. (2000). *Coherence in thought and action.* Cambridge, MA: MIT Press.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 53-76.

Wang, H., Johnson, T.R., & Zhang, J. (1998). UEcho: A model of uncertainty management in human abductive reasoning. In M. A. Gernsbacher & S. R. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1113-1118). Mahwah, NJ: Erlbaum.

How Similarity Affects the Ease of Rule Application

Ulrike Hahn (sapuh@cardiff.ac.uk)
School of Psychology; Cardiff University
Cardiff CF1 3YG - Wales, UK

Mercè Prat-Sala (M.Prat-Sala@wkac.ac.uk)
Department of Psychology; King Alfred's College
Winchester SO22 4NR - England, UK.

Emmanuel M. Pothos (E.Pothos@ed.ac.uk)
Department of Psychology, University of Edinburgh
Edinburgh EH8 9JZ - Scotland, UK

Abstract

There is good theoretical reason to believe that the application of explicit rules does not proceed in a strictly context free manner, whereby the features stipulated by the rule are simply checked one by one. The fact that specifications of general knowledge seem inherently prone to exception suggests that a more flexible approach is required. One way of balancing the simplicity of rules with the need for flexibility is through the combination of rule application with the monitoring of instance-similarity. As a test of this hypothesis, this paper reports an experiment which examines effects of instance similarity on the speed with which a simple explicit rule can be applied, both as a function of experience with the rule and its complexity.

Introduction

Since its very beginnings Cognitive Science has sought to establish the role of rules in human cognition. However, this work has focussed primarily on *internal*, often implicit, rules. With this we mean rules which are internal to the cognitive system and typically unavailable for conscious inspection, and which have no external, public manifestation. Hence this work has had little to say about how we reason with external, explicit rules such as legal rules or explanatory rules provided in educational settings.

This state of affairs is highly unsatisfactory last but not least because evidence for internal rules has been harder to come by than cognitive scientists originally thought, with any claim for rule-based behaviour typically countered by alternative –most frequently similarity-based- explanations, and supposedly supporting data. By contrast, the existence of vast numbers of external rules is beyond doubt. How we reason with these should thus be a central concern.

We present here an experimental investigation of the way in which explicit rules are applied, examining specifically the role of similarity in this process.

General Background

Past research within cognitive psychology has sought evidence for internal rules in a wide range of domains, from language, to implicit learning, reading and problem-solving. Each of these areas has seen intensive debate between proponents of rule-based accounts and proponents of alternative, similarity-based explanations (see e.g., Hahn & Chater, 1998a for an overview). For example, categorization might construed as the application of rules the learner has abstracted during learning (e.g., "If it is furry, four legged and barks, then the creature is a dog") or similarity comparison to known exemplars or a prototype (e.g., "This creature is so similar to Lassie, that it must be a dog"). In this way, rules and similarity have typically been viewed in opposition.

However, detailed computational considerations –which draw lessons from Artificial Intelligence – suggest that neither purely rule-based reasoning, nor purely similarity-based reasoning are optimal or even feasible strategies for real-world tasks (Hahn & Chater, 1998a, 1998b; Oaksford & Chater, 1991). Such considerations lessen the plausibility of any cognitive account that seeks to explain human performance solely in terms of one or the other.

However, similarity- and rule-based reasoning differ in their respective strengths and weaknesses. Hence, computational considerations suggest that 'blends', which combine the strength of both, are an extremely interesting class of account (Hahn & Chater, 1998a, 1998b). This is reflected in a recent interest in hybrid experts systems within AI which encompass both rule and similarity-based components (e.g., Rissland & Skalak, 1991).

Furthermore, it is highly suggestive that law, next to science the single most elaborate and explicit system we have developed for dealing with every-day life, displays both similarity- and rule-based reasoning in the form of precedent and statute. While legal systems differ regarding the relative weight they place on each of these factors (e.g., the Anglo-American legal tradition emphasises similarity to past cases, and the continental tradition emphasises rules),

the 'blend' of both is a robust finding for all western legal systems.

Together, these considerations suggest that research might profitably turn toward studying the potential interplay between rules and similarity in human thought.

The particular way we propose to do this also turns away from the focus on internal rules to external rules which, as the examples of both law and educational instruction show, are of indisputable significance in human cognition.

Previous Research

In comparison to the wealth of research examining either rules or similarity (see Hahn & Chater, 1998a for references), the body of previous experimental research examining a possible *interplay* of rules and similarity is tiny (e.g., Ross, Perkins & Tenpenny, 1990; Allen & Brooks, 1991; Nosofsky, Palmeri and McKinley, 1994). Of this work, only two studies consider explicit, external rules that are given to participants by the experimenter -Nosofsky, Clark & Shin, 1989 and Allen and Brooks (1991). Both find effects of exemplar similarity in the context of rule application. This is first positive evidence for a routine interaction between similarity and the application of explicit rules as might seem desirable in the light of computational considerations. Further examination, however, is required. In particular, the Allen and Brooks (1991) study does not provide the most robust test of the hypothesis that similarity generally influences rule application. Allen and Brooks provided participants with an explicit rule by which a set of training stimuli could be perfectly classified. Despite the fact that the rule was perfectly predictive, i.e. sufficient for classification, and that participants were both aware of the rule and the instruction to use it, their performance on novel items was significantly affected by the test items degree of similarity to items seen during training. However, the nature of the rule used raises worries about the robustness and generality of their findings.

Crucially, the explicit rule used in the experiment defined a prototype. The use of similarity might conceivably be an (artefactual) result of this, rather than a property of rule application in general. Specifically, the rule used in the Allen & Brooks study had the form "if X has 3 of the 5 features{a,b,c,d,e}, then X is a category member". In other words, the rule specifies a so-called **m-of-n concept** and thus is formally equivalent to a prototype (n, i.e. an item with *all* relevant features present) and a threshold (m, i.e. the number of matching features required) which determines the degree of similarity to the prototype which items must have in order to be category members (Langley, 1994; Hahn & Chater, 1998a). This equivalence makes Allen and Brooks' similarity effects rather less surprising. The rule effectively defines a prototype plus similarity threshold, thus virtually suggesting the general use of similarity to participants. Thus, the fact that even similarities irrelevant from the perspective of the rule influenced classification, Allen and Brooks' finding, might have arisen only because the nature of the rule pointed toward similarity in the first place.

To establish more generally a role for similarity in rule application we need to repeat the basic Allen and Brooks study with a rule that does not involve the specification of a prototype. It is crucial to see what happens if the rule is something like "choose all symmetric patterns," or "all patterns that have an even number of corners". Does one still find effects of similarity to training items on classification speed and accuracy with such a rule?

The second limitation of the Allen and Brooks' study concerns rule complexity. The rule used by Allen and Brooks is a fairly complicated one, so that participants might conceivably have had difficulty operationalising it. Thus Allen and Brooks similarity effects might be the result of alternative, "fall-back" strategies on the part of the participants or of errors arising from failure to use the rule, and not –as we would like to see– the result of genuine interaction between rule and similarity-based processing.

This suggests that it is crucial to look also at asymptotic performance, (that is performance after several experimental trials), to see whether influences of previously seen exemplars disappear as participants' accuracy with the rule grows.

Our study seeks evidence for a constructive interplay of similarity and rule application which avoids these limitations. To this end we conduct a simple replication with a different type of rule, varying rule complexity, and monitoring asymptotic performance.

Experimental Investigation

Our materials were simple geometric shapes as depicted in Figure 1 below.

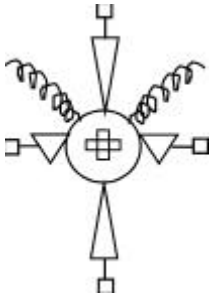


Figure 1: an example of a training material used in the experiment.

Category membership was governed by a simple, explicit rule "*is an A if it has an upside down triangle at the side"* in one-feature rule condition, and by a more complex rule, which made reference to three features –"*is an A if it has an upside down triangle at the sides, a cross in the centre, and a curly line at the top"*- in the complex rule condition. Neither of these constitute an m-of-n concept as did Allen & Brooks' rule. Testing and comparing both a simple and a more complicated rule would further allow us to ascertain the robustness of any putative similarity effect. Specifically, it would allow us to examine whether the similarity effects in the Allen & Brooks study were simply down to participants difficulties in operationalising the rule and, as a result, seeking out alternative, easier strategies.

To the same end, we also monitored participants' performance over four blocks of training –each of which comprised 24 exposures to the test stimulus items. This would allow us to examine whether similarity effects vanished as participants became more and more proficient with the rule.

At the heart of the study is its similarity manipulation. This was achieved by manipulating the additional features not referenced by the rule. Participants received a set of training items to illustrate the rule they had been provided with. The actual experimental test items where either high in similarity or low in similarity to these illustrative training items. The high-similarity test items differed from the training items in one feature; this feature was not referenced by the rule and consequently was irrelevant to the categorization task at hand. The low similarity items differed from the illustration items in 3 features, again features irrelevant to the application of the rule. Because the similarity manipulation concerned only stimulus aspects which were irrelevant to the rule, a difference in the way the high- and low-similarity items were treated would indicate an influence of similarity on the rule-based classification process, despite these differences in similarity being completely irrelevant to the application of the rule. This would provide evidence for an automatic monitoring and processing of similarity information in the context of rule application even where there were no task demands to necessitate this.

To make the classification task a genuine one, it was necessary to have both test items that complied with the rule and ones that didn't. The non-compliant test items, again, were either high- or low- similarity to the illustrative test items, differing from these items in a single feature which contradicted the rule, in the case of the high-similarity non-compliant items, or in a rule-feature and two further irrelevant features in the case of the low-similarity, non-compliant items. As a consequence, similarity effects could emerge both where the rule was applicable and where it did not apply.

The central prediction of this study was that there should be a difference in the speed with which the rule was applied between high- and low-similarity items. Specifically, compliant items which were also high in similarity to the training items should be classified more quickly than compliant items which were low in similarity. Differences should also emerge between the non-compliant high- and low-similarity items, though the direction of the difference is harder to predict here; low-similarity items might be rejected more quickly, but advantages at the decision-making stage are likely to be offset to greater or lesser extent by increased costs of processing a less familiar image. In addition, we would expect an effect of training on reaction time, and that the three feature rule should be slower to apply than the one feature rule because it requires a more complete scan of the stimulus, although these two predictions have no bearing on our experimental question.

With regards to similarity effects, we would also expect differences in the amount of errors elicited by high- and low-similarity items. Low-similarity compliant items, and high-similarity non-compliant items expected to be more error prone than their counterparts.

Method

Participants

Ninety-one undergraduate students from the University of Bangor (Wales) participated in this experiment as an extra credit option in a Psychology course. 45 participated in the simple rule condition while 46 took part in the complex rule condition.

Materials

The stimuli were line drawings of geometric shapes that varied in 6 aspects: Body, Side Ears, Top/Bottom, Inner, Antenna, and Hair. There were six alternative realisations of each of these aspects. We generated a total of 108 stimuli: 12 training items and 96 test items. Figure 1 gives an example of a training material used in the experiment.

The 96 test items were composed of four different types of items which formed to four conditions: 24 high-similarity (C-High) and 24 low-similarity (C-Low) (items that complied with the rule), and 24 high-similarity (N-High) and 24 low-similarity (N-Low) (items that did not comply with the rule). High-similarity items differed from (one of the items of) the training set in one value. Low-similarity items differed from (one of the items of) the training set in three values. For the items that complied with the rule, the differed value was always an irrelevant feature. For the items that did not comply with the rule, the differed feature was a value of the rule for the high-similarity items, and a value of the rule plus two irrelevant values for low-similarity items.

Apparatus

The experiment was controlled by the ERTS software run on a PC computer. The pictures were presented as black line drawings on a white background on a (640x480) VGA monitor. Participants used a two-key response pad attached to the ERTS EXKEY-logic connected to the computer to express their response. One of the key was labelled YES and the other was labelled NO. Participants used their dominant hand to press the YES key. Participants' key responses and time elapsed from the presentation of an item to the participants' response were recorded.

Procedure

Participants were tested individually. They were seated in a quiet booth at a comfortable viewing distance in front of a monitor. They received the instructions displayed on the monitor. The instructions told them that they would see a series of pictures of abstract objects. Some of the pictures corresponded to 'good objects' and some to 'bad objects'. Participants' task consisted of pressing the key labelled YES if the abstract object was a 'good object' and press the key labelled NO if the abstract object was a 'bad object'. Before the initiation of the experiment, participants were presented with 10 trials with either YES or NO displayed

on the screen to familiarise them with pressing the YES/NO keys. Participants were given a rule (see above) to determine whether an object was a good object or not. Immediately after the written description of the rule, participants received a graphic example of the rule.

The experimental session was divided into 5 blocks: there was 1 training block and 4 test blocks. During the training phase participants were given 36 trials made up of 12 training items seen 3 times each in a random order. Participants controlled the speed of presentation of each of the training items using the space bar.

After the training phase participants were given four blocks of test. Each test block consisted of 24 items selected randomly from the total of 96 test items, with the only constraint that each test block contained 12 items that comply with the rule (6 high-similarity items and 6 low-similarity items) and 12 items that broke the rule (6 high-similarity do non-compliant items and 6 low-similarity non-compliant items). The test items of each block were presented in a different random order to each participant.

After each of test block participants were reminded of the rule and were shown the 12 items from the training phase (this time they saw each training item only once). This was done with the aim to reinforce rule application. As during the training phase, this repeated training phase was controlled by the participants using the space bar. Participants were asked to respond as quickly as possible without compromising accuracy.

Results

We begin with the *error* analysis. In the case of the simple, 1 feature rule, 123 (2.84%) from a total of (96x45) 4320 responses were errors, that is the participant either pressed the YES key when the NO key was appropriate according to the rule or the other way round (see Table 1). For the complex, 3 feature rule, the data of 4 participants was not included because they failed to respond using the correct key for all the trials of at least one of the blocks in one of the conditions. From a total of (96x42) 4032 responses, 305 (7.56%) were errors were the participant pressed the YES key when the NO key was expected or the other way round (see Table 1).

Table 1: Errors for both 1 and 3 feature rules. "C" indicates items which comply with the rule, "N" items which violate it, "High"- and "Low" refer to the degree of similarity to the test items.

	C-High	C-Low	N-High	N-Low
Errors 1	26	33	36	36
Errors 3	46	61	92	107

We begin with the analysis of the simple, 1 feature rule. Though there were slightly more errors on the low similarity items which complied with the single rule, as we had expected, a paired t-test comparing the proportion of errors made by each subject in across high vs. low similarity comply did not reach significance ($t(44) = 1.26$, $p = 0.1$, one-tailed). The expectation that on the non-rule compliant items there should be more errors on the high similarity items than on the low similarity non-compliant items was not born out at all, with equal numbers of errors in both cases. In summary, the error analysis for the simple rule revealed no effects of similarity on rule application.

A different picture is presented by the error data for the complex, 3 feature rule. Again, there were more errors on the low-similarity compliant items than on the high-similarity compliant items as expected, but here the difference is statistically significant ($t(41) = 2.10$, $p < 0.02$, one-tailed). Again, the expectation that, of the non-compliant items, it should be the high similarity ones which elicit more errors was not born out in that participants made *more* errors on the low similarity, non-compliant items (Table 1); however, this unexpected difference, tested post hoc, was not significant ($t(41) = 1.48$, $p = 0.115$, two-tailed). In summary, at least for those items which comply with the rule, a significant effect of similarity on rule application emerged.

Finally, a comparison of the error proportions for the simple and complex rule revealed the expected higher level of errors in the complex rule condition: two 2-way Analysis of Variance (ANOVA) (one for the compliant items and another for the non-compliant items) with similarity (High vs. Low) within subject and rule (simple vs. complex) between subject analysis showed a main effect of rule complexity for the non-compliant data ($F(1,85) = 18.56$, $p < 0.0001$) but not for the compliant data ($F(1,85) = 2.21$).

In summary, the analysis of the error data revealed similarity effects for the complex rule, but not the simple rule. We next ask whether this finding is confirmed by participants' reaction times.

Reaction Time Analysis. We begin with the simple, 1 feature rule condition. For each participant we calculated the mean reaction time of response for each condition per block, giving us 16 data points per participant. These data point were transformed into their logarithm. This formed the bases for our analyses. Table 2 shows the mean reaction time per condition and block across all participants.

Table 2: mean reaction time across participants for one-feature rule.

	C-High	C-Low	N-High	N-Low
Block 1	565	587	662	626
Block 2	541	540	548	521
Block 3	514	515	530	583
Block 4	534	513	551	528

We analyse the compliant and the non-compliant items separately, not only because we expect different patterns of result, but because two different hands were used for YES and NO responses and therefore the RT's from the dominant hand are not directly comparable to those of the non-dominant hand.

A two-way ANOVA (fully within) was performed on the reaction time data for the *compliant* items. The variables were block (1-4) and similarity (high vs. low). A main effect of block was found ($F(3,132) = 7.15$, $p < 0.0001$). No main effect of similarity was found ($F(1,44) = .083$) and there was no interaction ($F(3,132) = .102$).

A further two-way ANOVA (fully within) was performed on the reaction time data for the *non-compliant* items. Again a main effect of block was found ($F(3,132) = 29.32$, $p < 0.0001$). No main effect of similarity was found ($F(1,44)= .804$) but the interaction was significant ($F(3,132) = 8.92$), $p < 0.0001$).

Repeating these analyses for the complex rule condition, we calculated, for each participant, the mean reaction time of response for each condition per block, giving us 16 data points per participant. These data points were transformed into their logarithm to minimise individual differences. This formed the bases for our analyses. Table 3 shows the mean reaction time per condition and block across all participants.

Table 3: Mean reaction time across participants for complex-rule condition.

	C-High	C-Low	N-High	N-Low
Block 1	1166	1276	1056	1166
Block 2	1018	1077	815	837
Block 3	934	976	723	736
Block 4	902	939	732	780

A two-way ANOVA (fully within) was performed on the reaction time data for the *comply* condition. The variables were block (1-4) and similarity (high vs. low). A main effect of block ($F(3,123) = 13.67$, $p < 0.0001$) and of similarity ($F(1,41) = 18.42$, $p < 0.0001$) was found. However, the interaction did not reach significance ($F(3,123)= 1.58$).

A further two-way ANOVA (fully within) was performed on the reaction time data for the *non-compliant* items. Again a main effect of block ($F(3,123) = 69.69$, $p < 0.0001$) and of similarity ($F(1,41)= 12.78$, $p < 0.001$) was found. However, the interaction did not reach significance ($F(3,123) = 1.72$).

Finally, comparing the simple and the complex rule condition, participants were significantly faster classifying on the basis of the simple rule, as predicted for both the compliant condition (($F(1,85) = 173.93$, $p < 0.0001$) and the non-compliant condition ($F(1,85) = 121.80$, $p < 0.0001$).

We illustrate the implications of the above analyses of reaction times with reference to two graphs plotting the mean reaction times for both kinds of rules, and high and low similarity items, with one graph each for the compliant (Figure 2) and the non-compliant items (Figure 3).

The main interest of the experiment, lies in potential differences between high- and low-similarity items, both as a function of rule complexity and of the amount of training.

As can be seen from Figure 2, the difference in reaction time for high and low similarity items is much greater for the complex rule than it is for the simple rule.

The difference also does not seem to change much with practice. These informal observations are confirmed in the above analyses in that there is no consistent effect of similarity in the case of the simple rule (the ANOVA found no main effect of similarity, nor any interaction between similarity and block) while there is an effect of similarity for the complex rule. However, this effect does not change significantly with practice, that is across blocks (the ANOVA showed a main effect of similarity, but no interaction).

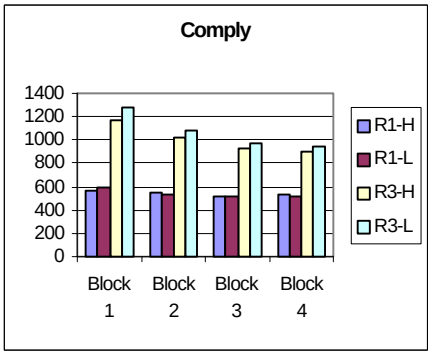


Figure 2: Plotted along the x- axis is the number of blocks, plotted along the y axis is the mean reaction time. R-1 refers to the simple rule, R-3 to the complex, and H and L stand for high and low similarity to training items.

The non-compliant items deviate from this picture only slightly. As can be seen from Figure 3, the magnitude of the difference between high- and low-similarity is again greater for the complex rule, which also shows no clear effect of practice. This is confirmed by the above statistical analysis: for the non-compliant items in the complex rule condition there is, once again, a significant main effect of similarity, but no interaction between similarity and block. The data for the non-compliant items of the simple rule are not quite as clean showing an anomalous increase on block three. As a result of this increase, there is a significant interaction between similarity and block, but as before with the compliant items, there is no main effect of similarity. Crucially, the failure to show a consistent similarity effect observed in the compliant items is thus replicated in the non-compliant items.

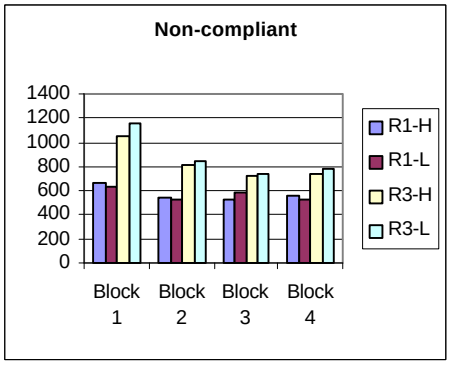


Figure 3: Plotted along the x- axis is the number of blocks, plotted along the y axis is the mean reaction time. R-1 refers to the simple rule, R-3 to the complex, and H and L stand for high and low similarity to training items.

In summary, then, we find similarity effects only for the complex rule, not the simple rule –a result which corresponds with the findings from the error analysis. Furthermore, there is no evidence that this similarity effect goes away with increased practice.

Discussion

The most important result of this experiment is that it manages to replicate the previously observed intrusion of similarity on rule-application even though there was nothing about the rules themselves which in any way made similarity seem relevant.

The similarity effect was apparent both in the error patterns and in participants reaction times and the finding that the similarity effects (where present) did not seem to abate with increased practice in applying the rule further underscore the generality of the observed interplay between rule-application and similarity assessment.

However, we failed to find a consistent effect of similarity in the simple rule condition. One must be cautious in interpreting such a null-effect in a single study in that an effect might well be observable given slight changes in experimental procedure or a considerably larger sample. However, the failure to find an effect does suggest that the role of similarity is not independent of rule-complexity and in that sense neither ubiquitous nor completely automatic. In contrast, the similarity effect found on the complex rule is very robust.

Why might one find effects of similarity in the context of rule application at all, and can our differential results for rules of different complexity be linked into an explanation? The weakness of rule-based reasoning lies in the fact that it is so exceedingly difficult to come up with perfectly predictive rules. Most regularities seem to be prone to countless exceptions, or hold only relative to certain background assumptions which are virtually impossible to capture. For example, the rule "birds fly" is true by and large, but there are some birds which don't. Making rules more specific doesn't eliminate the problem: "robins fly" seems true enough, but, again, will be false if the robin in question has broken its wing, has its feet stuck in concrete, or has eaten too many worms…The potential exceptions are endless and there seems to be no clean cut way of ruling them out in advance. These difficulties have dogged rule-based approaches within Artificial Intelligence which were once held to lead to "expert behaviour" within decades. The frame problem, the difficulties encountered by the naïve physics project (Hayes, 1979), and the difficulties in formalising defeasible inference (e.g., Reiter, 1980) are testimony to these difficulties (see also Pickering & Chater, 1995, for discussion).

Our assumption is that similarity might go some way toward alleviating these difficulties, thus allowing human beings to harness the undoubted power and clarity provided by explicit rules. Specifically, tracking the similarity of a potential candidate instance of the rule to previously encountered instances may provide a means whereby one is alerted to potentially deviating circumstances, where the rule –though seemingly applicable- does not, in fact apply. The fact that a novel instance seems dissimilar from previous instances might be a clue to the fact that it should actually be treated differently. This, we would argue, is one of the reasons why an interplay between rules and similarity seems profitable, and hence, is pervasive in large scale legal systems such as the law (see Hahn & Chater 1998a and 1998b for fuller discussion). If this is true, tracking similarity in the context of rule application would seem useful in most contexts, but we might expect greater reliance on similarity under conditions of increased uncertainty. One way of looking at the increase in rule complexity is as an increase –for participants- in uncertainty. This also seems compatible with Nosofsky, Clark and Shin's (1991) observation of similarity effects despite an explicit rule as their task required fine-grained perceptual distinctions along continuous valued dimensions, which it seems unlikely could be achieved exclusively by the rule. However, further experimentation is required to establish whether the uncertainty reduction is indeed central. The most obvious experimental path to pursue is that of increasing uncertainty in the case of the simple rule, for example by introducing exceptions, thus making the classification task "noisy", or by making the differences in appearance even more extreme. In the meantime, what our results suggest is that instance similarity has some role to play in rule application – a role which needs both further clarification and a satisfactory explanation.

Acknowledgements

The work reported in this paper was funded by a grant from the Nuffield Foundation.

References

Allen, S. & Brooks, L. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120, 3-19.

Hahn, U and Chater, N. (1998a). Similarity and Rules: Distinct? Exhaustive? Empirically Distinguishable? *Cognition*, 65, 197-203

Hahn, U. & Chater, N. (1998b). Understanding Similarity: a Joint Project for Psychology, Case-Based Reasoning and Law. *Artificial Intelligence Review*, 12, 393-429.

Hayes, P. (1979). The naïve physics manifesto. In D. Michie (Ed.), *Expert systems in the Micro-electronic Age*. Edinburgh University Press, Edinburgh.

Langley, P. (1994) Machine Learning. Addison Wesley

Nosofsky, R., Clark, S. & Shin, H. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 282-304.

Nosofsky, R., Palmeri, T. & Mckinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53-79.

Oaksford, M. & Chater, N. (1991). Against Logicist Cognitive Science. *Mind and Language*, 6, 1-38.

Pickering, M. and Chater, N. (1995). Why cognitive science is not formalized folk psychology. *Minds and Machines*, 5, 309-337.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13, 81-132.

Rissland, E. & Skalak, D. (1991). CABARET: rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies*, 34, 839-887.

Ross, B., Perkins, S. & Tenpenny, P. (1990). Reminding-based Category Learning. *Cognitive Psychology*, 22, 460-492.

Modeling Grouping with Recursive Auto-Associative Memory

Andreas Hansson (andreas.hansson@ida.his.se)
Department of Computer Science, University of Skövde, Box 408
541 28 Skövde, Sweden

Lars F. Niklasson (lars.niklasson@ida.his.se)
Department of Computer Science, University of Skövde, Box 408
541 28 Skövde, Sweden

Abstract

Sometimes humans have a need for storing long sequences of information in memory. Several experiments show that grouping the items in the sequence helps storing the sequence in auditory short-term memory. One architecture used by connectionist cognitive researchers when representing and processing sequences is Recursive Auto-Associative Memory. One of the aspects of it is that its capacity for storing sequences is limited, leading to that the longer the sequence the less likely it is that the entire sequence can be recalled; the deepest parts of the sequence are forgotten. Two experiments are performed to test if grouping affects storage in Recursive Auto-Associative Memories. We conclude that grouping affects the ability for storing sequences in Recursive Auto-Associative Memories much in the same way as it affects the human auditory short-term memory, i.e., using grouping increase the probability of that the sequence can be recalled correctly.

Introduction

One technique for studying how memory is constructed is serial recall (see for example Baddeley, 1999 and Bridges and Jones, 1996; Pickering, Gathercole, Hall and Lloyd, 2001). A sequence of items (for example objects, digits, letters, etc.) is presented to the research subjects. The task for the subjects is to recall all items in the sequence in the same order as they were presented.

One of the aspects studied using serial recall tasks is how to increase the ability of the subjects to recall longer sequences.

Martin and Fernberger (1929) performed such a study regarding the improvement of auditory short-term memory. They concluded that the ability to remember sequences increased, if the objects in the sequence were organized in groups. Wickelgren (1964) and Pollack, Johnson and Knaff (1959) among others later confirmed this.

Grouping entails that the sequence is divided into shorter sub-sequences. The actual grouping has in previous experiments been done in two ways. Either the size of the groups is decided by the experiment leader (cf. Wickelgren (1964); Pollack, Johnson and Knaff (1959)) or the size of the groups is selected by the subjects and can vary within a sequence (cf. Martin and Fernberger (1929); Baumann and Trouvain (2001)).

Inspired by the findings of Martin and Fernberger (1929); Pollack, Johnson and Knaff (1959) and Wickelgren (1964), that grouping can increase the ability to recall longer sequences, the interest here is to test how grouping affect the recall of sequences in Pollack's (1990) Recursive Auto-Associative Networks (RAAMs). A RAAM is a type of artificial neural network used for representing sequences and structures of unknown or dynamical size. It has previously been used in experiments involving manipulation of structured objects (e.g. sentence transformation (Chalmers, 1990), language translation (Chrisman, 1991). These experiments showed that RAAM uses the sequential order of the presented objects to develop highly structured (spatial) internal representations. In the experiments presented in the following a variant of these networks, called Extended Recursive Auto-associative Networks (ERAAM), originally suggested by Niklasson and Sharkey (1992), will be used.

Grouping experiments

The exact definition of what a *group* is varies. According to Wickelgren (1964), Fraisse (1945) defined a group as a cluster of correct items separated by one or more errors. Wickelgren noted that this way of defining groups assumes that the items in a group are rarely forgotten and that the subjects practically never can remember two groups in succession. Others, for example Baumann and Trouvain (2001), used rhythm of speech and intonations for finding groups when the subjects recalled the sequences. Wickelgren defined a *grouping method* as a method for *rehearsal*, i.e. grouping in twos means rehearsing the items in twos; grouping in threes means rehearsing the items in threes; etc.

A large number experiments on how grouping affects the ability to recall longer sequences have been conducted. Most of these experiments confirm each other in that they show similar results.

Martin and Fernberger (1929) trained subjects to remember long sequences. This was achieved by the subjects using increasingly larger groups (first the subjects used groups of size two, after a while groups of size three and so on). Martin and Fernberger noted that the performance increased up to groups of size 5, after that the performance decreased.

Pollack, Johnson and Knaff (1959) confirmed these results using temporal grouping (i.e. making a short pause between the groups when presenting the sequence for the subjects) for items presented in 1's, 2's, 3's, 4's and 6's. According to their experiment, using groups of size four resulted in the best overall performance. Note however, that they did not test groups of size five, which Martin and Fernberger (1929) concluded was better than other group sizes.

Wickelgren (1964) also conducted experiments on how different groupings affected recall. According to that study groupings in threes were the best, closely followed by grouping in fours. The worst was grouping in twos, closely followed by grouping in fives. Groups of one item resulted in a performance quite close to the five-grouping, see figure 1.

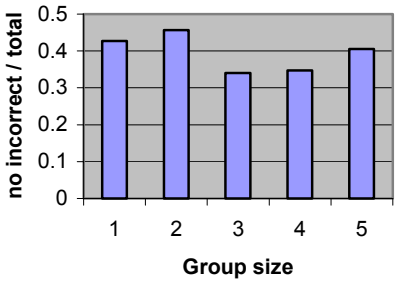


Figure 1: Results from Wickelgren's (1964) experiments for grouping in human short-term memory. Error rate is used to measure the subjects' performance.

In his experiments, Wickelgren used *error rate* to compare performance using different groupings. The error rate for a group is calculated by dividing the number of incorrectly recalled sequences (all items must be recalled in the correct order for the sequence to be regarded as correctly recalled) with the total number of sequences.

ERAAM

As its name suggests, the Extended Recursive Auto-Associative Memory (ERAAM), originally suggested by Niklasson and Sharkey (1992), is an extension of Pollack's (1990) Recursive Auto-Associative Memory (RAAM), which is a connectionist architecture able to represent dynamically large structures (for example sequences) in a fixed-sized artificial neural network. Due to several of its features it has been used by several cognitive researchers who want to investigate theories for sequence recall (see for example Adamson and Damper (1999); Blank, Meeden and Marshall (1992)). One of the features is that the network is trainable, i.e. that it can learn to represent, for example, a specific type of sequence. Another feature is that RAAM during recall is more likely to produce decoding errors the more complex the sequence is. The errors usually appear towards the end of the sequence, quite like humans (cf. Henson (1996), Baddeley, (1999) and Adamson and Damper (1999)).

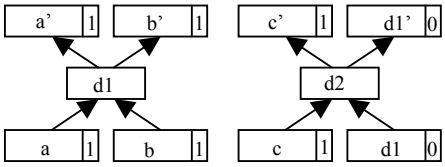


Figure 2: The structure of ERAAM. The extra bit in the input and output layers is used to classify if the representation is a terminal or a composite structure. *d1* is the compressed representation of *a* and *b*. *d2* is the compressed representation of *c* and *d1*. Due to decompressing errors we do not get *exactly* the same representation as we put in. To illustrate this, *a'* and *b'* are the representations from the decompressed *d1* and *c'* and *d1'* results from the decompressed *d2*. Note that the RAAM requires fixed valence on the structures represented.

In RAAMs two networks are involved: a compressor and a decompressor. The compressor consists of an input layer where an item in the sequence is presented together with a compressed representation of the previously presented items of the sequence and an output layer that contains the compressed representation of new item in combination with the previous (in figure 2 'd1' is a compressed representation of the combination of the terminals 'a' and 'b' and 'd2' is a compressed representation of the terminal 'c' and the non-terminal (i.e. previously compressed) 'd1'). The decompressor consists of an input layer containing the compressed representation and an output layer containing the (partially) decompressed representation. The compressor and decompressor networks respectively are used recursively to do further compressions or decompressions of sequences.

When decompressing, a terminal test is normally performed on the decompressed representations to see if any of them is a terminal and therefore should not be further decompressed. The terminals contain the representation of the items in the sequence. Different alternative methods for this terminal test have been proposed (see for example Chalmers (1990); Pollack (1990)). ERAAM is another suggestion for how to interpret if a decoded representation is a terminal or non-terminal. In ERAAM an extra bit of information has been added to each part of the sequence. If the representation is a terminal this bit is set to 1 and otherwise 0, see figure 2. This extra bit is compressed and decompressed along with the other information. When, during decoding, a 1 is encountered as the last bit in the representation, this is interpreted as the representation being a terminal. The representation is then compared to the representations for the known terminals and the terminal that is closest in Euclidean space is the one said to be present in the output.

Experiment 1

The purpose of this experiment is to see how grouping affect the ability to recall longer sequences in ERAAMs. In the experiments Wickelgren's (1964) definition of grouping as a method for rehearsal is used, i.e. the ERAAMs rehearse the groups in the sequence as well as the complete sequence. The group sizes vary from 1 to 6, but the same group size is used throughout the sequence. According to Baddeley (1999) grouping can be used as long as the subjects can notice the presence of the groups. In order to mark the boundaries of groups a 'nil' character is used marking the beginning of a new group.

Method

Sequences The sequences are entirely made up by digits between 0 and 9 and 'nil' representing spacing between groups. The digits and the 'nil' character have orthogonal representations so as not to give any benefit of grouping several representationally similar digits together. The sequences used are produced by picking a random digit for each position in the sequence. However, extra constraints are used. Each digit may only appear once during the first ten positions in the sequence, only once between the eleventh and twentieth position and only once between the twenty-first and thirtieth position, this to prevent that the probability for a specific digit to appear is larger than for the others thereby making the sequence a bit easier to learn. Furthermore, for groups of size two a specific digit may not occur in the same position *within* a group twice for the first twenty positions in the sequence. For groups of size three no digit may appear in the same position in a group during the first thirty items, and so on. This is done to ensure that the probability for a group to contain a specific digit in a specific position should be similar to the probability of other digits thereby giving as little advantage as possible to any digit combination in a group and as little advantage as possible to any specific group size.

Three sequences are generated following the above constraints:

- 97250831645982164073516,
- 58472096133509847261919,
- 25713680491368049257804

From each of these sequences twelve subsequences of are created varying in lengths from 12 to 23 digits. The subsequences of length 12 are constructed from the twelve first digits in the above sequences, the subsequences of length 13 are constructed from the thirteenth first digits and so on.

In order to separate the groups that the subsequences are divided into a 'nil' character was inserted marking the beginning of each group. This is done after the groups are constructed (see figure 3).

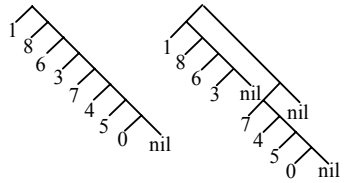


Figure 3: example of how the sequences are organized. To the left an ungrouped sequence, to the right a sequence with group size of four. Nil is used as an end marker.

ERAAMs 30 Extended Recursive Auto-associative memories are used having 2*11 sigmoidal input and output nodes and 10 sigmoidal hidden nodes. Each ERAAM has a unique and random initialization so as to minimize the probability that the experiments, by coincidence, give bias to some specific group size.

Procedure Each subsequence is trained on all the thirty ERAAM networks. Each network is trained using backpropagation on the given subsequence for 200,000 iterations. This number of iterations is established by several dry runs, where almost no changes in the learning of the network are detected after 50,000 iterations and no changes are detected after 150,000 iterations. The limit is set to 200,000 iterations to give the ERAAMs plenty of time to learn the subsequences, but still provide an upper limit to the training time.

The ERAAMs receive the subsequence one item at a time, but each group in the subsequence is trained to be reproduced separately.

The ERAAMs are tested after 200,000 iterations on whether or not they can compress and decompress the entire subsequence they have been training on.

The ERAAM is considered successful in this experiment if the entire recalled subsequence is compressed and decompressed correctly, i.e. there must not be a single error anywhere in the decompressed subsequence.

Results and Discussion

The proportion of errors in relation to number of trials in each group (i.e. the same measure as used by Wickelgren, 1964) can be seen in figure 4. For group size 1 the proportion of errors for sequences of length 12 to 23 is .748. For groups of size 2 it is .820. For groups of size 3 it is .768. For size 4 it is .745. For size 5 it is .729 and finally for groups of size 6 it is .769. It seems as if, analogous to Wickelgren's (1964) results, using a group size of 2 yields the worst result. We can also see that when using larger groups the error rate decreases until the use of group size 6. However, the performance is never much better than that achieved using an ungrouped sequence.

In the experiments the performance declines in a sigmoidal fashion from almost correct recall of ungrouped sequences of length 12 to none correctly recalled at sequence lengths of 20 or more.

The results acquired indicate that grouping *has* an effect on the ability to recall sequences in ERAAMs. It is better to

group a sequence into fours and fives than into twos with respect to the probability of recalling the sequence correctly. However, the ungrouped sequence is about as likely to be recalled correctly as using groups of size four and five. There are two possible reasons for this.

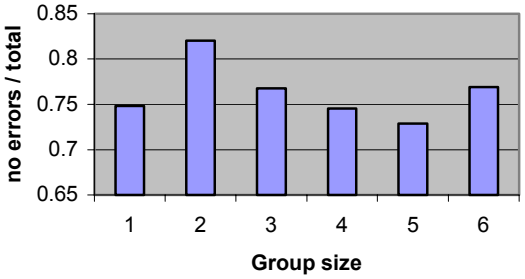


Figure 4: summary of results for experiment 1. As we can see, in almost 75% of the trials the ERAAMs fail to correctly encode and decode the entire sequence. We can also see that grouping for the most sizes makes this worse. However when using groups of sizes four and five there are fewer errors made than in the ungrouped sequences.

The first one is that grouping has no advantage in sequence recall compared to the ungrouped case.

The other reason is that in the grouped sequences there are extra characters inserted: the 'nil' characters. Since they are items as well as the digits this has the effect that the sequences get longer; more items need to be correctly recalled in the grouped sequences than in the ungrouped. This makes the task harder. If this is indeed the case, then it can be noted that the recall of sequences divided into groups of size four and five have the same performance as the ungrouped sequence, actually indicating that grouping has influence. To test whether this is the case, another experiment was performed where the 'nil' characters was removed.

Experiment 2

Whereas the performance in experiment 1 never gets much better for a grouped sequence than for an ungrouped sequence, an interesting effect can be seen in that the use of group size 2 is the worst, but, as the size of the groups increase, the error rate decreases leaving at the group size of 5 a result about the same as an ungrouped sequence before getting worse again with groups of size 6. What differs between the grouped sequences, other than the group size, is the number of 'nil's used, being the most in groups of size two and then gradually decreasing with larger group sizes. Since 'nil' is an extra character, in practice, the sequence gets as much longer as the number of groups used than if it had been ungrouped.

To test whether or not the 'nil' character, since it makes the sequences longer, makes it more difficult for the ERAAMs to learn the sequences another round of experiments are performed, this time without the 'nil' marking the boundaries of the groups.

Method

Sequences The same subsequences as the ones used in experiment 1 are also used here. The difference between the subsequences in this experiment and the previous is the grouping method. In this experiment 'nil's are not used to mark group boundaries, see figure 5. This leads to that the subsequences contain the same number of items regardless of the group size used, whereas in the previous experiment the 'nil's constitute extra items making the subsequences longer.

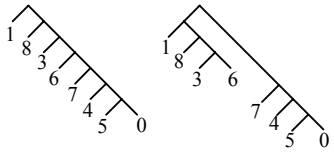


Figure 5: example of how the sequences are organized. To the left an ungrouped sequence, to the right a sequence with group size of four.

As in experiment 1 the size of the subsequences varies between 12 and 23.

ERAAMs The same ERAAMs as in experiment 1 are used; 30 differently initialized ERAAMs of size 2*11 sigmoidal input and output nodes and 10 sigmoidal hidden nodes.

Procedure As in experiment 1 the ERAAMs are trained to compress and decompress the given sequence for 200,000 iterations using backpropagation. The ERAAMs are reset to their initial configuration between each sequence so that the training of one subsequence does not affect the next.

After training, the ERAAM is tested whether or not it can compress and decompress the entire subsequence. As before, the entire subsequence must be correctly reproduced for successful result, otherwise the ERAAM is considered to have failed in representing the subsequence.

Results and Discussion

The error rates for different groupings can be seen in figure 6. For groups of size 1 the overall error rate (the number of correctly compressed and decompressed sequences divided by the total number of sequences used) is .577, for size 2 it is .482, for size 3 it is .402, for size 4 it is .358, for size 5 it is .398 and for groups of size 6 it is .370. It seems as if grouping the sequence leads to better performance, continuously improving until group size 4 when the performance cease to improve any further, see figure 6.

There is now a clear benefit shown of using grouping when trying to recall longer sequences. Using the group size of four yields a 41% better result (the difference between the number of incorrectly recalled sequences for an ungrouped sequence and the number of incorrectly recalled sequences using group size four divided by the number of

incorrect ungrouped sequences) than trying to represent an ungrouped sequence.

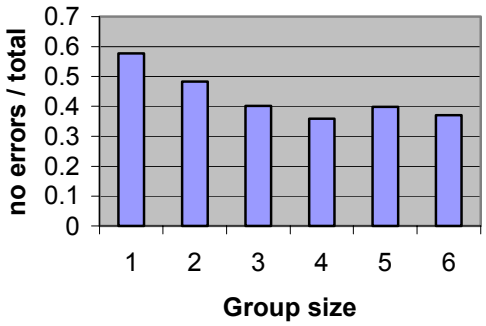


Figure 6: summary of results experiment 2. As we can see, not grouping the sequence here leads to that fewer sequences are recalled than in the grouped cases. We see that using groups of sizes four, five and six results in that more sequences are correctly recalled.

We can also see that this way of grouping the sequences results in that more sequences are correctly recalled than when using 'nil's as group separators.

General Discussion

Sometimes there is a need to store dynamically long sequences in a memory. One of the techniques used by connectionist cognitive researchers, is to compress the sequences using recursive auto-associative memories (RAAMs). However, as in human memory, an aspect of the RAAM architecture is that the longer the stored sequence is the more probable it is that the deepest parts of the sequences cannot be recalled correctly, due to some compressing and decompressing errors. Since grouping a sequence increases the probability that it is correctly recalled in human memory we set out to test if grouping has a similar effect when recalling sequences in Extended Recursive Auto-Associative Memories (ERAAMs).

The experiments reported here show that grouping has a large impact on the probability that the sequences are correctly recalled. The probability that a sequence between the size of 12 and 23 is recalled correctly is on average 41% better using the group size of four instead of not using grouping at all.

The problem when recalling long sequences is that the deeper into the sequence, the more likely it is to miss-recall an item due to cumulating compressing and decompressing errors. We believe that grouping, dividing the sequence into smaller sub-sequences that are linked, works since it decreases the depth that needs to be recalled. The larger the groups the more the depth is decreased. After a while, however, the sub-sequences get so large that they also start to suffer from errors. This means that, as in humans, using increasingly large groupings the performance starts to decrease again.

In many situations there is a need to store long sequences. It is well known that grouping the sequence makes it easier to recall in auditory short-term memory. We see that this is also true for ERAAMs; grouping the sequences that are to be stored leads to an increased probability that they are correctly recalled.

What is suggested here is that RAAM like architectures can be used to model the human ability to store and recall sequences. However, many questions demand answers before networks of this kind can be said to model all the aspects of human memory. This includes questions concerning biological plausibility, a performance more closely matching the human, especially when using meaningful subsequences, etc. The results presented here show that RAAM like architectures indeed have a promising potential for supplying the answers to these questions.

References

Adamson, M. J. & Damper, R. I. (1999). B-RAAM: A Connectionist Model which Develops Holistic Internal Representations of Symbolic Structures. *Connection Science*, 11(1), 41-71.

Baddeley, A. D. (1999). *Essentials of Human Memory*. Psychology Press Ltd.

Baumann, S. & Trouvain, J. (2001). On the Prosdy of German Telephone Numbers. *Proceedings Eurospeech 2001 Scandinavia*, 557-560.

Blank, D.S., Meeden, L.A., and Marshall, J. (1992). Exploring the Symbolic/Subsymbolic Continuum: A case study of RAAM. In The Symbolic and Connectionist Paradigms: Closing the Gap.

Bodén, M. (1994), On Biased Learning for Generalisation, in *Proceedings of the International Conference on Neural Information Processing*, Seoul.

Bridges, A.M., & Jones, D.M. (1996). Word-dose in the disruption of serial recall by irrelevant speech. *Quarterly Journal of Experimental Psychology, 49A*, 919-939.

Chalmers, D. J. (1990). Syntactic transformations on distributed representations, *Connection Science, 2*, 53-62.

Chrisman, L. (1991). Learning Recursive Distributed Representations for Holistic Computation. *Connection Science 3*, 345-366.

Fraisse, P. (1945). L'influence de la vitesse de presentation et de la place des éléments. La nature du present psychologique. *Annee psychoogique, 45*, 29-42.

Henson, R. N. A. (1996). Short-term memory for serial order. Unpublished doctoral thesis, University of Cambridge.

Martin, P. R. & Fernberger, S. W. (1929). Improvement in memory span. *American Journal of Psychology, 41*, 91-94.

Niklasson, L. F. (1999). Extended encoding/decoding of embedded structures using connectionist networks. In *Proceedings of the 9th International Conference on Artificial Neural Networks (ICANN99)*, IEE, 886-891.

Niklasson, L. F. & van Gelder, T. (1994). Can Connectionist Models Exhibit Non-Classical Structure

Sensitivity?, Proceedings of the Sixteenth Annual Conference of the Cognitive Society –94, Atalanta, Lawrence Erlbaum Associates, Hillsdale, NJ, 664-669.

Pollack, J. B. (1990). Recursive Distributed Representations. *Artificial Intelligence 46(1)*, pp 77-105.

Pollack, I., Johnson, L. B. & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology, 57*, 137-146.

Pickering, S. J., Gathercole, S. E., Hall, M., & Lloyd, S. A. (2001). Development of memory for pattern and path: Further evidence for the fractionation of visual and spatial short-term memory. *Quarterly Journal of Experimental Psychology, 54A*, 397-420 .

Wickelgren, W. A. (1964). Size of rehearsal group and short-term memory. *Journal of Experimental Psychology, 68*, 413-419.

Holographic Reduced Representations for Oscillator Recall: A Model of Phonological Production

Harlan D. Harris (hharris@uiuc.edu)
University of Illinois at Urbana-Champaign
Department of Computer Science, MC-258
Urbana, IL 61801

Abstract

This paper describes a new computational model of phonological production, Holographic Reduced Representations for Oscillator Recall, or HORROR. HORROR's architecture accounts for phonological speech error patterns by combining the hierarchical oscillating context signal of the OSCAR serial-order model (Vousden, Brown, and Harley 2000; Brown, Preece, and Hulme 2000) with a holographic associative memory (Plate 1995). The resulting model is novel in a number of ways. Most importantly, all of the noise needed to generate errors is intrinsic to the system, instead of being generated by an external process. The model features fully-distributed hierarchical phoneme representations and a single distributed associative memory. Using fewer parameters and a more parsimonious design than OSCAR, HORROR accounts for error type proportions, the syllable-position constraint, and other constraints seen in the human speech error data.

Introduction

The phonological production subsystem is the part of the language production apparatus that sequences the sounds in individual words and groups of words. Phonological production is the mapping from lexical units, morphemes and words, to sequences of phonological units, phonemes. This paper presents a new model of the phonological production system, a model that offers a new explanation for errors and serial order in speech.

Speech Error Effects

Numerous constraints and patterns have been observed in speech error patterns, including error type proportions (see Table 1), the syllable position constraint, the C-V category constraint, the distance constraint, the phonological similarity effect, and the phonotactic regularity effect (Fromkin 1971). Unless otherwise specified, the numbers in the descriptions below are from the (Vousden et al. 2000) analysis of the (Harley and MacAndrew 1995) error corpus.

A strong constraint on movement errors (the first three error types in Table 1) is the *syllable position constraint*, or SPC. 89.5% of movement errors retain their position in the syllable (onsets move to onsets, vowels to vowels, etc.).

Type	Rate	Example
anticipations	35.1%	**d**et the dog
perseverations	26.0%	pet the **p**og
exchanges	10.6%	**d**et the **p**og
non-contextual slips	17.3%	pet the **l**og
mixed errors	11.0%	**l**et the **p**og

Table 1: Error type proportions. Target utterance is "pet the dog." Mixed errors include any error not in the other categories.

An even stronger constraint is the *consonant-vowel category constraint*, or C-V constraint. Errors very rarely involve the replacement of a consonant by a vowel or vice versa. A superset of these errors, those that violate language-specific rules (the *phonotactic regularity effect*), occur in less than 1% of errors (Stemberger 1983).

The *distance constraint* is the observation that phonemes tend to move only short distances (one or two syllables) in movement errors.

When movement errors occur, they are more likely than chance to involve phonemes that share phonetic features. For example, "pig bull" for the intended "big pull" is a more likely exchange than "bill pug," since [p] and [b] are more similar than are [g] and [l]. This is the *phonological similarity effect*.

Language Sequencing Models

Phonological production models can be categorized by how they generate serial order. I follow Vousden et al. (2000) and use the terms *associative chaining model*, *frame-based model*, and *control signal model*.

Associative chaining models account for serial order by having each subsequent phoneme be triggered by a combination of the pattern of previous phonemes and a representation of the target utterance (Dell, Juliano, and Govindjee 1993). These models successfully account for phonotactic regularity effects and the C-V constraint, but they do not generate anticipations and exchanges well, nor do they account for SPC effects.

Frame-based models (Dell 1986; Roelofs 1997) use strict phonological frames to slot phonemes into pre-

specified positions, such as the onset, nucleus, and coda positions of a syllable. These models often use chains of sequencing nodes to activate the slots sequentially (Eikmeyer and Schade 1991). Although frame-based models are influential, sequencing nodes are often criticized as being poorly motivated.

To address this point, control signal models (Burgess and Hitch 1992; Hartley and Houghton 1996; Vousden et al. 2000) replace discrete syllable frames with continuous time-varying signals. Prior to production, different parts of the word are associated with different parts of the signal. Then, as the signal changes during production, the associated phonemes are output sequentially. Simple control signal models explain how phonemes could be produced in order, but don't account for SPC effects.

The OSCAR model (Vousden et al. 2000), described below, is a complex control signal model that accounts for SPC effects by using a multidimensional control signal with biological motivation. It contains an implicit frame in the way that the control signal is structured, but does not require explicit slots or sequencing nodes for production.

Building Blocks

The HORROR model combines elements of two previously existing models: the OSCillator-based Associative Recall (OSCAR) model of serial-order and phonological production (Brown et al. 2000; Vousden et al. 2000), and the Holographic Reduced Representations (HRR) model of hierarchical associative memory (Plate 1995). Prior to describing HORROR, I review its two ancestral models.

OSCAR

OSCAR works by associating item vectors (phoneme representations) and phonological context vectors (PCVs) in a Hebbian associative memory. The PCVs are inspired by oscillating signals in the brain, and have an important hierarchical self-similarity pattern, described below. As the PCVs are iteratively presented to the associative memory, the original item vectors are recalled and become available for production. The self-similarity pattern generated by the oscillators, when combined with noise, generates patterns of errors that previously required the use of syllable frames.

In OSCAR, there are 30 oscillators in two groups of 15. In the non-repeating group, the oscillators generate sinusoidal values at frequencies ranging from very slow to very fast. Initial phases and frequencies are generated with sufficient randomness that the non-repeating group's state does not repeat for many steps. In the repeating group, the initial phases of the oscillators are random, but the frequencies are identical. The state of this group repeats precisely every three time steps, representing the period of a three-segment CVC syllable.

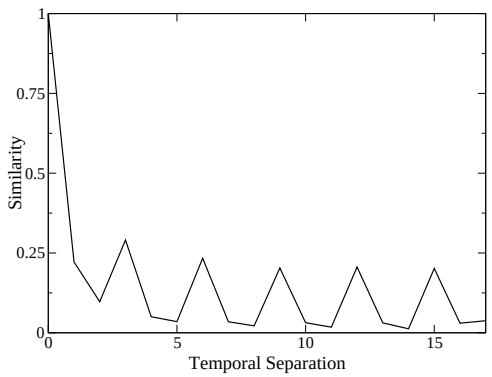


Figure 1: HORROR's PCV self-similarity function.

The PCV itself is generated by multiplying together selected oscillator signals to form a 32-element vector. Each element is a product of four oscillator signals, all of which are selected from the same group (repeating or non-repeating). The pattern of multiplications results in an automatically normalized PCV, allowing easy comparisons for similarity.

A key feature of OSCAR is that the PCV is self-similar in a hierarchical manner. Each state of the PCV is most similar to states that are multiples of three time-steps away, but nearby states are also somewhat similar (see Figure 1).

The process for producing a "word" (a randomly generated 18-segment sequence of six CVC syllables) is as follows. A PCV is initialized, and starts to change with time. At each time step, the PCV is associated with a phoneme feature vector in a Hebbian weight matrix. Each time step uses a *separate* weight matrix. This entire process is performed nine times (in parallel), to create a total of 81 weight matrices, nine replications of nine time steps. To produce the sequence, the PCV is re-instated to its initial state, then sequentially re-produces each step's state. The PCV is usually associated with the correct weight matrices to generate an approximation of the phoneme feature vector. In addition, a probabilistic process is used to generate errors. 70% of the time, segments which are associated with PCVs that are similar to the current PCV are combined with the output from the correct weight matrix. The result is an output vector, a potentially noisy approximation to the correct phoneme. Also, a post-output suppression mechanism is used to reduce excessive perseveration and facilitate exchange errors. The generated output vectors for each of the nine replications are compared to an item memory containing each phoneme, such that each phoneme is activated to an extent proportional to the similarity with the nine output vectors. The most active phoneme is then produced in a winner-take-all process.

OSCAR's Pros and Cons In many ways, OSCAR is important work in the literature of phonological production and speech error modeling, but it has significant problems that may limit its applicability. Its contributions include making good use of an independently-motivated context signal to create serial order, accounting for SPC effects without position-specific phonemes, and using an implicit rather than explicit syllable frame. Overall, it accounts for various error patterns better than do chaining models.

However, several limitations lead me to question the extent of the model's successes. Most importantly, the noise-addition procedure is unprincipled. As well, the artificial words did not include repeated phonemes, the associations between the context and phonemes are stored separately, and there are a concerning number of parameters.

Consider the noise-addition procedure. Cognitive models should use reasonable sources of noise to generate error phenomena. Many models add Gaussian noise, while others use intrinsic noise from distributed representations and imprecise network computation. Although OSCAR uses well-motivated oscillator signals to provide serial-order effects, its noise-generation procedures are much more weakly motivated. As described above and in Appendix C of Vousden et al. (2000), phonemes associated with states of the PCV that are selected by their similarity to the correct PCV are recalled in parallel and used to corrupt the winner-take-all process.

That this procedure generates impressive error results is not surprising. The noise in OSCAR is generated only by interference from particular phonemes in the current sequence, not by any sort of random numerical noise or other natural interference. OSCAR claims to explain why most errors are movement errors – in their model, it's because the generated noise *is* movement noise.

A related concern with OSCAR is that the associations between the PCV and phoneme vectors are stored separately. Although it is reasonable to use Hebbian learning to associate a PCV signal with phoneme representations, it is difficult to explain why each segment need be stored in entirely separate sets of weights. A more parsimonious solution would use a single set of weights and would treat the resulting noise as an asset, not a weakness.

HORROR adopts the oscillating PCV system from OSCAR, but replaces the movement-based noise-creation system with the noise inherent in an associative memory system with overlaid weights. It also uses a more parsimonious unified memory system, allows repeated phonemes within a sequence, and requires fewer free parameters[1].

[1]In addition to the five listed in Table 7 of Vousden et al. (2000), there are these four: the ratio of correct-to-incorrect activation, 0.6; the number of redundant associations, 9; the similarity threshold for allowing a phoneme to be added as noise, 0.5; and the similarity exponentiation factor in the item memory, 3.4.

	$\mathbf{T_1} = \mathbf{a} * \mathbf{b}$
$* : \mathbf{I} \times \mathbf{I} \Rightarrow \mathbf{T}$	$\mathbf{a} \# \mathbf{T_1} \rightarrow \mathbf{b} + \text{noise}$
$\# : \mathbf{I} \times \mathbf{T} \Rightarrow \mathbf{I}$	$\mathbf{T_2} = \mathbf{a} * \mathbf{b} + \mathbf{c} * \mathbf{d} + \mathbf{e} * \mathbf{f}$
$+ : \mathbf{T} \times \mathbf{T} \Rightarrow \mathbf{T}$	$\mathbf{d} \# \mathbf{T_2} \rightarrow \mathbf{c} + \text{noise}$
	$\mathbf{T_3} = \mathbf{g} * \mathbf{T_1} + \mathbf{h}$
	$\mathbf{g} \# \mathbf{T_3} \rightarrow \mathbf{T_1} + \text{noise}$

Figure 2: Holographic Associative Memory. $\mathbf{a} - \mathbf{h}$ are item vectors; $\mathbf{T_i}$ are memory vectors. $*, \#$, and $+$ symbolize circular convolution (encoding), correlation (decoding), and addition (composition).

Distributed Associative Memories

For several decades, mathematical psychologists have looked at distributed representations for models of memory (Murdock 1982; Eich 1982), and have accounted for many recognition and recall effects. Compared to localist connectionist models, where representations consist of features and microfeatures, distributed representations use long quasi-random vectors. These vectors are generated and manipulated such that similarity between two representations is defined by the dot product or cosine. Distributed representations can be combined in various ways. Two symbols may be associated by operations such as convolution or the outer product, resulting in another large vector. Retrieval from memory vectors is performed by inverting the association operation, correlation. Distributed memories can store a number of associations at once, simply by adding the vectors together. As vectors are overlaid, the amount of noise increases. This intrinsic noise is part of the model, and resulting simulations can account for list-length and item-similarity effects.

A limitation in much of the work on distributed memories is that the operations that generate associations greatly expand the size of the vector, with the result that hierarchies of associations are impractical. HORROR utilizes one of several approaches that overcome this problem, the Holographic Reduced Representations (HRRs) of Plate (1995).

With HRRs, the representations and associations are always fixed-length vectors. A circular version of convolution is used to associate vectors. The resulting memory vectors are the same length as the input vectors, at the cost of increased noise. The greatest benefit is that hierarchies of associations can be easily generated and stored. See Figure 2 for simple examples and notation. An auto-associative item memory (a Hopfield network or a nearest-neighbor search through a list) is necessary to identify the result of each correlation.

HORROR

HOlographic Reduced Representations for Oscillator Recall (HORROR) is a model of serial-order processing that combines the self-similar context vectors of OSCAR with the hierarchical representations and memory of HRRs. The result is a fully-distributed phonological production model that accounts for errors in the serial-order part of the system by using the intrinsic noise from the associative-memory part of the system.

OSCAR and HRRs fundamentally both represent similarity by distance between vector representations. In OSCAR, the extent to which pairs of context vectors which are near in time are also near in space determines retrieval accuracy and error patterns. With HRRs, capacity and noise levels are determined by the extent to which composed vectors are near (not orthogonal) to each other. In OSCAR these similarity metrics can be complex and hierarchical, determined by the oscillator patterns, and in HRRs, the similarity metrics can also be hierarchical, by the process of overlaying associations. In both models, item memories are used to clean up and to select a single item.

HORROR is a new model based on the general framework of OSCAR. It takes a variation of the PCV from OSCAR, and combines it with an HRR associative memory, replacing the simple associative memory used by OSCAR. In addition, the feature-vector phonological representations used in OSCAR are replaced with fully-distributed hierarchical representations in HORROR. A critical aspect of HORROR is that all of the phoneme-context pairs that make up a sequence are stored together in a single large vector, rather than in OSCAR's many separate weight matrices. The noise in this memory vector, combined with representational similarity and the PCV structure, provide sufficient opportunities for appropriately distributed error patterns to arise.

Experiments

A major goal of this work is to account for the same human speech error data as does OSCAR, using a simpler structure, more parsimonious procedures, and fewer parameters.

As with OSCAR, PCVs are generated sequentially and convolved with phoneme representations to form memory traces. Unlike OSCAR, these traces are summed to form a single vector representing the entire sequence. To produce the sequence, the vector is correlated with the PCVs in order, resulting in noisy versions of the phonemes. The phonemes are cleaned up in an item memory, and the results are analyzed for various types of errors.

The oscillators used to generate the PCV were the same as used by OSCAR. HORROR additionally includes a parameter, *nrep*, that specifies the proportion of repeating versus non-repeating oscillators. The procedure of generating the PCV from the oscillators in HORROR is very similar to the procedure used in OSCAR, but since HORROR's PCV is very wide (2048 elements), the process was repeated with different random initial phases and frequencies in order to fill up the vector, which was then normalized. See Table 2 for the list of PCV and other parameters used in the experiments described below.

Param.	Value	Description
nrep	17	# of non-repeating oscillators
vw	2048	Representation vector width
cc	3	Repeating oscillator inv. freq.
D	4	Speech-rate (larger = slower)
Inhib	.121	Post-activation inhibition level
InDec	.5	Inhib. decay (lower = faster)
ds	3	Phoneme dis-similarity factor

Table 2: Free parameters in HORROR

Vousden et al. (2000) use an articulatory-feature representation of phonemes. Each phoneme is 17 elements long, with binary features representing place and manner of articulation, nasality, voicing, and vowel position and tenseness. We converted these localist features into distributed features for the fully-distributed representations used in HORROR.

Phonological representations were built in a fully-distributed manner by generating random Gaussian vectors (of width *vw*) for each feature, then summing the appropriate features together and normalizing. Each vector thus has an intrinsic similarity metric, defined by the number of shared features. In order to partially "drown out" the similarity between otherwise very-similar phonemes, additional random vectors (*ds*) were added to each phoneme vector.

Decoding consists of sequentially correlating each time-step of the PCV with the single stored memory vector. The result is a series of approximations to the target phonemes, corrupted by the noise intrinsic to a holographic memory. Each recalled vector is compared to an item memory containing possible phonemes. The phoneme that is most similar to the recalled vector is then produced.

The item memory has three features that help it best account for the error patterns. First, each item in the item memory has a persistent activation level, a. Activation is added to similarity to determine which phoneme is selected. At each step, each item's a is increased by the item's distance from the recalled vector, weighted by the $Inhib$ parameter. Second, after a phoneme is selected, it is suppressed by setting a to be the negation of $Inhib$. Post-output suppression is a common feature of this type of model (Vousden et al. 2000; Dell 1986). Finally, at every time step, activation decays toward zero according to the decay constant $InDec$.

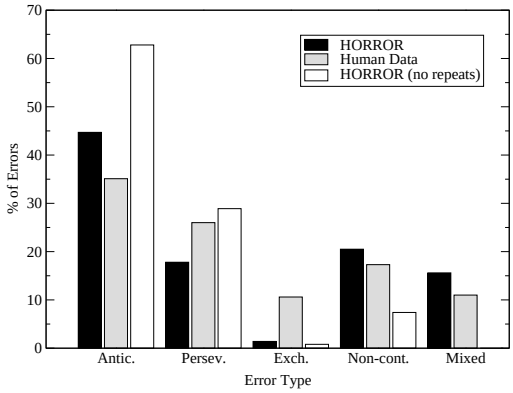


Figure 3: HORROR's error type proportions.

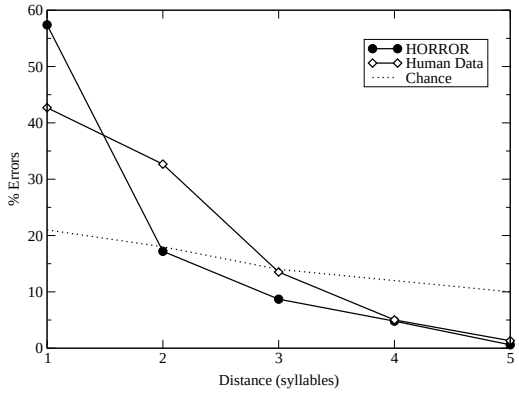


Figure 4: Distance gradients of the anticipation errors produced by HORROR, compared with the human data and chance baseline of Vousden et al. (2000). Adjacent syllables have a distance of 1. Same-syllable errors (separation 0) are not shown.

Experimental Results

2000 6-syllable "words" were generated, associated, and output. Errors were determined by an automatic categorization process. Error type proportions, SPC violations, distance constraint statistics, and phonetic similarity constraint statistics were counted.

Error proportions 1538 errors occurred during production of 36,000 segments, resulting in an overall error rate of 4.3%. Figure 3 shows the proportions of error types. The results compare fairly well with the human data reported in Vousden et al. (2000). Exchanges, however, were under-represented in the model, raising the question of whether HORROR's exchanges are true exchanges or merely the joint event of independent anticipations and perseverations. Other speech error models (Dell et al. 1993; Roelofs 1997) are unable to produce true exchanges, and are therefore seen as incomplete.

To address this, I calculated the expected number of exchanges, assuming that they are coincidental. This number, 0.33 per 2000 sequences, was more than sixty times smaller than the number of exchanges actually observed (22), demonstrating a true tendency for exchanges. Exchanges in HORROR occur because post-activation inhibition helps to prevent an erroneously anticipated phoneme from then appearing in its correct location. Instead, the earlier, replaced phoneme may be triggered via the PCV, turning an anticipation into an exchange.

Distance constraint The model's movement gradients parallel the distance constraints seen in human data, with disproportionately small separations. Exchange errors shifted least, an average 0.95 syllables, followed by anticipations, averaging 1.4 syllables, and perseverations, averaging 2.9 syllables. Figure 4 shows a comparison on anticipations between HORROR and human data. The shorter movements made by exchanges, compared to other error types, has been observed in human error data (Nooteboom 1973).

Error type	n	Mean shared features
Movements	895	2.4
Exchanges	22	3.1
Non-contextual	306	3.1
Chance		1.9

Table 3: Average similarity for consonant errors. Movements are anticipations and perseverations.

Phonetic similarity constraint Vousden et al. (2000) concentrate their analysis of the phonetic similarity constraint on consonant exchanges. HORROR produced only 22 consonant exchanges, 20 of which shared 75% or more of their phonetic features. Table 3 compares the categories of consonant errors to chance. Chance was determined by randomly selecting 1000 pairs of consonants, and counting the number of shared phonemes for each pair. The phonetic similarity constraint is clearly present in these results. Note that exchange errors were significantly more similar than were other movement errors. This is true for human exchanges, and also lends further support to the observed exchanges being real.

Syllable-position constraint 29.0% of the model's errors violated the SPC, compared to 10.5% of errors in human data (Vousden et al. 2000). To confirm that this number still reflects a constraint, and is not just the chance rate of violations, it's necessary to look at the probabilities of errors being in each syllable position. In this set of data, 50.7% of errors were in the onset, 8.6% in the vowel, and 40.8% in the coda. To

calculate the expected rate of SPC violations, assume that the consonant-vowel constraint is never violated, and that consonant errors have a 50% chance of movement from onsets and codas. Therefore, the expected SPC violation rate is $1 - (.086 + .507 * .5 + .408 * .5) = 45.7\%$. Although the SPC is violated more often by the model than it is in human data, it is still a real effect.

Consonant-vowel constraint Only 2.3% of the errors violated the C-V constraint, showing that the model is generally respecting the consonant-vowel categorical distinction seen in natural errors.

Repeated phonemes In order to investigate the role of repeated phonemes in the model, the same experiment was re-run with repeated phonemes disabled. Since repeated items are known to strongly affect performance in distributed memories, it was expected that the effects on HORROR would be significant as well. The error rate without repeated items was reduced to 1.3%, and the proportion of non-contextual errors was greatly reduced (see Figure 3). HORROR is more error prone when repetitions occur, as in human data (Dell 1986). Repeated phonemes appear to be an important trigger for speech errors, including non-contextual errors.

Discussion

The HORROR model combines the best features of OSCAR, a serial-order phonological model with a hierarchical context signal, and HRRs, a holographic associative memory using hierarchical representations. Its aim is to account for speech error patterns using more parsimonious mechanisms than previous related models.

HORROR succeeds in a number of ways. It allows repeated phonemes in the sequences, it combines associative memory traces into a single distributed association vector, and its error mechanism relies entirely on the intrinsic noise from the associative memory with no generated noise at all. It uses fewer parameters than does OSCAR, and accounts for a number of error patterns in human data. Specifically, the model's error type proportions, distance constraint, phonological similarity constraint, and C-V category constraint results were largely similar to human data. The SPC results were real, if modeled less accurately. HORROR accounts for these major speech error patterns by using fully-distributed hierarchical representations, a single intrinsically-noisy associative memory, and an oscillating phonological context signal.

Acknowledgments

I appreciate the advice and contributions of Gary Dell, Janet Vousden, and Franklin Chang. This work was supported by NSF grant SBR 98-73450 and NIH grant DC-00191.

References

Brown, G. D. A., T. Preece, and C. Hulme (2000). Oscillator-based memory for serial order. *Psychological Review 107*(1), 127–183.

Burgess, N. and G. J. Hitch (1992). Toward a network model of the articulatory loop. *Journal of Memory and Language 31*, 429–460.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review 93*(3), 283–321.

Dell, G. S., C. Juliano, and A. Govindjee (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science 17*, 149–195.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review 89*(6), 627–661.

Eikmeyer, J.-J. and U. Schade (1991). Sequentialization in connectionist language-production models. *Cognitive Systems 3*(2), 128–138.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language 47*(1), 27–52.

Harley, T. A. and S. B. G. MacAndrew (1995). Interactive models of lexicalization: Some constraints from speech error, picture naming, and neuropsychological data. In D. Bairaktaris, J. Bullinaria, and D. Cairns (Eds.), *Connectionist models of memory and language*. London: UCL Press.

Hartley, T. and G. Houghton (1996). A linguistically restrained model of short-term memory for non-words. *Journal of Memory and Language 35*, 1–31.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative recall. *Psychological Review 89*(6), 609–626.

Nooteboom, S. (1973). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech Errors as Linguistic Evidence*. The Hague: Mouton.

Plate, T. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks 6*(3), 623–641.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition 64*, 249–284.

Stemberger, J. P. (1983). *Speech errors and theoretical phonology: A review*. Bloomington: Indiana University Linguistics Club.

Vousden, J. I., G. D. A. Brown, and T. A. Harley (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology 41*, 101–175.

Similarity and Difference Judgments Under Perceptual and Non-Perceptual Conditions

Uri Hasson (uhasson@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544, USA

Vladimir Sloutsky (Sloutsky.1@osu.edu)
Center for Cognitive Science & School of Teaching and Learning, The Ohio State University
21 Page Hall, 1810 College Road, Columbus, OH 43210, USA

Abstract

It has recently been suggested that knowledge is represented in the form of perceptual symbol systems (Barsalou, 1999). According to this view, perceptual states may be used to support higher cognitive processes without being transduced into a representational language. Since the ability to recognize difference and similarity is fundamental for cognition, we examined to what extent it might be based on perceptual information. In three experiments, participants made judgments of similarity and difference for simple items under three presentation conditions: Words Only, Words and Pictures, and Pictures Only. Reaction times for judgments in the Words-Only condition were consistently slower than in the other presentation conditions. However, judgments of perceived similarity and perceived difference did not markedly differ between presentation conditions. The results suggest that participants recruited perceptual information when evaluating similarity in the Word-Only condition. Additionally, presentation condition had an effect on the relation between the similarity and difference scales: for a given degree of similarity, more extreme difference judgments were found under those conditions where words were displayed. We offer an explanation for this effect, and present a further research program.

Introduction

Similarity, or psychological resemblance of entities, is a fundamental aspect of cognition. Similarity plays a critical role in perception, memory, learning and transfer, categorization, analogical reasoning, problem solving, and language comprehension. It has also been suggested that recognizing differences between entities is fundamental to cognition; for instance, in discriminating category members from non-members. An extensive body of research then has outlined the processes for which similarity and difference are important. However, much less is known about that nature of the information recruited for evaluating similarity and difference themselves.

Some theories of similarity hypothesize a process whose inputs are representations in the form of feature lists with most of the properties represented as unitary attributes (Nosofsky, 1986; Tversky, 1977). According to others, conceptual knowledge plays such an important role in similarity, that similarity is taken to be akin to analogy; operating on an interconnected system of relations and their arguments (Gentner & Markman, 1997). The crux of this view is that processes of similarity, as well as difference (Markman, 1996), operate by aligning the structures of the compared entities. The structures themselves are described as a system of relational predicates and their attributes. The research program of structural alignment (see Gentner & Markman, 1997) convincingly demonstrated that the use of structural knowledge is an integral part of making similarity judgments.

It is an open question, however, whether the *only* kinds of information used for these judgments are in the form of such a-modal representations as relations or feature lists. A recent proposal (Barsalou, 1999), suggests that perceptual states - modal and analog forms of representation - are *also* used to support higher cognitive processes. Perceptual states are taken to maintain a part of the perceptual nature of their referents, without being transduced into a representational language. If this is the case, judgments of similarity and/or difference might make use of perceptual information, in addition to other forms of knowledge.

We examined this issue by evaluating the effects of various types of presentation conditions on similarity and difference judgments, which were made for a variety of natural stimuli. We presented participants with pairs of items for similarity and difference judgments under three presentation conditions which were manipulated between groups: pairs were presented as Words-Only (WO), as words accompanied by pictures (WP) or as pictures only (PO). The instructions given to all three groups were the same, and did not make any reference to speed of response. Most critically, the instructions given to participants in the Word-Only condition did *not* mention envisioning the objects depicted by the words. If participants in this

group took longer to reach their decision, but ultimately made judgments resembling those made in the picture conditions, this would suggest that perceptual information was recruited when making these judgments.

The reported research has two major goals: (1) to estimate effects of perceptual and non-perceptual aspects of the stimuli on making judgments of similarity and difference; and (2) to examine whether perceptual and non-perceptual aspects of the stimuli affect the relation between the similarity and difference scales.

Calibration Experiment

The purpose of the calibration study was to choose pictorial material that would be easily and unambiguously recognized as typical examples of the intended items. Twenty-seven undergraduate students from the Ohio State University participated in the study. The photographs chosen corresponded to 100 items, taken from 10 categories of Battig and Montague's (1969) category norms with 10 items being selected from each category. The ten categories belonged to two ontological domains, living things and artifacts, with five categories in each domain. In addition to the experimental items, 25 items that appeared to be bad examples of their types were also added as negative anchors. Participants were presented with photographs of objects followed by words denoting these objects, and decided whether the photographs were good examples of the entities depicted by the words ("something that immediately reminds you of that thing"). For example, a picture of a car was displayed, followed by a blank screen and then by the word *CAR*. Participants pressed 1 if the photo was indeed a good example and 0 if it was not.

The proportion rating for the experimental items was 0.85. We included only those photographs that were found to be good depictions by more than 67% of the participants. This procedure resulted in the removal of 10 items.

Experiment 1a: Judging Similarity and Difference of Objects Denoted by Words

Method

Participants There were 29 participants in the similarity-judgment condition and 25 participants in the difference-judgment condition. The participants in both conditions, and in all following experiments, were undergraduate students from the Ohio State University who participated to fulfill a psychology course requirement or in return for payment.

Design and Materials The experiment had a mixed design with Judgment type (similarity judgment or difference judgment) as a between-subjects variable and Pair-type (same-superordinate, within-ontological domain or across-ontological domains) as a within-subjects variable. The 90 items chosen from the calibration test were used. Items were paired to construct 240 pairs of items such that: (a) 80 pairs were of items from the same superordinate category, (b) 80 pairs were of items from different superordinate categories but from the same ontological domain (i.e., both were either artifacts or living things) and (c) 80 pairs were of items belonging to different ontological domains. We decided to use such pairs since there was ground to suppose that they differ considerably in their perceived similarity (see Markman & Wisniewski, 1997).

Procedure The participants worked alone in an experimental room. Participants making similarity judgments were told that their task was to decide how similar these items were. They were instructed to press 3 if items were very similar, 2 if they were not so similar, and 1 if they were not similar at all. Participants making difference judgments were told that their task was to decide how different these items were. They were instructed to press 3 if items were very different, 2 if they were not so different, and 1 if they were not different at all. The experiment was run in three blocks, with a 1-minute break between blocks. The experiment began with a few practice pairs, which were followed by 14 hidden practice pairs whose purpose was to bring participants up to speed. Practice items were not analyzed.

Results and Discussion

Reaction time data and ratings for similarity and difference judgments are presented in Table 1. Note that ANOVA results will be presented in the overall-analysis section, whereas the results of t-tests are presented in Table 1.

Table 1. Reaction times and ratings for similarity and difference judgments in the Word-Only condition

	Reaction Time		Ratings	
Pair type	Sim	Dif	Sim [x]	Dif [y]
Across-domain	1232_a	1245_a	1.04_a	2.99_a
Within-domain	1315_b	1317_b	1.20_b	2.90_b
Within-superordinate	1400_c	1511_c	2.32_c	2.06_c
Average	1315	1358	1.52	2.65

Note. Numbers with different subscripts in a given column differ at $p < .0001$.

[x] High numbers reflect greater similarity

[y] High numbers reflect greater difference

As expected, similarity and difference ratings for the different pair types were different, with Within-

Superordinate pairs (e.g., dog – cat) being more similar than Within-Domain (e.g., shark – horse) pairs, and the latter being more similar than Across-Domain (e.g., cow – spoon) pairs. The difference ratings showed the same pattern. More importantly, reaction times in both the similarity and difference judgment groups were fastest for Across-Domain pairs, slower for Within-Domain pairs and slowest for Within-Superordinate pairs. As will be shown, this pattern of results replicated in the Word-Picture and Picture-Only presentation conditions as well.

Experiment 1b: Judging Similarity and Difference of Objects Depicted by Words and Pictures

Method

Participants There were 18 participants in the similarity-judgment condition and 18 participants in the difference-judgment condition.

Design, Materials and Procedure Design, Materials, and Procedure were identical to those in Experiment 1a, except that items were denoted by words and by pictures, such that the words were printed above the pictures.

Results and Discussion

Reaction time data and ratings for similarity and difference judgments are given in Table 2. The overall pattern of reaction time and rating results was similar to that in experiment 1a.

Table 2. Reaction times and ratings for similarity and difference judgments in the Picture-Word condition

	Reaction Time		Ratings	
Pair type	Sim	Dif	Sim [x]	Dif [y]
Across-domain	1002_a	1180_a	1.04_a	2.96_a
Within-domain	1123_b	1177_b	1.23_b	2.84_b
Within-superordinate	1270_c	1376_c	2.37_c	1.82_c
Average	1131	1211	1.55	2.54

Note. Numbers with different subscripts in a given column differ at $p < .0001$.

[x] High numbers reflect greater similarity

[y] High numbers reflect greater difference

Initial inspection demonstrates that reaction times in this presentation mode were faster than in the Word-Only presentation condition. Though several theoretical explanations for this may exist, it was important to examine one in particular; namely, that participants were ignoring the words and basing their decisions solely on the pictures.

To examine whether participants were reading the words, we modified the experiment slightly to include 12 inconspicuous spelling mistakes. If participants were paying attention to the words, reaction times for these changed items should be higher than reaction times to the same items in the original experiment. Thirteen participants participated in the modified version of the similarity-judgment condition. The average reaction time for the modified items (1453 ms) was significantly higher than for the original, items (1170 ms.; one-tailed t-test, $t_{22} = 2.37$, $p < .01$). In addition, we asked participants whether they had noticed anything during the study. All but two noticed a few spelling mistakes. The results indicate that participants were indeed reading the words presented above the pictures, and that the reduced reaction times could not be attributed to such neglect.

Experiment 1c: Judging Similarity and Difference of Objects Depicted by Pictures

Method

Participants There were 25 participants in the similarity-judgment condition and 26 participants in the difference-judgment condition.

Design, Materials and Procedure Design, Materials, and Procedure were identical to those in Experiment 1a and 1b, except that items were now depicted only by pictures. Participants were told that they would be presented with pictures referring to objects in the world, and that their task is to determine how similar these objects are.

Results and Discussion

Reaction time data and ratings for similarity and difference judgments are given in Table 3. The overall pattern of results was as in experiments 1a and 1b. The fact that the similarity ratings increased and difference ratings decreased the more conceptually related the objects were, testifies to the fact that participants used conceptual knowledge in their judgments of pictorial material, and did not rely exclusively on perceptual information. Reaction time data and ratings of similarity and difference are given in Table 3.

Table 3. Reaction times and ratings for similarity and difference judgments in the Picture-Only condition

	Reaction Time		Ratings	
Pair type	Sim	Dif	Sim [x]	Dif [y]
Across-domain	1067_a	1159_a	1.03_a	2.92_a
Within-domain	1154_b	1171_b	1.21_b	2.74_b
Within-superordinate	1334_c	1316_c	2.22_c	1.85_c
Average	1185	1215	1.49	2.50

Note. Numbers with different subscripts in a given column differ at $p < .0001$.

[x] High numbers reflect greater similarity

[y] High numbers reflect greater difference

Combined Analysis of experiments 1a – 1c

Analysis of Response Times for Difference and Similarity Judgments Under Three Presentation Conditions.

The data from all experiments were combined to assess the effects of the presentation-mode on reaction times in the similarity- and difference-judgment tasks. Average response times were calculated and entered into a 3 (Presentation) X 2 (Judgment) X 3 (Pair-Type) mixed ANOVA, with Presentation (Word-only, Word-Picture and Picture-only) and Judgment (Similarity and Difference) as between subjects variables, and Pair-Type (Across-domain, Within-domain, Within-superordinate) as a within subjects variables.

As expected, the main effect of Pair-Type was significant, $F(2, 250) = 251.5$, $p < .0001$. More important, the main effect of Presentation was significant, $F(2,125) = 10.91$, $p < .0001$. Reaction times in the Word-Picture condition (1138 ms) were faster than in the Picture-Only condition (1183 ms) and the latter were faster than reaction times in the Word-Only condition (1337 ms). Scheffe's post hoc analysis revealed that reaction times in the WO condition were significantly slower than reaction times in the PO and WP presentation conditions ($p < .001$). Judgments in the WO condition were slower than in the other conditions even though the WP condition presented participants with more information, and the PO condition amounted to a naming task. There are several explanations for this effect: it might be that under the different presentation conditions, participants made differential use of perceptual and conceptual information in their judgments; relying more heavily on pictorial data whenever it was available, thus being faster in those conditions that contained pictures. If this is the case, then judgments in the PO and WP conditions should be more similar to each other than to the WO condition. The second possibility, which is consistent with the perceptual-symbols hypothesis, is that participants under the three presentation conditions eventually constructed the same representation, but that this process took longer in the Word-Only condition. If so, judgments in all three modes are expected to be quite similar. An analysis of the similarity and difference judgments was conducted to test these possibilities.

Analysis of Similarity and Difference Judgments Under Three Presentation Conditions.

We now examined the effects of presentation mode on (a) judgments of similarity and difference and (b) the relation between the similarity and difference scales. We separately analyzed similarity and difference judgments given under the three presentation conditions. Due to the ordinal nature of the response scale, we used the variance of response proportions as the measure of test in the following ANOVAs. However, since the data are more easily encapsulated in the form of averages, we present them in Table 4.

Table 4: Mean similarity and difference ratings under three presentation conditions.

Condition	Mean Similarity	Mean Difference
Picture Only	1.489	2.507
Word Picture	1.548	2.543
Word Only	1.522	2.654

We first analyzed responses given in the *similarity-judgment* groups. The proportions of 1, 2 and 3 responses were analyzed in 3 separate one-way ANOVAs, with Presentation as a grouping factor and response proportion of each response as the dependent variable. No effect was found in any of these analyses ($Fs < 1$). There were practically no differences in the response proportions in the three presentation conditions. Proportions of the three responses varied minimally between presentation conditions; in the range of 1.5%.

A similar analysis was performed on the response proportions in the *difference-judgment* groups. The proportions of 1 and 2 responses did not vary significantly between presentation conditions. However, an ANOVA of '3' responses was significant, $F(2,62) = 5.08$, $p < 0.01$. In the WO condition, there were more '3' responses (74%) than in the other WP and PO presentation conditions (69%, 68% respectively).

The similarity-rating data are congruent with the hypothesis that participants in all presentation conditions performed the similarity judgment on the same representation. Note however that while judgments across all conditions were similar, participants in the WO condition, who were not presented with the perceptual component in the stimuli, were slower to make their judgments.

In addition to this converging measure, the data that are perhaps most suggestive of the use of perceptual information in the WO condition is the fact that participants in the WO condition were also slowest in judgments of Across-domain items (e.g., Pan-Dog). One does not *need* to envisage a Pan and a Dog in order to determine that they are not similar at all, or very different. Such a decision could easily be made on the basis of categorical knowledge alone. However, judgments of similarity and difference for such pairs in the WO condition were slower than such judgments in

the WP and PO conditions (1232, 1002, 1067 and 1245,1080,1059 ms. for similarity and difference judgments respectively). The difference-judgment data are also congruent with the possibility that participants evoked perceptual data. As reported, difference judgments to Across-domain items in the WO condition were slower than in the other presentation conditions, suggesting that perceptual information was used here as well. However, the slight tendency to find more items 'very different' in the WO conditions may suggest that another mechanism was also at work here. We discuss the issue subsequently.

Analysis of Scale Equivalence Under Three Presentation Conditions

We also computed a measure relating the similarity and difference ratings under the three presentation conditions. To recap, in this analysis we are interested in the relation between an items' perceived similarity and its perceived difference under the three presentation modes. Since it is only sensible to talk of a degree of similarity in cases where similarity was found, we included in this analysis only items whose average similarity was greater than 1.

To compute this measure, for each item whose average similarity score was greater than 1, we compared its location on the similarity scale to its location on the difference scale. Let S_i and D_i denote, respectively, the distance of an item from the "least similar" end of the similarity scale and the "most different" end of the difference scale, in units of standard deviation. In the case of equivalence of the similarity and different scales, the items' difference between S_i and D_i should be equal to 0 (i.e., an item that was judged "most similar" should be also judged "the least different").

Delta, a measure of deviation from scale equivalence, is the average of the differences between these two parameters when computed across all items. The total deviation from equivalence between the scales then is $[\sum (S_i - D_i)]/N$. If scales are equivalent, Delta is equal to 0. Delta is greater than 0 when the judgment of similarity is more conservative than the judgment of difference (e.g.,, if a pair is judged as "somewhat similar" and also as "very different"). Delta is less than 0 when the judgment of difference is more conservative than the judgment of similarity (e.g., if a pair is judged as "somewhat similar" and also as "least different").

Delta measures for all items, under the three presentation conditions were subjected to a one-way ANOVA with Presentation as a between-groups factor. The main effect of Presentation was significant, $F(2,511) = 40.3$, $p < .001$. Scheffe' s post hoc comparisons revealed that all Delta measures differed significantly both from each other and from zero, p's < .05. Delta measures, the number of pairs on which Delta was greater than zero and the number of pairs on which Delta was lower than zero for each presentation condition are given in Table 5.

Table 5: Delta ratings and proportions for three presentation conditions.

Condition	Delta	Pairs where Delta > 0	Pairs where Delta < 0
Picture-Only	-0.08_a	57	108
Word-Picture	0.09_b	109	54
Word-Only	0.16_c	146	38

Note. Numbers with different subscripts in a given column differ at $p < .05$.

In the Picture-Only condition, items tended to be located 'farther' from the "very different" end of the difference scale than they were from the "not similar" end of the similarity scale. However, for the Word-Picture condition, and especially the Word-Only condition, items tended to be 'closer' to the "very different" end than they were to the "not similar" end. In short, negative Deltas in the Picture-Only conditions stem from more conservative ratings of difference compared to the similarity ratings given.

A possible explanation is that the presence of words resulted in more weight being given to conceptual knowledge in judgments. It has been suggested that in certain conditions, difference judgments are based on a comparison that involves aligning the structure of the items in the pair (Markman, 1996). Since conceptual knowledge affords a basis for structural alignment, it might be that the presentation of words resulted in enhanced attention to these differences. For instance, a pair such as 'dog – cat' might be judged highly similar when presented under all presentation conditions. However, when greater weight is given to conceptual knowledge, one might remember that dogs bark and cats meow, that dogs can become attached to people and cats are territorial, and that dogs bite and cats scratch. Minding these differences could lead to more extreme difference ratings.

General Discussion

We set out to examine the effects of various presentation modes on perception of similarity and difference in order to establish what information is used in such judgments, and how the presentation modes affect the similarity and difference scales themselves. We used Reaction times, and similarity and difference ratings to address the first issue, and used Delta, a measure of asymmetry between the scales to address the second.

In short, two major findings stem from the three reported experiments: (1) When pictures were not

present, reaction times were slower than in the other presentation conditions, but the judgments were comparable with them and (2) Deltas were positive whenever words were present, and were negative whenever words were absent. These data may be indicative of several important regularities.

First, people may rely on both perceptual and conceptual input when making similarity/difference judgments. In particular, their responses are faster when words are accompanied by pictures than when words are presented alone. It seems reasonable to hypothesize that they try to envision objects when those are depicted by words alone. Particularly suggestive of this, is the fact that compared to the Picture-Only and Picture-Word conditions, participants in the Word-Only condition were also slower in responding to Across-domain pairs. Other research on the role of perceptual information in higher-level cognitive tasks supports this possibility. Recent studies have demonstrated that accessing perceptual knowledge is an integral part of such tasks as property generation and property verification (Solomon & Barsalou, 2000; Wu & Barsalou, 2001). Though the data support the hypothesis that people did use perceptual information, the issue can be conclusively resolved by obtaining activation measures from brain areas implicated in imagery during the performance of this task.

Second, the Delta measure refers to what may be different processing considerations under the Word-Only and Word-Picture conditions, on the one hand, and the Picture-Only condition on the other. Particularly, in the former two conditions, Deltas were largely positive, whereas in the latter condition they were largely negative. Recall that the more that difference is underestimated per a given similarity rating, the larger the Delta. Therefore, negative Deltas in the Picture-Only condition stemmed from more conservative ratings of difference compared to respective similarity ratings. These data may point to the effect of knowledge on the perception of difference. They also indicate that regardless of the pair type, judgments of similarity and difference are affected by the modality of input.

Overall faster response times for different pairs than for similar pairs are indicative of the fact that judgments "not similar" or "very different" are made whenever no sufficient similarity is found. At the same time, computation of the degree of similarity is a more lengthy process. The results reported here seem to establish a boundary condition for situations where alignment is used in similarity and difference judgments. Participants were faster to decide that objects are not similar than to decide that they are; suggesting that judgment of difference was not based on finding specific differences. We are currently in the process of setting up research for examining brain activity under the conditions reported here to evaluate whether brain areas involved in imagery are recruited during similarity judgments to words only.

References

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(4), 577-660.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology, 80(3/2), 1-46.

Gentner, D., & Markman, A. B. (1997). Similarity mapping in analogy and similarity. American Psychologist, 52(1), 45-56.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. Cognition, 65, 263-297.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. Cognitive Psychology, 23, 222-262.

Markman, A. B. (1996). Structural alignment in similarity and difference judgments. Psychonomic Bulletin and Review, 3(2), 227-230.

Medin, D. L., Goldstone, R. L., Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254-278.

Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. Journal of Experimental Psychology: Learning, Memory and Cognition, 23(1), 54-70.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.

Solomon, K. O., & Barsalou, L. W. (2000). Grounding concepts in perceptual simulation: II. Evidence from property verification, Cognitive Psychology. 43(2), 129-169

Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327-352.

Wu, L. L., & Barsalou, L. W. (2001). Grounding concepts in perceptual simulation: I. Evidence from property generation, Under Review

Acknowledgments

We would like to thank Zachary Estes, Phil Johnson-Laird and two anonymous reviewers for their comments on this paper. This research has been supported in part by grants from the James S. McDonnell Foundation and the National Science Foundation (BCS # 0078945) to Vladimir M. Sloutsky.

The /s/ morpheme and the compounding phenomenon in English

Jenny Hayes[1] (J.Hayes@herts.ac.uk)
Victoria Murphy[1] (V.A. Murphy@herts.ac.uk)
Neil Davey[2] (N.Davey@herts.ac.uk)
Pamela Smith[1] (P.M.Smith@herts.ac.uk)
Lorna Peters[2] (L.Peters@herts.ac.uk)
Departments of Psychology [1] and Computer Science [2], University of Hertfordshire, College Lane, Hatfield, AL10 9AB, United Kingdom.

Abstract

Compound words with irregular plural nouns in first position (e.g. mice-eater) are produced far more frequently than compound words with regular plural nouns in first position (e.g. *rats-eater), (Gordon, 1985). Using empirical evidence and neural net modelling, the studies presented here demonstrate how a single route, associative memory based account might provide an equally, if not more, valid explanation of this phenomenon than the standard dual mechanism based theory (Marcus, Brinkmann, Clahsen, Weise & Pinker, 1995).

1. Introduction

1.1 The Compounding Phenomenon

Psycholinguistic research has shown that English compound words with irregular plural nouns in first position (e.g. mice-eater) are produced far more frequently than compound words with regular plural nouns in first position (e.g. *rats-eater), (Gordon, 1985).

1.2 The Dual Mechanism Model's Explanation of Compounding

The dual mechanism model (Pinker, 1991), proposes that irregular nouns and their plurals are stored as memorised pairs of words in the mental lexicon (e.g. mouse-mice) but that regular plurals are produced by the addition of the /s/ morpheme to the regular stem at a post lexical stage (e.g. rat + s = rats). Compounds are created in the lexicon. Thus as irregular plurals are stored in the lexicon they are available to be included within compound words. However, as only the singular stems of regular nouns are stored in the lexicon the plural form is never available to be included within compound words (Marcus et al, 1995).

1.3 A Single Route Associative Memory Based Explanation of Compounding

An alternative explanation of this compounding phenomenon based on the frequency and patterns of occurrence of items in the linguistic input has not been explored fully. However an explanation of this sort may explain the treatment of both regular and irregular plurals in compounds (Murphy, 2000). Frequency counts of a sample of the CHILDES (Child Language Data Exchange System) corpora (McWhinney & Snow, 1985) have shown that the plural /s/ morpheme is a perfect predictor of word finality and furthermore, the plural /s/ morpheme is never followed by a second noun. Importantly, the reverse pattern is found with the possessive /'s/ morpheme since it is always followed by a second noun. Therefore, it might be that a noun rarely follows the regular plural /s/ morpheme (i.e. patterns such as "*rat/s/ chaser" do not occur) because the pattern "noun – morpheme /s/- noun" is reserved for marking possession (such as rat's tail). Interestingly in other languages that do not have this competition between the plural and possessive morpheme such as Dutch (Schreuder, Neijt, van der Weide & Baayen, 1998) and French (Murphy, 2000), regular plurals are allowed within compounds. Irregular plurals may, however, appear in English compounds as they are not formed by the addition of the plural /s/ morpheme. Thus, irregulars do not compete with the possessive structure and as such may be followed by a second noun in a compound. This polyfunctionality of the /s/ hypothesis is explored here using three neural net simulations and an empirical study.

2. Neural Net Modeling

An associative memory-based account of inflectional morphology has been investigated in numerous connectionist models. Several models have successfully simulated the putative dissociation between regular and irregular inflection for both verbal morphology (Daugherty & Seidenberg, 1994) and plural morphology (Plunkett & Juola, 1999) using a single learning mechanism and no explicit rules. Furthermore, as well as being able to learn mappings from input to output, connectionist models have also been able to learn sequential mappings (Elman 1990). Thus it is predicted that a single route associative memory system could learn that the inclusion or omission of the regular plural morpheme /s/ is influenced by where that /s/ morpheme occurs in a sequence of language input. Three neural net

models are considered here. The first investigates whether the presence of /s/ predicts word finality. The second and third models analyse whether learning about the word that follows an /s/ morpheme is sufficient to drive learning about compound formation in English.

2.1 Experiment 1.

Experiment1, was designed to test the degree to which /s/ indicates word finality in a stream of concatenated letters. A neural network was trained on a concatenated stream of 200 sentences of child directed speech taken from CHILDES (MacWhinney & Snow, 1985). A word-ending marker was attached to each word and the words (including a word-ending marker) were concatenated to form a stream of 3596 letters. The network was trained on 200 passes through the sequence of letters. Each letter was encoded using one of 26 random 5-bit vectors (one for each letter in the alphabet). The word-ending marker was encoded using a 27th 5-bit vector. The network was required to predict the next letter it expected to occur given the letters it had seen previously. The network consisted of 5 input units, 30 hidden units, 5 output units and 35 context units. A fully recurrent and a SRN (Elman, 1990) architecture were tested and produced similar results. It was hypothesised that on a next letter prediction task of this kind, a neural network would learn that after the input /s/ there was a high probability that the next input would be a word ending marker.

Test Set And Results: As predicted, at the beginning of a word the error was high but as more letters were presented to the network the error decreased until it was at its lowest at the end of the word. The network's ability to learn that [–s] is a good predictor of word finality was tested using 19 unseen words that ended in /s/ and 19 unseen words that ended in other letters. The network was more accurate (i.e. the error was lower) at predicting a word ending marker after an /s/ than after all other letters combined ($t=-2.08$, $df=18$, $p<0.05$). This simulation was completed to confirm that a model with a single learning mechanism and no explicit rules, trained on child directed speech, could learn that after /s/ there was a high expectancy that the next item would be a word-ending marker. Interestingly, /s/ is associated with word finality even though /s/ appears in the middle of numerous words. This overwhelming pattern of /s/ at the end of a word may influence language learners to omit /s/ from the middle of words such as compounds.

2. 2 Experiment 2.

The aim of this experiment was to examine how highly consistent patterns in the input (i.e. that a plural noun is never followed by another noun while a possessive noun is always followed by a second noun) might drive learning about how to manipulate plurals within noun-noun compounds. The network was required to predict the next word it expected to occur given the words it had seen previously. It was impossible for the network to predict the exact word that followed in the input. However, the network was expected to learn which syntactic category the next item would come from. Thus the network was expected to make a first order distinction between the function of nouns and verbs, determiners and adjectives (Elman, 1990). Furthermore from these induced syntactic categories the network was expected to learn a second order distinction that only "verbs" could appear after some /s/ morphemes and only "nouns" could appear after other /s/ morphemes. It was impossible for the network to distinguish between the possessive and the plural /s/ as both were encoded in exactly the same manner in the input. However, the network was trained on one group of words that were represented as having the properties of possessives, plurals and singulars, a second set was only represented as singulars and plurals and a third group was only represented as singulars and possessives. It was predicted that the tokens making up these three groups of words would cluster together as three distinct sets in the hidden layer representations. The network was trained on a concatenated stream of 2000 legitimate English sentences constructed from a lexicon of 38 words. A sentence-ending marker was attached to each sentence and the sentences (including the sentence-ending marker) were concatenated to form a stream of 14,600 words. The network was trained on 10 complete passes through the sequence of words. Each word (including the sentence-ending marker) was encoded using a 39-bit localist coding scheme. The presence or absence of /s/ at the end of a word was also explicitly coded. A simple recurrent network was used so that at any point in time the state of the hidden units at the previous time step were used as additional input (Elman, 1990).

Results: Figure 1, shows a typical representation of the first two principal components of the hidden unit representations. The dotted line superimposed on the PCA diagram shows the divide between the way nouns and verbs are represented in the hidden units. The network has also represented determiners and adjectives separately. Most interestingly, nouns which were included in the training set as both "plurals and possessives", items that were only included as "possessives" and items which were only included in the "plural" form are all represented separately within the cluster of words ending in /s/. Therefore, Experiment 2 showed that a neural net was able to make some differentiation between the plural and possessive /s/ depending on the words that followed it in the input even though the two types of /s/ had exactly the same encoding characteristics.

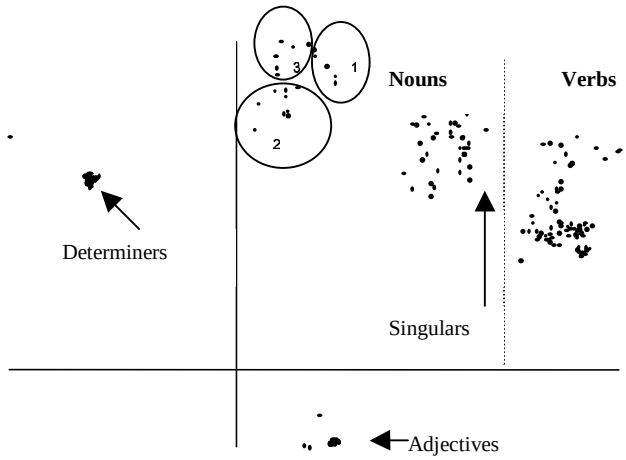


Figure 1. First two principal components of the hidden layer representations in Experiment 2. Area 1 corresponds to the representational area occupied by items that appeared in the context of both plurals and possessives. Area 2 corresponds to the representational area occupied by items that appeared only in a possessive context and area 3 corresponds to the representational area occupied by items that appeared only in the plural context.

2. 3 Experiment 3.

In Experiment 2, the network was able to group nouns that in the training set were behaving as "plural and possessive" or as "plural" or "possessive" only. However, the network could not totally disambiguate plurals from possessives. In this third simulation, the network that was used in Experiment 2 was amended to include an extra input unit that encoded whether the subject of the sentence in which the word occurred was either a plural or a singular noun. Hence, although both "plural" and "possessive" words were coded as ending in /s/, only plural items were encoded as ending in /s/ **and** being plural, as possessive words were encoded as ending in /s/ but being singular. The same training set and task from Experiment 2 was employed. With the addition of this minimal semantic information, the network is expected to disambiguate "plural" nouns from "possessive" nouns. It was predicted that in the hidden units, the plural and possessive nouns would be represented separately.

Results: Figure 2, shows a typical representation of the first two principal components of the hidden unit representations. From the PCA it is evident that once again nouns, verbs, determiners and adjectives are represented separately in the hidden units. With the addition of the semantic information it is now evident that singular, plural and possessive nouns are all represented separately. Singular and possessive nouns (both of which are actually singular nouns) are located in a similar position but plural nouns are now represented quite separately. Interestingly, both plurals and singulars i.e., items that may be followed by a verb lie in similar positions on the x axis, while the possessives are clustering with adjectives i.e., with other items that are followed by nouns. Therefore, Experiment 3 shows that learning about the different functions of the /s/ morpheme is enhanced with the addition of the very minimum of semantic information

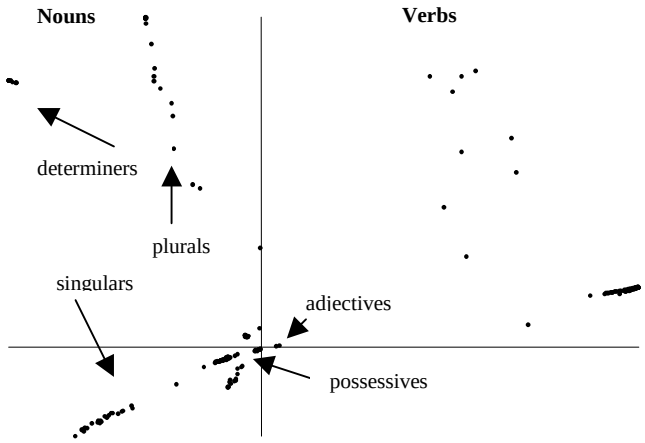


Figure 2. First two principal components of the hidden layer representations in Experiment 3

2.4 Discussion of Neural Net Modelling

From Experiment 1, it is evident that a neural net model trained on child directed speech was able to learn that /s/ is strongly associated with word-finality (even though /s/occurs frequently in the middle of words). This overwhelming pattern of /s/ at the end of words might influence English language learners to omit /s/ from the middle of words such as compounds. Experiment 2, showed that the net was able to learn that /s/ followed by one set of words was different from /s/ followed by a different set of words even though the /s/ was encoded in exactly the same way in the input. The same might be true for the language learner. Both the possessive /s/ and the plural /s/ sound the same phonetically but the patterns in which the two different types of morpheme appear in the input may be sufficiently distinct as to indicate that one type of morpheme performs a specific linguistic function and the other performs another type of linguistic function. From Experiment 3, it is evident that learning that the plural and possessive morphemes are only legal in certain sequences may be refined as the child learns that semantically, the plural morpheme refers to many things while the possessive morpheme usually refers to one thing.

Table 1: Examples of Compounds used as stimuli in Experiment 4.

Group	Example of contextualising sentence	Examples of compounds
(1) Possessive nouns	Last week, I left my purse in a London taxi. Luckily, I managed to signal to the	taxi's driver
(2) Regular plural nouns	I feed four cats, a Burmese, a Siamese and two lovely old Persians. I enjoy being a	cats feeder
(3) Irregular plural nouns	Women always get lowly jobs. In the nursery rhyme the farmer's wife is nothing more than a	mice chaser
(4) Comparatives or superlatives	Greg is very modest. He was amazed to hear that his song is still the record company's	biggest seller
(5) singular nouns ending in phoneme /s/	We'll have a larger lawn and mowing the grass will take longer. I'm thinking of employing a	grass cutter
(6) singular nouns ending in a phoneme other than /s/	Stephen is so skilled at mixing cocktails that the hotel want him to work permanently as a	drink server

3. Experiment 4:Compound Processing Study

The compounding phenomenon was further tested by asking 22 native adult English speakers to process "noun-noun" compounds as part of an "on-line" lexical decision (LD) task. This is important as most research has focussed on production (e.g., Gordon, 1985; Murphy, 2000). In this experiment, participants were required to categorise 216 compounds as being words or non-words having seen them presented visually on a computer screen and proceeded by a contextualising sentence. The mean response time for processing different types of compound (see stimuli) was examined in a within subjects design. Two hypotheses were tested to examine the associative explanation of compounding. These were:- (1) If, as the first neural net simulation (Experiment 1) confirmed, language users associate /s/ with word finality, then compounds in which the first noun ends in /s/ should be processed more slowly than compounds that do not include a first noun ending in /s/. (2) If, as the polyfunctionality hypothesis indicates, /s/ appearing in the middle of a compound made up of two nouns is interpreted as indicating possession rather than plurality, then compounds containing possessive nouns should be processed more quickly than compounds containing plural nouns**.**

Two further hypotheses were investigated to test the dual mechanism model's explanation of compounding. (1) Pinker (1991) stated that:

> "because it categorically distinguishes regular from irregular forms, the rule-association hybrid predicts that the two processes should be dissociated from virtually every point of view…..[including] reaction time ……. " (p 253).

However, the dual mechanism model makes no directional prediction as to which type of morphology might be processed more quickly. Beck (1997) asked native adult English speakers to supply the past tense of a series of present tense regular and irregular verbs. Beck found that both low (mean response time 477 msec) and high (mean response time 508 msec) frequency regulars were produced more quickly that both low (mean response time 581) and high (mean response time 535 msec) frequency irregulars. By collecting reaction times in Experiment 4, it was possible to test the speed at which the two types of morphology were processed within compounds in a lexical decision task. The following two hypotheses were tested, (1) compounds containing irregular morphology and compounds containing regular morphology will be processed at different speeds (2) more specifically, compounds containing irregular plurals and compounds containing regular plurals will be processed at different speeds.

3.1 Stimuli

The first noun in each compound was taken from one of six groups. These were: - (1) possessive nouns (2) regular plural nouns (3) irregular plural nouns (4) comparative or superlatives (5) singular nouns which ended in phonetic /s/ (6) singular nouns which ended in a phoneme other than /s/. Each group of first nouns were matched for frequency. The second noun in each compound was a deverbal noun, i.e., a noun that is formed from a verb (e.g.s, walker, chaser). All second nouns were matched for frequency. Table 1.shows examples of each type of compound tested. The apostrophe was omitted from all the possessive nouns thus making it impossible to

distinguish between the plural and possessive nouns (cf. the neural net used in Experiment 2). However, each compound was preceded by a contextualising sentence, (cf. the neural net used in Experiment 3) which pilot work had shown would lead the first noun in the compound (e.g., rats in *rats eater) to be interpreted appropriately. To ensure uniform treatment of all stimuli, contextualising sentences also preceded the first noun even where they were not taken from the plural or possessive groups (see Table 1. for examples of contextualising sentences).

3.2 Results

Table 2. Mean reaction times

	Mean reaction time in milliseconds	Standard deviation of Mean reaction time	Difference between means in milliseconds
Comparisons to test the associative account			
All groups ending in /s/	1285	465	
Final phoneme other than /s/	1205	455	80*
Regular plurals	1277	492	
Possessives	1191	437	86*
All groups containing regular morphology	1231	450	
Irregular plurals	1339	470	108*
Regular plurals	1277	492	
Irregular plurals	1339	470	62
Comparison to test the time difference between processing plurals and processing other types of morphology			
Regular and irregular [illegible]	1291	479	
All other items of morphology	1188	424	103*

* Difference reliable at alpha = 0.05

Mean reaction times in milliseconds are shown in Table 2. Two planned comparisons were conducted to test the associative explanation of compounding. Firstly, compounds with a first noun ending in /s/ were processed more slowly than compounds where the first noun ends in another letter (a mean difference of 80 milliseconds) ($t=4.41$, $df=21$, $p<0.05$). It took participants an average of 86 milliseconds longer to process compounds containing regular plurals than compounds containing possessive nouns ($t=2.195$, $df=21$, $p<0.05$). These two findings support the outcomes of the neural net simulations in Experiments 1,2 and 3. Two planned comparisons were conducted to test the dual mechanism model's explanation of compounding. All types of regular morphology were processed more quickly that irregular plurals (mean difference of 108 milliseconds) ($t=3.22$, $df=21$, $p<0.05$). It took participants an average of 62 milliseconds longer to process irregular plurals than regular plurals but this difference was not found to be reliable. A post hoc analysis was also conducted to test the difference between the time taken to process compounds containing both types of plural and the time taken to process compounds containing other types of morphology (mean difference of 103 milliseconds). A Tukey's HSD test found this difference to be reliable ($F=11.29$, $df=21$, $p<0.05$).

3.3 Discussion

Consider first the two hypotheses that tested the associative explanation of compounding. It took longer to process compounds in which the first noun ended in /s/ than compounds which did not include a first noun ending in/s/. Language users, like the network in Experiment 1, found it harder to process /s/ in the middle of a word. Furthermore, possessive nouns are easier to process than plural nouns in the middle of compounds even though they share exactly the same phoneme. The /s/ morpheme in the middle of a word seems to indicate possession not plurality. Consider next the two hypotheses that tested the dual mechanism's explanation of compounding. Similar to Beck's (1997) production data, it took participants in this experiment less time to process all types of regular morphology than it took them to process irregular plurals (the only type of irregular morphology tested). However, no difference was found in the time taken to process regular and irregular plurals, despite Pinker's (1991) prediction that the two types of morphology should be dissociated "from virtually every point of view" (p 253). Interestingly, a reliable difference was found when reaction times to both types of plural were collapsed together and compared with reaction times to comparatives and superlatives and possessives (all items of regular morphology) collapsed together. Adult language users find it relatively difficult to process either type of plural in the middle of compounds. However, contrary to the predictions of the dual mechanism model, adults seem to have no difficulty processing other items of regular morphology (i.e., items which are produced at a post-lexical stage) within compounds (cf. Marcus et al, 1995). It has been argued elsewhere, that due to the competition with the possessive structure, language users omit regular plurals from compounds. Furthermore, guided by this template, i.e., that plurals do not occur within compounds, mature language users also begin to

omit irregular plurals from compounds, (Hayes, Smith & Murphy, unpublished manuscript)

4. General discussion

Experiment 1 showed that /s/ is associated with word finality. Furthermore, participants in the empirical study took longer to process compounds which contained /s/ than compounds that did not contain /s/ (regardless of what type of /s/ it was). Both strands of evidence would seem to indicate that /s/ is linked to word finality despite the fact that /s/ occurs frequently in the middle of many words. This overwhelming pattern of /s/ at the end of words might influence language learners to omit /s/ from the middle of words such as compounds. Evidence from Experiment 2, showed that a neural network was able to learn that /s/ followed by one set of words was different from /s/ followed by a different set of words even though the /s/ was encoded in exactly the same way in the input. From Experiment 3, it was evident that learning that the plural and possessive morphemes are only legal in certain sequences may be refined as the child learns that semantically, the plural morpheme refers to many things while the possessive morpheme usually refers to one thing. The empirical evidence also showed that one type of /s/ morpheme was processed more quickly within compounds than another type of /s/ morpheme, although it was denoted by same phoneme. Both the possessive /s/ and the plural /s/ sound the same phonetically but the patterns in which the two different types of morpheme appear in the input seem to be sufficiently distinct to indicate that one type of morpheme performs a specific linguistic function and the other performs another type of linguistic function. As well as providing support for the associative account Experiment 4 also calls into question the dual mechanism model's explanation of compounding. No difference was found between the time taken to process regular and irregular plurals. However, participants were able to process some items of regular morphology within compounds relatively quickly. It seems to be plurals (of either kind), rather than items of regular morphology, that adults find difficult to process within compounds.

The three models taken together with the empirical work provide evidence for an associative account of compounding. In this associative account the language user relies on properties of the linguistic input itself and not on distinct ways of representing "rules" versus associations to drive linguistic behaviour. More specifically in the case of compounding, the language user learns that the /s/ morpheme tends to nearly always occur at the end rather than in the middle of a word. Furthermore, the language learner is sensitive to the fact that the same /s/ morpheme occurs in different patterns in the input. With the addition of the absolute minimum of semantics, namely the numerical context in which the phrase is uttered, the language learner seems able to differentiate between the plural and the possessive morpheme. The possessive morpheme may be followed by a second noun but the plural morpheme may not be followed by a second noun. When faced with a noun-noun compound the language user may delete the plural morpheme from the end not because regular items of morphology are different in kind from irregulars and represented as rules but simply because this pattern is used to denote possession not plurality. Thus the dissociation between the treatment of regular and irregular morphology in compounds may result form the fact that one type of morphology is subject to competition with the possessive morpheme but the other is not. As this alternative hypothesis is explored further, it may become apparent that this plural dissociation in compounds is not good evidence to support the dual mechanism model.

References

Beck, M-L., (1997). Regular verbs, past tense and frequency: tracking down a potential source of NS/NNS competence differences. *Second language Research, 13,* 93-115.

Daugherty, K. G. & Seidenberg, M. S. (1994). Beyond rules and exceptions. In S. D.Lima, R. L. Corrigan & G. K.
Iverson (Eds.), *The reality of linguistic rules.* Amsterdam: John Benjamins.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Gordon, P. (1985). Level-ordering in lexical development. *Cognition, 21* , 73-93.

Hayes, J.A., Smith, P.M., Murphy, V.A. (Unpublished manuscript). Modality effects in compounding with English inflectional morphology.

Marcus, G.F., Brinkmann, U., Clahsen, H., Weise, R. & Pinker, S., (1995). German inflection: The exception that proves the rule. *Cognitive Psychology, 29,* 189-256.

MacWhinney, B. & Snow, C. E. (1985). The Child Language Data Exchange System. *Journal of Child Language, 12,* 271-296.

Murphy, V. A. (2000). Compounding and the representation of inflectional morphology. *Language Learning, 50,* 153-197.

Pinker, S. (1991). Rules of language. *Science, 253, 530-535.*

Plunkett, K. & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science, 23,* 463-490.

Screuder, R., Neijt, A., van der Weide, F. and Baayen, R.H., (1998). Regular plurals in Dutch compounds: Linking graphemes or morphemes? *Language and Cognitive Processes*, 13, 551-573.

Interactional Context in Graphical Communication

Patrick G. T. Healey
Information, Media and Communication Research Group
Department of Computer Science,
Queen Mary University of London.

Simon Garrod, Nicholas Fay
HCRC
Department of Psychology,
University of Glasgow.

John Lee and Jon Oberlander
HCRC and Department of Architecture,
Division of Informatics,
University of Edinburgh.

Abstract

A body of empirical evidence indicates that interactional context has a key influence on the form and interpretation of language. This paper reviews a series of experiments which indicate that interactional context also plays a key role in the interpretation of drawings and sketches. Two graphical communication tasks, analogous to definite reference tasks, are described. The findings from these tasks show significant parallels between the mechanisms of co-ordination in graphical dialogue and natural language dialogue. Specifically; the coordination of graphical representation types by 'dialogue' participants, the contraction of recurrent 'graphical referring expressions', effects of direct interaction on the use of abstract drawings, and the development of community-specific graphical conventions.

Interactional Context in Dialogue

Conversation is a, if not *the*, key context of understanding for language. People's use of language to represent objects, events and situations is sensitive to, amongst other things; who they are speaking to, the mutual availability of referents, the history of their conversation and their (dis)joint membership of cultural and linguistic sub-communities (Hymes, 1972; Clark, 1998). Evidence for the direct influence of interactional context on interpretation and understanding comes from a variety of sources (see Krauss and Fussell, 1996, for a review). One example is provided by work on the Collaborative Model of dialogue (Schober and Clark, 1989; Clark and Wilkes-Gibbs, 1986). Wilkes-Gibbs and Clark (1992) have shown that full understanding of referring expressions depends on the degree of active participation in conversation by speaker and addressees. Non-active participants in a conversation, such as passive side-participants, overhearers, or bystanders, have more difficulty in interpreting referring expressions than active participants. This is observed even when, in gross informational terms, they are equivalent to active participants.

A second example of the influence of interactional context comes from studies of conceptual and linguistic co-ordination in dialogue. Garrod and Anderson (1987) have shown that conversational partners tend to match or 'entrain' on the form and interpretation of utterances during interaction. Where several types of semantically distinct referring expressions are possible for describing a location, people show a strong preference for matching the type of expression used by their conversational partner. Brannigan, Pickering and Cleland (2000) have observed similar entrainment effects with syntax. Garrod and Anderson (1987) argue that these dialogue phenomena reflect the operation of a basic dialogue co-ordination mechanism which simplifies the processes of production and comprehension in interaction.

Interactional Context in Graphical Dialogue

Intuitively, it might be supposed that graphical representations would be less sensitive to interactional context. One reason for this is that the production and use of drawings and sketches is normally treated, and analysed, as an activity more akin to monologue than dialogue (cf. Scaife and Rogers, 1996). There is evidence, however, that this underestimates both the actual and potential use of drawing activities as a mode of interaction. Anecdotally, drawings are often incrementally produced and modified as part of a conversational exchange. For example, sketch maps and explanatory diagrams form a familiar extension of many routine conversations.

van Sommers (1984) provides evidence from a questionnaire study that approximately half of routine, non-work, drawing activities take place with or for an audience. Although van Sommers does not report how often these interactions involve direct graphical exchanges, his findings demonstrate the variety of interactional contexts in which drawing occurs. The most frequently cited category is the production of sketch maps of a local area, either as part of an explanation or in order to give directions. The second most frequently cited category relates to activities with children including; games and amusements, teaching or helping with homework and helping children learn to draw. Additional categories of 'public' drawing include; sketching of hair, makeup and clothing, sketching house plans, drawing to express feelings, defacing pictures and drawing people.

The collaborative development and modification of sketches is a feature of many specialised work related interactions, such as architect-architect and architect-client (Neilson and Lee, 1994). We estimate that in the architects' practice studied by Healey and Peters (2001) approximately 30% of daily drawing activities occurred as an integrated part of a conversational exchange. Engle (1998) provides experimental evidence that graphics, gesture and language combine in explanatory dialogues to create composite communicative signals (cf. Clark, 1996). Overall, there is a *prima facie* case that sketches and drawings are often closely integrated into interaction and that this may have significant implications for their interpretation.

A second source of scepticism about the role of interactional context in the interpretation of sketches and drawings is the intuition that drawings and sketches are easier to interpret than language. Arguably, many of the interactional influences on language interpretation are associated with the conventional nature of linguistic representation. Coordinated interpretation of utterances requires the concerted application of conventions. Interaction is used to maintain and modify those interpretations. Drawings and sketches can exploit iconicity to provide a less arbitrary form of representation. Consequently, we might suppose that they would be less dependent on interaction to secure their interpretation. This idea is most plausible for, say, sketches of buildings or people but it does not cover the range of uses to which sketches and drawings are put. Explanations involving sketches of Venn diagrams or Euler circles provide perhaps the most obvious counterexample.

Experiments on Graphical Dialogue

The present paper summarises the findings from a series of experiments which, considered together, provide evidence that the interpretation of drawings and sketches is sensitive to interactional context. In particular, that interactional context affects the form, interpretation and understanding of sketches; and that the mechanisms and processes that give rise to these effects show important parallels to those identified for natural language dialogue.

The findings reported below are drawn from experiments involving two basic referential communication tasks, the Concept Drawing Task and the Music Drawing Task, in which pairs of subjects communicate about a variety of concepts using exclusively graphical means. These tasks can be thought of as two-way or conversational variants of the party game 'Pictionary' (TM).

The Experimental Tasks

The basic Concept Drawing Task uses an ordered list of twelve concept words drawn from the categories; places (e.g., "theatre", "art gallery", "museum"), people (e.g., "Robert de Niro", "Arnold Schwarzenneger", "Clint Eastwood"), television programmes (e.g., "drama", "soap-opera", "cartoon"), objects (e.g., "television", "computer" "microwave"), and abstract concepts (e.g., "loud", "homesick", "poverty"). One participant, the 'Drawer', is given an ordered list of twelve words. Their partner, the 'Chooser', is presented with an unordered list of the same twelve words plus four distractors. The task is for the Drawer to take each word in turn from their list and produce a sketch of it so that their partner, the Chooser, can identify the concept depicted. The aim is for the Chooser to determine the original ordered list of twelve concept words that the Drawer started with.

The basic Music Drawing Task is similar to the Concept Drawing task but uses pieces of music in place of concept words. The pieces are relatively unknown 30 second piano solos in a variety of genres and styles. In the typical procedure, the Drawer and the Chooser are seated in separate rooms. The Drawer listens to a target piece of piano music and produces a sketch of it. The Chooser has two pieces of music, the target and a distractor, and tries to select which piece is the one depicted by the Drawer. Playback of the pieces is self-paced and all drawing takes place on a shared virtual whiteboard which logs the drawing data for analysis (Healey, Swoboda, King, forthcoming).

In both tasks, subjects are free to draw anything they like; the only restriction is that they do not use letters or numbers. The types of drawing produced for each concept or piece of music varies substantially between pairs, some examples are provided in Figures , 2 and 3. All things being equal, each pair tends to establish their own conventional solutions to the communication problems posed by the task. Subjects appear to find both tasks enjoyable and engaging and perform them with above chance accuracy.

Effects of Interactional Context

A number of experiments have been performed using these tasks which suggest important parallels between the effects of interactional context on graphical and verbal dialogue. Here we provide an overview of the findings from these experiments and discuss their implications for investigations of graphical representation and models of human interaction.

Interactional Entrainment. One of the simplest pieces of evidence for effects of interactional context on the use of drawing comes from the Music Drawing task. Participants in this task produce drawings that can be reliably classified into two basic types[1]; 'Abstract' and 'Figurative' (Kappa = 0.9, N = 287, k= 2). Abstract drawings, illustrated in Figure 1,

[1]For ease of exposition a third, 'Composite, type is not discussed here

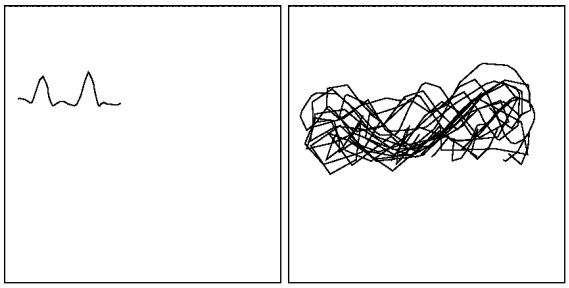


Figure 1: Example Abstract Drawings from Two Trials of the Music Drawing Task

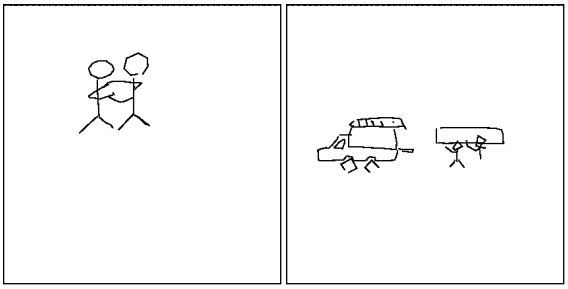


Figure 2: Example Figurative Drawings from Two Successive Trials of the Music Task

typically involve graph-like representations of e.g., pitch, melody, rhythm or intensity. By contrast, Figurative drawings, illustrated in Figure 2 typically depict recognisable objects, figures or scenes. Where pairs of participants in the task both take the role of Drawer (either by alternating roles or in manipulations in which both participants draw at the same time) they show a reliable tendency to match each another in their use of drawing the Figurative and Abstract drawing types (Healey, Swoboda, Umata, & Katagiri, 2001). As noted above, this pattern of entrainment between the participants in an interaction is also established for semantic and syntactic aspects of utterances in dialogue (Garrod and Anderson, 1987; Brannigan, Pickering and Cleland, 2000). Garrod and Anderson (1987) argue that entrainment constitutes a basic mechanism through which conceptual co-ordination is achieved in dialogue.

Contraction of Recurrent References. The procedure for the Concept Drawing task typically requires pairs to repeat the same set of twelve target words, in different orders, over several trials. This manipulation ensures that each word is drawn, and identified, several times by each pair. This is designed to reproduce the procedure followed by Clark and Wilkes-Gibbs (1986) who investigated the production of recurrent (verbal) referring expressions by conversational partners. Clark and Wilkes-Gibbs found that both the average number of words and average number of turns used to refer to a target item (in their case a tangram figure) rapidly declined with the number of repetitions. Experiments with the concept drawing task show the same pattern of reduction. This is illustrated in Figure 3 which shows a sequence of four trials (ordered left to right and top to bottom). Where target concepts recur, the drawings that represent them quickly become simplified. This is indexed both by simple quantitative measures such as the amount of 'ink' and number of lines used, and by their complexity as estimated by human judges and calculations of their visual complexity.[2]

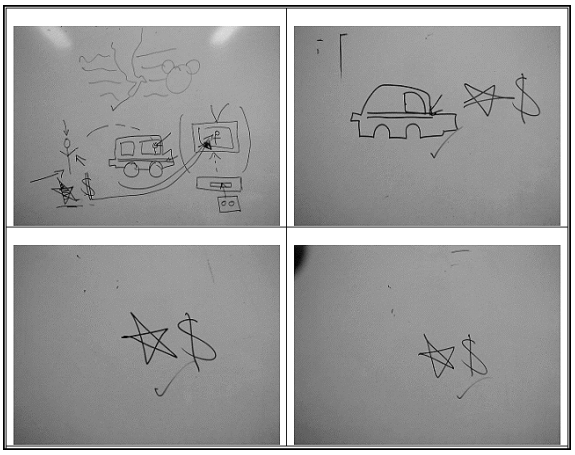


Figure 3: A Sequence of 'Robert deNiro's from the Concept Drawing Task

Experiments to evaluate the effects of these contractions on the intelligibility of the drawings for non-participants are currently in progress. However, it appears likely that they will have a substantial effect. The first drawing in Figure 3 has a number of elements that might allow a non-participating observer to guess the identity of the individual depicted. For example it includes a sketch map of Italy, sketches of a TV and VCR, and an image of a taxi (which refers to a de Niro film). By contrast, the last sketch in the sequence, consisting of a star and a dollar sign would be much harder to decipher.

Effects of Direct Interaction. Experiments with the Music Drawing task have investigated the influence of level of communicative interaction between participants on the type of drawing (Abstract or Figurative) that they produce. The basic contrast is between an interactive and non-interactive version of the task (Healey et al., 2001; Healey, Swoboda, Umata and Katagiri, forthcoming). In the

[2]The analysis of visual complexity is based on a psychophysical measure developed by Pelli Burns Farell and Moore (in press) and is based on the formula: Complexity = Perimeter^2 / Ink.

non-interactive version, subjects alternate between acting as Drawer and Chooser on each trial and the whiteboard is configured to prevent the Chooser from drawing. In this version of the task each trial approximates to a single turn in the communicative exchange. In the interactive version the task is altered so that both members of a pair draw at the same time. They have one piece of music each and must determine, using only drawing, whether their pieces are the same or different. In this case there is a richer communicative exchange. In addition to producing drawings of their pieces, subjects employ devices such as arrows, underlining, and circling to query and revise various aspects of their drawings. Each trial in the interactive task thus approximates to a number of 'conversational' turns.

The effect of the difference in level of communicative interaction can be seen in Table 1 (the 'Composite' category refers to drawings that combine Figurative and Abstract elements). Where both members of a pair can interact directly on the whiteboard, they rely primarily on the Abstract drawings. In the non-interactive task, where they are alternating between drawing and choosing, they rely primarily on Figurative drawings.

Table 1: Distribution of Drawing Types in the Music Drawing Task

	Drawing Type		
Task	Abstract	Figurative	Composite
Interactive	59%	21%	16%
Non-Interactive	27%	64%	8%

Further evidence for the importance of direct interaction comes from analysis of the logs of drawing activity (Healey, Swoboda, Umata and Katagiri, forthcoming). The Abstract and Figurative drawing types are not distinguishable in terms of the number of lines or ink (pixels) involved in producing them, nor in terms of the accuracy of responses associated with their use. Considerations of the efficiency or effectiveness of the two drawing types alone do not appear to explain their pattern of use. However, drawing activities overlap approximately 20% more when subjects produce Abstract drawings than when they produce Figurative drawings. This suggests it is the availability of specific mechanisms of communicative interaction, such as the circling and underlining of each others drawings, that is critical to the co-ordinated use of the Abstract drawings.

Community-based Conventions. Perhaps the most interesting parallel between graphical and verbal dialogue comes from experiments on the emergence of graphical conventions in experimental 'sub-communities' (cf. Garrod and Doherty, 1994).

Data from an unpublished experiment with the Music Drawing task demonstrates that, for this task at least, the patterns of co-ordination in drawing style that emerge within sub-communities are specific to those sub-communities (cf. Healey, 1997). The experiment takes place in two phases. In the first 'convergence' phase experimental sub-communities consisting of sub-groups of six people are formed. Subjects themselves are unaware of this sub-group manipulation, from their perspective the experiment consists of a series of rounds of Music Drawing with a different partner each time. During the convergence phase, the composition of pairs is controlled so that they are always made up of individuals from within the same sub-group. This continues for four rounds thus allowing for a history of interactions to build up within each sub-group. On each round subjects perform the interactive version of the Music Drawing Task for 12 trials.

The second, experimental, phase occurs in the fifth round. In this round two conditions are compared; same-group pairs who are composed, as before, of subjects from within a single sub-group and cross-group pairs who are composed of subjects drawn from different subgroups.[3] Same-group and cross-group pairs have equivalent task experience and, as noted, are unaware of any sub-group manipulation. Nonetheless they are reliably different in their use of the Drawing types. Multinomial regression analysis shows a reliable effect of the group manipulation on the distribution of Drawing types ($Chi^2_{(3)}$=25.44, p=0.00, n=516). The percentages are shown in Table 2.

Table 2: Use of Drawing Types in Pairs Drawn from the Same or Different Subgroups

	Drawing Type		
Task Version	Abstract	Figurative	Composite
Same-group	62.7%	11.1%	18.1%
Cross-group	41.3%	32.9%	15.8%

These results indicate that the co-ordination on particular drawing types that develops within the experimental sub-communities is community-specific. Subjects in the cross-group interactions use a more mixed profile of drawing types. This suggests that the graphical conventions established within sub-communities do not readily transfer to interaction outside those communities. Healey (1997) reports parallel results for verbal dialogues about spatial locations. In this case the types of spatial referring expressions established within sub-communities

[3]The original design employed three experimental subgroups but for ease of exposition only two are reported here.

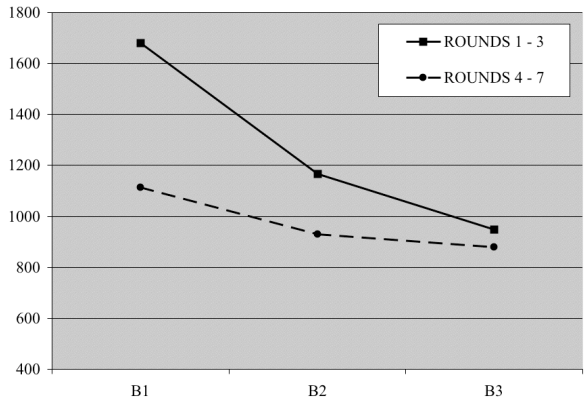


Figure 4: Reduction in Visual Complexity of Concept Drawings with Repetition in a Round (B1-B3 = Blocks of recurring items within a round)

during the corresponding convergence phase also proved unstable in 'cross-group' interactions.

Data from a community-based version of the Concept Drawing Task also suggests parallels between communities of graphical and linguistic communicators. The task requires a community group of 8 participants to communicate with each of the other 7 over an extended period of time. In the first round of the experiment they work in 4 pairs with both participants drawing each concept 3 times over the course of the round. In the second round the 8 participants are re-paired and again draw the concepts 3 times. After each round they are re-paired again until every participant has encountered each of the others once and only once.

Figure 4 shows how drawings become increasingly simple (on the Pelli et al. measure) as the experiment proceeds. In the first 3 rounds this simplification process occurs across repetitions of the drawings (shown along the x axis of the figure). However, as the shared interaction within the community begins to develop (i.e., after round 4) the initial drawings in a round become as simple as the final drawings in the round. A similar pattern of results emerges for the communicators accuracy at identifying the concepts conveyed by their partners drawings. These findings are consistent with the idea that as a community becomes established through a common history of interaction so the drawings become conventionalised within the community: Drawings become simpler and more readily interpreted by the members of the community.

The implication of these results is that the processes which establish the conventions for producing and interpreting drawings and verbal descriptions operate in a manner that is directly tied to the character and pattern of interactions in which they were developed and used.

Discussion

The aim of providing an overview of a number of experimental results means that much important detail has been elided from the descriptions of experiments and results provided above. Nonetheless, the results summarised above consistently point to the importance of interactional context in graphical communication.

Like referring expressions in conversation, the form and interpretation of drawings is systematically influenced by the character of the interaction in which they occur. Participants in interactions show a strong tendency to match each others representational style and type. If items recur in an interaction, pairs also tend to develop increasingly abbreviated ways of representing them that are difficult for third parties to interpret. These patterns of change in the form of drawings obtain independently of the particular concept or item being represented. In addition to these basic co-ordination processes of entrainment and abbreviation, there is evidence that level of direct graphical interaction available to participants affects the form of representations they use. In particular, the ability to mark up and modify elements of each others' drawings appears to be important to the sustained use of more abstract representations. Lastly, this paper has presented evidence that interactions within sub-communities lead to the development of community-specific conventions for graphical interaction.

The programmatic rationale for investigating tasks, such as those described above, that involve exclusively graphical communication is the potential they offer for comparison with other modes of interaction. The results summarised above suggest significant parallels between the mechanisms that underpin communicative co-ordination in exclusively graphical and verbal exchanges. As noted above, some of these findings can be accounted for in terms of the collaborative model of grounding (Clark and Wilkes-Gibbs 1986; Clark, 1996) and input-output coordination model (Garrod and Anderson 1987, Garrod and Doherty 1994). The importance of interactional mechanisms, such as localisation, to graphical communication also suggests possible parallels with the mechanisms of conversational repair (Sacks, Schegloff and Jefferson 1974; Schegloff 1992). The viability of applying these explanations to the details of graphical communication is the subject of further work.

Acknowledgements

We gratefully acknowledge the support of ESRC/EPSRC under the PACCIT programme for the project MAGIC: Multimodality and Graphics in Interactive Communication (L328253003), and ATR Media Information Science Laboratories. We are especially grateful to James King, Nik

Swoboda, Ichiro Umata and Yasuhiro Katagiri for their contributions to this research. An earlier version of this paper was presented under the title "Interactional Context in Sketch Understanding" at the AAAI Spring Symposium, Stanford, 25th-27th of March, 2002.

References

Brannigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(B), 13-25.

Clark, H. H. (1996). *Using language.* Cambridge: Cambridge University Press.

Clark, H. H. (1998). Communal lexicons. In K. Malmkjoer & J. Williams (Eds.), *Context in language learning and language understanding* (3rd ed., pp. 63–87). Cambridge: CUP.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Engle, R. (1998). Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In M. Gernsbacher & S. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society* (pp. 321–326).

Garrod, S. C., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Garrod, S. C., & Doherty, G. (1994). Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181–215.

Healey, P. (1997). Expertise or expert*ese*?: The emergence of task-oriented sub-languages. In M. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual conference of the cognitive science society* (pp. 301–306).

Healey, P., & Peters, C. (2001, 18th-20th April). *Notes on turn-taking and topic in drawing-in-interaction.* (Paper presented at the *The 4th International Workshop on Gesture and Sign Language based Human-Computer Interaction* City University, London)

Healey, P., Swoboda, N., & King, J. (forthcoming). *A tool for performing and analysing experiments on graphical communication.* (Paper to be presented at *HCI2002: The 16th British HCI Group Annual Conference*, September 2nd- 6th, South Bank University, London)

Healey, P., Swoboda, N., Umata, I., & Katagiri, Y. (2001). Representational form and communicative use. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 411–416).

Healey, P., Swoboda, N., Umata, I., & Katagiri, Y. (forthcoming, October). *Graphical representation in graphical dialogue.* (Paper to appear in a special issue of the *International Journal of Human Computer Studies* on Interactive Graphical Communication)

Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35–71). New York: Holt, Rinehart and Wilson.

Krauss, R., & Fussell, S. R. (1996). Social psychological models of interpersonal communication. In A. Higgins, E.T. nad Kruglanski (Ed.), *Social psychology: Handbook of basic principles* (pp. 655–701). London: Guildford Press.

Neilson, I., & Lee, J. (1994). Conversations with graphics: implications for the design of natural language/graphics interfaces. *International Journal of Human-Computer Studies*, *40*, 509–541.

Pelli, D. G., Burns, C. W., Farell, B., & Moore, D. C. (in press). Identifying letters. *Vision Research.*

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, *50*, 696–735.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, *45*(2), 185–213.

Schegloff, E. A. (1992). Repair after the next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, *97*(5), 1295–1345.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211–232.

Sommers, P. van. (1984). *Drawing and cognition: Descriptive and experimental studies of graphic production processes.* Cambridge: Cambridge University Press.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183–194.

Diagrams and Descriptions in Acquiring Complex Systems

Julie Heiser and Barbara Tversky (jheiser,{bt}@psych.stanford.edu)
Department of Psychology
Jordan Hall, Building 420
Stanford, CA 94305 USA

Abstract

Complex systems such as a car brake, circulatory system, or legislative system can be conveyed by language or diagrams. Such systems can be presented from structural or functional perspectives. In three experiments, we examine communicating structure and function of mechanical systems (bike pump, car brake, pulley system) by text and diagrams in relation to mechanical ability. By adding arrows, structural diagrams can be enriched to convey functional information. Inferring structure from function was easier than inferring function from structure. Participants high in mechanical ability outperformed low participants except when text perspective matched question perspective. Those with low mechanical ability are at a disadvantage, especially for inferring function from diagrams. Comprehension of complex systems depends in sensible ways on perspective, medium, and ability.

Conveying Complex Systems

When we learn about a new digital camera, attempt to troubleshoot a broken-down car, or try to understand a new finding in neuroscience, we need to understand a complex system. Despite the ubiquity of contact with complex systems, understanding them or interacting with them can be frustrating. The frustrations are due not only to the complexity of the systems but also to the inadequacy of instructions and explanations.

Effective explanations of complex systems have a complexity of their own. Effectiveness depends on the perspective of the information to be conveyed, on the medium of conveying the information, and on the ability and expertise of the learner. Some of these complex interactions have been examined in previous work, though finding generality in the conclusions has been elusive (e.g., Hegarty, et al., 1990; Hegarty, et al., 1993; Mayer & Gallini, 1990; Morrison and Tversky, 2000). More clarity may be achieved by an analysis of the information to be conveyed relative to characteristics of the media and to qualities of individual differences.

Information about complex systems is of two types: structural information, the configuration of parts or topology of the system, and functional information, the sequence of operations and outcomes. The configuration of parts has a spatial or metaphorically spatial structure, and the sequence of operations has a temporal, causal structure. The primary media for conveying complex systems are language and diagrams. With an increasing emphasis on visual displays of information, we found it important to investigate the success that diagrams have in comparison to text in communicating this information. Structural information should be effectively conveyed in diagrams because diagrams use elements and relations in space to convey actual topology. Furthermore, arrows indicating the sequence of operations can be added to a diagram to convey functional information.

There are conflicting results on the relations between medium and ability. Some studies show that people with low ability benefit from diagrams and others show that people with low ability have difficulties extracting information from diagrams (Hegarty 1992; Larkin & Simon, 1987; Mayer, 1989) An analysis of information perspective may reconcile these conflicting findings. In particular, low ability participants or novices may be able to extract structural but not functional information from diagrams. Functional information must be inferred from diagrams, in contrast to structural information, which is explicit.

Three experiments examine the interactions of medium, content, and ability in the comprehension of complex systems. We use three systems that have been used with success in previous literature, a pulley system (adapted from Hegarty & Just, 1993), car brake and bicycle pump (both adapted from Mayer & Gallini, 1990).

Experiment 1: Descriptions from Diagrams

Diagrams of complex systems are excellent for conveying structural information as they use space and the elements in it to convey real or conceptual elements and the relations among them. Adding arrows may facilitate conveying functional information as arrows indicate the temporal sequence of operations. Participants were asked to describe what is depicted in a diagram of a complex system, without and with arrows.

Method

Participants

Participants were 67 psychology students fulfilling a course requirement. Thirty-four participants described diagrams without arrows; 8 a car brake, 14 a bicycle pump, and 12 a pulley system (see Figure 1 for example of car brake diagram with arrows). Thirty-three participants described diagrams with arrows; 8 a car brake, 12 a bicycle pump, and 13 a pulley system.

Procedure

Participants were first asked to rate their mechanical ability and prior knowledge of the device given to them on a 1 to 7 scale, 1 = poor, 7 = excellent. In the 3 experiments reported here, participants self-rated their mechanical ability and their prior knowledge of the complex system presented. We chose a self-rated measure as it has been found to correlate with actual mechanical ability and spatial ability (Hegarty & Just, 1993; Heiser & Tversky, in prep).

Participants were shown one of three diagrams: a car brake (Figure 1), bicycle pump, or a pulley system, either without or with arrows and asked to "Please examine the diagram above. On the lines below, write a description of the system in the diagram."

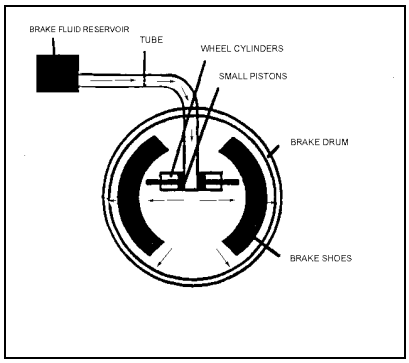


Figure 1: Car brake diagram with arrows (adapted from Mayer & Gallini (1990))

Coding

Self-rated mechanical ability and self-rated prior knowledge of system correlated highly ($r = .78$, $p < 01$). They were averaged to provide a single ability score, ranging from 1-7 (poor-excellent).

Descriptions of diagrams were coded blindly. Two coders divided each description into propositions. Coders were told that a proposition is the smallest unit of meaning in a sentence. For example, in the sentence "the liquid brake fluid travels down the tube," there are two propositions. First, "the brake fluid is liquid" and second, "the brake fluid travels down the tube."

Coders categorized each proposition as structural or functional. Descriptions of the system structure or of the features of the components (i.e. the shape of a part) counted as structural information. Descriptions of the function of the system, the function of individual parts or the way the parts work together, counted as functional information. In the previous example, the first proposition, "the brake fluid is liquid" was coded as structural. The second proposition, " the brake fluid travels down the tube" was coded as functional. Coders agreed 94%, and disagreements were settled through discussion.

Results and Discussion

There were no main effects for diagram content, self-rated ability or total number of propositions across conditions.

As predicted, participants who described diagrams with arrows produced significantly more functional units of information ($M = 2.24$, $SD = 1.3$) than participants who described diagrams without arrows ($M = 1.26$, $SD= 1.1$), $F(1,61) = 10.9$, $p < .01$. Similarly, participants who described diagrams without arrows generated significantly more structural units ($M = 1.65$, $SD = 1.65$), than those who described diagrams with arrows ($M = .52$, $SD = .62$), $F(1,61) = 13.67$, $p < .01$.

The presence of arrows in a diagram of a mechanical system indicates the sequence of operations of the system. From the temporal sequence, participants readily made inferences to the function of the system, and described it in those terms.

Experiment 2: Diagrams from Descriptions

Is the use of arrows to convey temporal, causal sequence so established that producers of diagrams will comply? Participants read either a structural or a functional description of a complex system, and produced a diagram.

Method

Participants

240 students in an introductory psychology course at Stanford University participated for course credit. Forty-four participants either did not draw a diagram or did not complete the questionnaire, leaving 93 participants in the functional description group and 103 in the structural description group, distributed fairly evenly across the three systems.

Stimuli

Structural and functional descriptions were written for each of the three systems, the car brake, bicycle pump, and pulley system. Those for the car brake appear in Table 1 and 2. Structural descriptions contain details of parts and their spatial relations, primarily using forms if the verb "to be" or verbs of fictive motion. Functional

descriptions contain actions and consequences primarily using active verbs of motion.

Table 1: Car brake structural description

"The brake or brake drum is a circular structure. Directly inside the sides of the brake drum are two thick semicircular structures called the brake shoes. The brake fluid reservoir is located above and to the side of the brake drum. From the brake fluid reservoir, a tube runs down sideways and then down to the middle of the brake drum. Extending from both sides of the tube in the middle of the brake drum are wheel cylinders surrounding small pistons. Brake fluid can move from the reservoir through the tube to the pistons. The small pistons can move outward toward the brake shoes. The brake shoes can move outward toward the brake drum."

Table 2: Car brake functional description

"From the brake fluid reservoir, brake fluid enters and travels sideways and down the tube. As the brake fluid accumulates at the bottom of the tube, pressure is exerted on the small pistons inside the wheel cylinders. This causes the pistons to push outward toward the brake drum. The outward movement of the shoes causes friction along the inside of the brake drum, slowing the rotation of the wheel."

Procedure

Participants first rated their mechanical ability on a 1 to 7 scale, 1 = poor and 7 = excellent and their specific knowledge of the depicted mechanical system on the same scale. Participants then read a description of one of three labeled systems (car brake, bicycle pump, or pulley) and were asked: "In the space provided below the description, please construct a diagram of what you think the description is trying to convey."

Coding

Self-rated mechanical ability and self-rated prior knowledge of the device were highly correlated, $r = .72$, $p < 01$. They were averaged to produce a single ability score for each participant. Diagrams were coded blindly by two coders for conventional elements, such as arrows or lines, that conveyed either structure or function. There were no disagreements in coding.

Results

As before, there were no effects of mechanical system or for self-rated ability. Of the 196 depictions coded, the only graphical element added was arrows. The arrows were to indicate direction of motion of the mechanical system. As predicted 62/93 (66.7%) who depicted functional descriptions used arrows in their depiction to indicate sequence of operation, whereas 16/103 (15.5%) participants who depicted structural descriptions included arrows, X^2 (N = 196) = 9, $p < .01$. All 16 who included arrows in depictions from structural descriptions were high mechanical ability participants (see Figure 2 for examples).

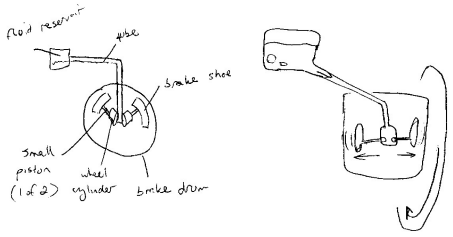


Figure 2. Drawing from structural description (left) and functional description (right) of car brake.

Discussion: Communicating With Diagrams

Experiment 1 and 2 showed that people readily interpret and produce arrows in diagrams to suggest functional properties of complex systems. For a car brake, bicycle pump and pulley system, diagrams without arrows elicited structural descriptions. Conversely, for structural descriptions participants drew diagrams without arrows but for functional descriptions they drew diagrams with arrows. Moreover, low ability participants were as likely as high ability participants to comprehend and produce arrows to convey function.

The finding that structural diagrams can be effectively enriched by the simple addition of arrows is important, as making inferences from structure to function is one of the major difficulties of understanding complex systems. The next study will examine the roles of structural and functional descriptions and diagrams with and without arrows in comprehending and making inferences about complex systems.

Experiment 3: Learning structure and function from diagrams and text

Complex systems can be described from structural or functional perspectives. The structural aspects of a system are typically easier to convey. However, it is often the functional aspects that are critical for understanding how the system operates and for trouble-shooting, error-recovery and high-level problem solving. The previous experiment showed that a simple enrichment of structural diagrams, an arrow, enable functional inferences. Here we examine directly and in detail the relative efficiency of structural and functional text and of diagrams with and without arrows in conveying structural and functional aspects of complex systems. We do this for both high and low mechanical ability participants. This experiment will provide insight into the effects of medium, text or diagram; perspective, structural or functional; and ability, high or

low, on transmission of structural and functional information about complex systems.

Method

Participants and design

Participants were 147 students in an introductory psychology course at Stanford University participating for course credit. Each participant was randomly assigned to one of 8 conditions. 31 to the no arrow diagram condition, 40 to the arrows diagram condition, 33 to the structural text condition, and 43 to the function text condition. Approximately equal proportions of the participants studied the car brake and the bicycle pump. The pulley system was not used in this study. Study time was recorded for all subjects; however, timing was inaccurate for 34 subjects (in random conditions) leaving 113 study times in the analysis.

Procedure

As in Experiments 1 and 2, participants first rated their general mechanical ability and prior knowledge of the specific device (car brake or bicycle pump) on a scale from 1 to 7, 1 = poor, 7 = excellent.

Participants studied a description or diagram of either the bike pump or car brake. In the text conditions, participants read and studied the description four times. In the diagram condition, participants studied the diagrams completely four times. Study time was self-terminating. Immediately after study, participants answered 16 true/false statements, half structural, half functional. The questions varied in difficulty. An example of a structural T/F statement is "The small pistons are adjacent to the brake shoes." An example of a functional T/F statement is "The pistons put pressure on the brake shoes." Participants were told to respond quickly and accurately.

Ability measurements

Participants' scores from the self-rated prior knowledge of device and mechanical ability scales correlated significantly ($r = .68$, $p < .01$) and were averaged to form a mechanical ability score. A median split gave low and high ability students.

Results

Does the medium, text or diagram, or perspective, structural/no arrow or functional/arrow affect performance on structural and functional questions? How does ability affect performance? Because of their natural mapping to space, it is predicted that diagrams will be superior to text for structural questions. In regards to conveying structural or functional perspectives of complex systems, it is predicted that structural descriptions or diagrams should facilitate structural questions and functional presentations should facilitate functional questions. Finally, which is easier, making inferences from structure to function or from function to structure? To assess these effects and others, we performed four analyses of variance on errors and response times for structural and functional questions. Each ANOVA had medium, text or diagrams; perspective, structural or functional; and mechanical ability, high or low, as factors**.**

Study time

There was wide variability in study time, but it did not correlate with any of the measures of interest-medium, perspective, or ability.

Learning Structural Information

Effect of Ability

High mechanical ability participants outperformed low ability participants on structural questions. Low ability participants made more errors ($M = 2.5$, $SD = 1.51$) than high ability participants ($M = 1.59$, $SD = 1.14$), $F(1, 139) = 15.7$, $p < .01$. There were no significant differences in response times between high ($M = 4.6$s, $SD = 1.5$) and low mechanical ability ($M = 4.5$s, $SD = 1.3$s). There were also no significant interactions between ability, medium, and perspective for errors on structural questions. Figure 3 illustrates that low ability participants perform close to that of high ability participants when structural text was studied, however this did not elicit a significant interaction.

Effect of Medium

There were no effects of medium for structural questions. Fewer errors were made after a diagram was studied ($M = 1.76$ out of 8, $SD = 1.08$) than after text was studied ($M = 2.28$, $SD = 1.62$), however this difference was not significant, $p > .1$ (see Figure 3). Structure was conveyed equally well by text and diagrams. Response time, however, was significantly longer on structural questions after studying a diagram ($M = 5.1$s, $SD = 1.4$s) than after studying a text ($M = 4.2$s, $SD = 1.3$s), $F(1, 131) = 13.6$, $p < .01$. This effect may be due to extra time required to translate a visual representation into a sentential representation in order to answer the verbal questions.

Effect of Perspective

There were no effects of perspective (structural or functional) on errors or response time on structural questions. Participants made similar numbers of errors on structural questions if a structural perspective was studied ($M = 1.89$, $SD = 1.39$) than if a functional perspective was studied ($M = 2.13$, $SD = 1.41$), $p > .05$ Though in this analysis, diagrams have a clear advantage because structure remains explicit in the diagram with arrows, the interaction between presentation and perspective was not significant, $p > .1$.

The finding that both high and low mechanical ability participants did not differ significantly on structural errors regardless of study perspective indicates that they were able to efficiently make inferences from function to structure.

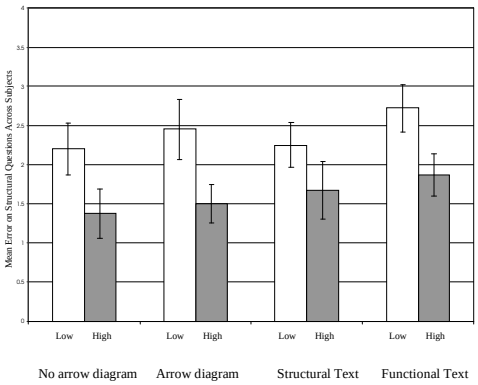


Figure 3. Errors on structural questions by ability, perspective and medium.

Learning Functional Information

Effect of Ability

For functional questions there was a main effect for ability, where high mechanical ability participants made fewer errors (M = 1.44, SD = 1.3) than low ability participants (M= 2.75 SD= 1.6), F(1, 145) = 29.6, p<.01. There were no significant differences in response times between high mechanical ability (M = 5.2s, SD = 1.9s) and low mechanical ability (M = 5.3s, SD = 1.8s), p >.1.

Mechanical ability interacted with medium. See the following section for details.

Effect of Medium

There were no overall effects of medium on errors and response times on functional questions. However, medium and perspective interacted, F(1, 139) = 8.02, p < .01. High ability participants made fewer errors on functional questions when diagrams were studied (M = 1.1, SD = 1.1) than when text was studied, whereas low ability participants made fewer errors when text was studied (M = 2.6, SD = 1.6) than when diagrams were studied (M = 3.0, SD = 1.6). This effect however, could be driven by interaction between perspective of study and medium, where participants performed extremely well if functional text was studied, but not structural text. This is further discussed in the next section.

Interestingly, high ability participants outperformed low ability participants on functional questions in all conditions except when functional text was studied (see Figure 4). These results indicate that low ability participants have difficulties making functional inferences from structural descriptions and diagrams, with or with out arrows. When functional information is presented verbally, low ability participants are no longer disadvantaged.

Effect of Perspective

There was a slight benefit for functional questions from studying functional material, however this effect was only marginally significant, F (1,139) = 3.5, p = .06. Performance was higher on functional questions after studying functional text or diagrams with arrows (M = 1.73, SD = 1.48), than after studying structural text or diagrams (M = 2.45, SD = 1.69). There were no differences in response times.

There was an interaction between perspective and medium. Errors on functional questions were higher after studying a structural text (M=3.0, SD=1.7) than after studying a diagram without arrows (M=1.87, SD=1.5), functional text (M = 1.71, SD = 1.27) or diagram with arrows (M = 1.75, SD = 1.68), F(1, 139) = 17.48, p < .01. In general, participants were better at making functional inferences from diagrams, than from structural text.

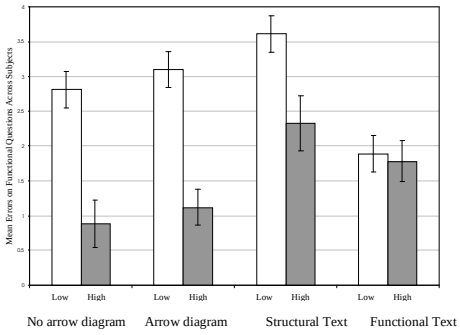


Figure 4: Errors on functional questions by ability, perspective and medium.

Experiment 3 Discussion

Structural information was effectively conveyed by well-constructed diagrams and text, from the perspective of the system's structure or function. Mechanical ability of the participant, however, is important for predicting errors on structural questions.

The results for functional information were quite different. Again, high mechanical ability participants outperformed low ability participants. This result was conditional upon presentation, however, where low ability participants performed as well as high ability participants when functional text was studied. Low ability participants were at their worst when functional inferences had to be made from diagrams.

The results from Experiment 3 help to clarify the relationship between ability and comprehension of diagrams, illustrating the importance of the information

to be conveyed. Low mechanical ability participants were able to learn structure from both diagrams and text, but needed functional text to aid learning functional information.

General discussion

Complex systems consist of structural information, a configuration of parts, and functional information, a sequence of operations and outcomes. The present research investigated the effects of medium, text and diagram, perspective of presentation, structural or functional, and ability on acquisition of complex systems.

Diagrams use elements and relations in space to convey elements and relations in real or conceptual space. Thus, diagrams are especially suited to convey structural information. Experiment 1 and 2 showed that a simple addition to diagrams, arrows, enables a static diagram to convey functional information effectively. Participants spontaneously interpreted diagrams with arrows functionally, and produced diagrams with arrows for functional descriptions. In global acquisition of complex systems, however, arrows were sufficient for participants with high mechanical ability but not for those with low mechanical ability.

In Experiment 3, participants were more adept at inferring structural information from functional than functional from structural. Apparently, function imposes more constraints on structure than structure imposes on function, in accordance with the design adage that form follows function. This means that function is not necessarily transparent from form. This fact is substantiated by the performance of low ability participants, who, in contrast to high ability participants, had trouble making functional inferences from diagrams. Similarly, Suwa and Tversky (2001) found that experienced architects were more likely to extract functional information from their sketches than novices. Low ability participants reached the level of high ability participants when the perspective of the questions matched that of the studied text. This suggests that the text guides the learner in forming a mental model of that information, especially for low ability learners. For this type of complex systems, including car brakes and bicycle pumps, the disadvantages of low ability can be overcome by the addition of explicit functional information.

These results have implications for theories of diagrammatic reasoning. The findings indicate that learners of all abilities are able to extract essential parts and their interrelations from diagrams; however the advantage of diagrams disappears when learners with low mechanical ability are asked to make inferences beyond what is conveyed explicitly in the diagram.

In addition, these results have implications for design of instructions and explanations as well as comprehension of complex systems. Instructions and texts depending solely on diagrams will be ineffective for some users, especially for functional information. Instructional illustrations of mechanical, scientific, or abstract systems such as governmental legislation need to include explanatory text. Taking into account the ability of the learner, the perspective of the information, and the medium in which it is portrayed, will dramatically increase the accuracy and amount of information that can be acquired from a portrayal of a complex system.

Acknowledgements

The first author is supported by an Eastman Kodak Inc., fellowship. The research is funded by Office of Naval Research, Grants Number NOOO14-PP-1-O649 and N000140110717 to Stanford University.

References

Hegarty, M., Carpenter, P. A. & Just, M. A. (1990). Diagrams in the comprehension of scientific text. In R. Barr, M. L. Kamil, P., Mosenthal, & P. D. Pearson (Eds.) Handbook of reading research. New York: Longman.

Hegarty, M. and Just, M.A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Larkin, J.H. & Simon, H.A. (1987) Why a diagram is (sometimes) worth a thousand words. *Cognitive Science, 11*, 65-99.

Mayer, R.E. (1989) Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 89(2)*, 240-246.

Mayer, R.E. and Gallini, J.K. (1990) When is an Illustration Worth Ten Thousand Words? *Journal of Educational Psychology*, Vol. 82 (4), 715-726.

Morrison, J. B. (2000) *Does Animation Facilitate Learning? An Evaluation of the Congruence and Equivalence Hypotheses.* Unpublished Doctoral Dissertation. Department in Psychology, Stanford University.

Morrison, J. B., and Tversky, B. (2001). The (In) effectiveness of animation in instruction. In Jacko, J. and Sears, A. (Editors), Chi 01: Extended Abstracts. Pp. 377-378. Danvers, MA: ACM.

Suwa, M.,& Tversky, B. (1997). What architects and students perceive in their sketches: A protocol analysis. *Design Studies, 18*, 385-403.

Do argumentation tasks promote conceptual change about volcanoes?

Joshua A Hemmerich (joshh@uic.edu)
Jennifer Wiley (jwiley@uic.edu)
Department of Psychology
The University of Illinois at Chicago
1007 W. Harrison Street (MC 285)
Chicago, IL 60607, USA

Abstract

In the present studies, we assessed college undergraduate research participants' models of the earth's composition and dynamics, both without and with access to a web site on plate tectonics. In previous studies, it has been found that argument writing tasks promote better understanding from web pages, with the best comprehension of texts observed when students write arguments using a two-window browser. In the present investigation, we are interested in whether or not students in this condition acquire more advanced conceptual models of the subject matter than naïve students, or students in other reading/ writing conditions.

In previous studies (Wiley & Voss, 1999; Wiley 2001) the task of writing an argument and the presentation of web pages in two side-by-side windows were found to lead to the most comparison, integration and explanation in student essays. This resulted in better understanding of the subject matter, as measured by inference and analogy tasks. Theoretically, presenting information in multiple sources as well as asking students to construct their own arguments both seem like conditions which may especially prompt active processing, and demand that readers try to develop their own models of the text. (Wiley & Voss, 1999, Perfetti, 1997; Kintsch, 1998). The present studies first investigated what Earth Science concepts students held, specifically pertaining to the causal nature of volcanic eruptions, without receiving any instruction at all. Second, we advanced a taxonomy of student concepts about volcanoes and plate tectonics, and we investigated whether manipulating the writing instruction (essay or argument), as well as the type of web interface (one window or two windows) the materials were presented in, had any effect on the quality of students' causal models of a phenomenon in Earth Science, the eruption of Mt. St. Helens

Earth Science Concepts

Students' understanding of Earth Science concepts is a historically neglected topic that has only begun to receive the necessary attention. Consequently, mental representations of the complexity of our planet and the causes of its natural phenomena is an appropriate topic for conceptual change researchers to focus on as well as an important goal for educators. By the age of 13, most children have acquired a spherical earth concept. They have developed a model of the Earth that corresponds to a planet (a huge sphere surrounded by space). Vosniadou and Brewer (1992) delineated a series of models that many children hold as they approach a mature understanding of the earth's shape. Most children will acquire a round earth concept by fifth grade. This knowledge alone is an important building block for understanding of many Earth Science concepts, however, there are still important conceptual developments in Earth models that need to occur in order for students to understand many other topics in Earth Science.

For many Earth Science topics, following the adoption of a spherical Earth model, the students need to refine their understanding of the compositional properties and surface features of that model. Specifically, students need to develop models that can account for rock cycles, mountain formation, sea floor dynamics, and geological disasters such as earthquakes and volcanoes. They need to develop models that explain geological data relating to changes in the Earth's surface. This seems to be problematic for many students as they are presented with new Earth Science information.

Generally around fifth or sixth grade, the composition and dynamics of the Earth are included in an Earth Science curriculum. Ross and Shuell (1993) found that students in grades K through 6 had many misconceptions about the causes of earthquakes. Some examples of young children's misconceptions are: that earthquakes are caused by wind or weather; that volcanoes are caused by the heat of the sun or by mountains; and that volcanoes and earthquakes can have animistic/humanistic explanations, like the earth is "upset", and that these events somehow reflect the earth's mood or temperament.

The American Geological Institute (1991) has prescribed the understanding of how the Earth's crust is moving and the Theory of Plate Tectonics as essential questions to be answered by students in grades 9-12. However, even after their first instruction on these topics, students have many misconceptions about the causes of earthquakes and volcanoes. Marques and Thompson (1997) found that sixteen and seventeen year-old Portuguese students held numerous misconceptions about the Earth's continents, magnetic field, and tectonic plate movements. For instance, some students believed that tectonic plates rotate around a

plate axis, while others believed that there is a progressive cooling of the Earth, which causes the crust to crack. Barrow and Haskins (1996) have shown that Earth Science misconceptions extend beyond grade 12 and are exhibited by college students in an introductory geology course, of which less than 7% believed their earthquake knowledge to be good or excellent. Many adults think that earthquakes can be predicted by the weather or the tides. Some still have a model of the earth's surface with continents floating on top of oceans, while others see volcanoes and earthquakes as the result of the earth building up too much pressure, heat or other internal stuff, making the Earth like a balloon or pimple. Clearly, Earth Science is a domain where mastery among children and young adults is rare and misconceptions are widespread.

In one of the few investigations that have been done on people's understanding of advanced Earth Science concepts, Gobert (2000) has found that when younger students attempt to produce causal explanations about Earth Science phenomena, they tend to demonstrate incomplete or distorted knowledge. Gobert (2000) classified student models using typologies of the interior of the Earth and the causal mechanisms of volcanic eruptions. Gobert's typology of explanations for volcanic eruption consists of type 1a models in which mechanisms are heat-related only (like the earth core gets too hot), type 1b models, which involved movement-related causal mechanisms and are void of heat-related causal concepts (like the magma rises or pushes up through the crust and causes a volcano), type 2 mixed models, which contain some elements of heat and movement causal mechanisms but are not elaborate or integrated (the inside of the earth is hot, magma pushes up), and type 3 models, which consist of multiple, well integrated, heat-related and movement-related mechanisms. At level 3 the idea that heat causes movement, and more specifically that convection currents in the earth's core cause tectonic plate movements, is important.

Gobert's typology is a useful one in analyzing and categorizing the models of Earth Science students. With some minor additions and fine-tuning it is utilized in the data analysis in the present studies.

Argumentation Tasks

It is possible that part of the problem with students understanding of Earth Science is that they don't integrate the information that they are presented with into a coherent and complete model. The achievement of this goal could be aided by encouraging students to engage in tasks that facilitate the integration of relevant concepts that are presented to them.

Past work has indicated that argument writing is a task that requires students to integrate information, particularly when it is necessary to coordinate information from different sources to make a cohesive representation of a phenomenon. Wiley and Voss (1999) found that when students were asked to write arguments about the causes of the Irish Potato Famine from multiple sources it resulted in essays with more transformation, integration, and explanation of the presented information, than when students were asked to write narratives from the same set of sources. Furthermore, students who wrote arguments were better able to identify correct inferences and underlying principles about the causes of the Potato Famine after the writing task. In comparison to students who wrote narratives from textbook chapters, students who wrote arguments from the multiple sources in a web site demonstrated a better understanding of the subject matter. Based on this evidence, Wiley and Voss (1999) concluded that tasks which require students to construct their own representation of a situation yield the most conceptual learning in web-like environments; and the argument writing task promoted understanding because it required students to integrate information from across multiple sources as they created support for a thesis. This result is consistent with other studies demonstrating that tasks that require learners to engage in active, constructive and integrative tasks lead to the best understanding of text (e.g., Chi, de Leeuw, Chiu & LaVancher, 1994; Goldman, 1997; McNamara, Kintsch, Songer, & Kintsch, 1996, Scardemelia & Bereiter, 1987) as well as studies on collaborative discourse which have found that students who engage in more argumentation-related behaviors develop a better understanding from peer discussion (Anderson, et al. 2001; Chinn, Anderson & Waggoner, 2000).

There has been little work studying how students use multiple windows, or looking at optimal conditions for multiple window use (Foss, 1989; van Oostendorp, 1996). Recently, Wiley (2001) found that when readers were given explicit instruction on how to use the browser there were some learning benefits for a two-window interface, while there was an even more resilient facilitation for argumentation task. There appears to be growing evidence that engaging in argumentation, and similar tasks, facilitates conceptual learning and integration of new material.

Present Studies

In the present studies, we assessed undergraduates' models of the Earth's composition and dynamics. In the first study we asked undergraduates for their understanding of what causes volcanic eruptions without providing them with any reading material. In a second study, we tested whether undergraduates would display more mature models after engaging in argumentation tasks. Undergraduates were asked to read documents from a web site about earthquakes and volcanoes either with the general instruction to learn the information so that they could write an essay about what caused the eruption of Mt St Helens, or the specific instruction that they should read the site in order to write an argument of what caused the eruption of Mt St Helens. In addition, students either read the information presented in a single-window or two-window browser. In general, past

work has found that both the two-window design of the browser as well as the argument writing task are responsible for promoting understanding, with the best comprehension of the text observed when students write arguments using a two-window browser. In the present investigation, we are interested in whether or not students in this condition acquire more advanced conceptual models of the subject matter.

An accurate understanding of the nature of the eruption of Mt. St. Helens would entail the following information: Mt. St. Helens is a subduction zone volcano, which means that it is located on a tectonic plate boundary and not on a hotspot. The Earth's tectonic plates are known to move, due to convection currents in the Earth's liquid layers. The plates that lie underneath Mt. St. Helens pushed together, or converge, leading to subduction. Consequently, this subduction (one tectonic plate sliding underneath the other) causes solid mantle from the bottom plate to be pushed down to areas of higher temperature. This solid mantle melts in the high temperature and became viscous liquid magma. Viscous magma builds up and causes an increase in pressure, which is not released until the magma shifts and an eruption occurs.

Study 1

Method

Participants. 28 undergraduates at the University of Illinois at Chicago participated in this experiment.

Procedure. The participants were asked to answer the question "What caused the eruption of Mt. St. Helens on May 18, 1980?". Students were asked to write at least a paragraph.

Measures. Student concepts were assessed by coding answers for the kind of models that students had of how volcanic eruptions happen. The coding scheme was originally based on Gobert's (2000) typology, but several categories needed to be added or amended to account for the models we observed in our protocols. The different levels of our typology are described below.

Student Models of Volcanic Eruptions

It should be noted that some models are not necessarily incorrect explanations of volcanic eruptions per se, because they could account for certain types of volcanoes. But many are not sufficient explanations of why Mt. St. Helens, a stratovolcano, erupted as it did.

Type 0 Incorrect, Superficial Models

Students were assigned a 0 if their explanation of the cause of volcanoes was related to an irrelevant surface feature of the earth. Examples of explanations at Level 0 are that volcanoes are caused by surface conditions, such as wind, avalanches, landslides, mountains, weather, sun, the orbit of planets, tides, faults, time, dormancy or too much lava, as well as non-explanations. Essays that did not include any of the major causal agents identified below received a 0.

Type 1 Local Models

Models that mentioned single, local causes of movement or heat, as in Gobert's (2000) typology, or the concept *pressure* were assigned a '1'. Models were given this rating if they expressed the idea of one of these three as being the causal agent in the eruption of the volcano. The addition of the concept *pressure* as a type 1 causal agent was made because this concept is a separate notion from heat or movement and is relevant to the eruption of a stratovolcano, such as Mt. St. Helens, in which no gas escapes from the volcano before a violent eruption. A second amendment from Gobert's typology was splitting the movement category into two separate categories, one specifically related to magma or lava movement, and the second related to plate movement. After proposing this coding analysis, none of the students had an explanation related to magma movement alone, so a single movement category was retained.

Explanations of Type 1A tended to mention hot, melting or molten magma, the temperature of the magma, and the heat of the earth's core. Explanations of Type 1B mentioned the movement, shifting, colliding, rubbing or interacting of plates. Explanations of Type 1C tended to mention that the volcano or Earth was full of gas, the magma had too much gas, that there was pressure or that the magma was being kept under force.

Type 2 Mixed Models

Models that included plate movement with either heat, pressure, force or chemical processes were assigned a '2'. In these models, multiple factors were mentioned but not causally related.

Type 3 Integrated Models

Only models that causally related heat or pressure and movement in either direction (i.e. convection currents cause plate movement; or plate movement causes plates to subduct and melt, forming magma that rises under volcanoes) were coded as level 3 models.

Examples of explanations included in the naïve student models along with frequency of occurrence are included in Table 1.

Table 1: Frequency of Naïve Models with Examples

Model	Examples	Frequency (%)
Type 0	I assume that the eruption was due to a geological disturbance such as a sudden misalignment of orbits.	7 (25%)
Type 1A	The eruption of Mt. St. Helens was caused by the heat build up	3 (10.71%)

	in the earth's core…	
Type 1B	The eruption of the volcano was caused by a sudden shifting in the earth's tectonic plates. This shifting caused a disruption of the mountain…	8 (28.57%)
Type 1C	The eruption of Mt. St. Helens occurred because there was an enormous amount of pressure on the volcano… It couldn't keep the lava in, and erupted.	7 (25%)
Type 2	The eruption of Mt. St. Helens was caused by movement in the plates.... The lava is heated to the point where it has to escape.	2 (7.14%)
Type 3	Volcanic eruptions are the result of the earth's tectonic plates shifting below the surface. As the plates move past each other, friction builds up and hot magma forms. Once the plates are pushed to a certain point, the magma is forced up through volcanoes.	1 (3.57%)

Implications

A surprising number of undergraduates have incomplete or incorrect models of why Mt. St. Helens erupted. Consistent with anecdotal reports, there are a number of persistent misconceptions about why volcanic eruptions and earthquakes occur. In study 2, we address the extent to which students may undergo conceptual development as they construct arguments from evidence presented in a web site.

Study 2

Method

Participants. 40 undergraduates at the University of Illinois at Chicago participated in this experiment.

Design. There were two manipulations in this experiment. The first manipulation was in the instructions that students were given. Twenty of the students were asked to read the web site in order to "write an essay of what caused the explosion of Mt St Helens in 1980", while the other twenty were given the exact same instructions except the word *essay* was replaced by the word *argument*.

The second manipulation was in the format of the browser in which the web site was presented. In this experiment, students either read the documents from a single-window browser, meaning they chose documents from a document list that was presented at the start of the experiment, and viewed the documents one at a time. The other half of the students were given a two-window browser, but they also got specific instructions about why they were being given two windows "in order to compare across documents". Further in this condition the list of documents was split in half, so that in order to read all of the information readers had to use both windows. All students received explicit instruction on how to use the browser environment.

In each of the two presentation conditions, half the students received an essay writing instruction while half received the argument writing instruction. This yielded a 2x2 (writing task x browser format) design with 10 students in each of the four conditions.

Materials. The contents of the page were taken from the USGS web page**.** Pictures and diagrams were presented with captions, but in their own windows (as documents). There were no hyperlinks between documents other than navigational links back to the overview lists, and between the overview list and the documents.

Procedure. The participants were asked to read documents from a web page about Geological Hazards in order to write either an essay or an argument. All participants were given 30 minutes to read the documents and write their essays.

Measures. Students' concepts were assessed by evaluating the quality of the essays. We coded essays using the coding scheme developed in Study 1. Two raters independently coded each essay, blind to condition. Inter-rater reliability was above .90. Discrepancies were resolved through conversation.

Additionally a demographics questionnaire was administered at the end of the experimental session that included the question, "How much did you know about Mt. St. Helens and its relation to plate tectonics before reading this site?" Participants answered this question on a scale of 1-10, with 1 meaning "not much" and 10 meaning "a lot".

Examples of Student Explanations

The following examples are excerpts of the participants' written essays. Two examples of each category are provided. Only the portions of the essays containing relevant ideas are included.

Type 0 Incorrect, Superficial Model

… the climate has a dramatic effect on volcanoes.

Type 1A Local Heat Model

… in certain locations around the world volcanism has been active for a long time which means there are a hot spots under the plates which are exceptionally hot regions that provide localized high heat energy to use.

…Below some plates there are hot regions which give off high heat energy, thus sustaining volcanism.

Type 1B Local Movement Model

…The earth is built around a dozen plates...As the plates move, it causes the plates to rub against each other, causing

the explosion of the volcano.
…In the case of Mt St Helens, an oceanic-continental boundary formed. A dense plate from the ocean floor meets a less dense plate of continental land, creating the mountainous area around Mt St Helens....The material of the dense plate goes deep into the earth and eventually transforms into magma, a product of a volcano…

Type 1C Local Pressure Model
…The fierce explosion of Mt. St. Helens was due to the fact that gas was trapped inside the magma. This gas can't escape until magma enters the throat of the volcano...
…Mount St Helens violent explosion was due to great amount of silica (in the magma).... These are what stop gases from escaping at the proper time…

Type 2 Mixed Models
…There could be several reasons why Mt. St. Helens erupted. However, I believe a collision of oceanic and continental plates caused the earthquake that caused Mt. St. Helens to erupt...Eventually the Juan de Fuca plate and the North American plate, smashed into each other, causing a great disturbance underneath Mt. St. Helens volcano. The gas inside the volcano could not escape... the pressure built up inside the volcano and grew too strong and came out as one big burst.
…Three plates come into play underneath Mt. St. Helens...The movement of these plates and the added build up of pressure cause a seismic zone to form under Mt. St. Helens...

Type 3 Integrated Models
What produced the explosion of Mt. St. Helens? The explosion could have been caused by the collision of oceanic and continental plates... As the subducting oceanic crust melts within the asthenosphere the new magma rises to the top of the surface and forms volcanoes. Shallow earthquakes are associated with high mountain ranges when intense compression is occurring. Most volcanic eruptions occur near plate boundaries.
The eruption of Mt. St. Helens was caused by the unsettled magma and gas pressure. As plates meet, the denser heavier plate will be forced to sink below the lighter plate. As it moves below, magma is formed as extremely high temperatures below the mantle melt the plate. Gas and magma flow to the surface, pushing until mountains and volcanoes, which will eventually erupt, are formed. They erupt due to this pressure...

Distribution of Models across conditions
The distribution of models by reading, writing and window condition are presented in Table 2. A 2x2 (writing condition x window) ANOVA was conducted on the groups' numerical self-ratings of previous Mt. St. Helens knowledge and there were no significant differences across the experimental groups, ($F<1$).

Table 2: Frequency of Models by Writing and Window Condition

Model level	0	1	2	3
Narrative				
1 window	3	6	1	0
2 window	1	5	2	2
Argument				
1 window	1	3	5	1
2 window	1	3	4	2
Total	6	17	12	5
Reading	15%	42.5%	30%	12.5%
No Reading	25%	65%	7%	4%

Chi square analysis on the frequency of models by writing condition indicated that models were not evenly distributed. Narrative writers had more models at levels 0 and 1 versus 2 and 3, while argument writers had more models at levels 2 and 3 if anything ($X^2(1)=4.44$, $p<.03$). (An overall chi square analysis on the eight cells was not possible due to low cell size). No effects of number of windows were seen in a chi square on the frequency of models by the windows condition, ($X^2<1$).

Implications:
The results of this study indicate that young adults have incomplete models of the Earth's composition and dynamics as indicated by their observed models of the eruption of Mt. St. Helens. Although several students were able to exhibit some understanding of volcanic eruptions in general, many of their models did not show any understanding of the importance of the Theory of Plate Tectonics or that there are different kinds of volcanoes. Generally, students could not accurately describe Mt. St. Helens as a subduction zone volcano, even after reading several documents about the topic that contained all of the necessary information.

Consistent with Wiley (2001) it was found that argument writing did facilitate better understanding and model building from scientific electronic text, while providing a two-window browser only showed a beneficial trend. These results, based on a small number of participants, further suggest that conceptual development, in a domain such as Earth Science, is aided by tasks that encourage integration. But the degree of learning (and the fact that not all students achieved an understanding of plate tectonics and volcanic eruption in this condition) suggests that there may be other pieces of the puzzle needed to advance students beyond their misconceptions in this domain.

It may be that in areas such as Earth Science, where students lack concrete experience with observing and dissecting planets, some concepts are particularly difficult to learn about from text. Additionally Gobert (2000) asserts that plate tectonics concepts are difficult for children to learn due to the large size scales of the agents involved, and the

extremely long temporal scales that extend far beyond a human lifetime. Students may also find it difficult to integrate and visualize these concepts in order to understand the structure and behavior of the planet. Based on the presented studies this holds true for young adults as well. Mastery of this domain essentially requires understanding dynamic spatial information, which may make images, animations and simulations quite important. Although in general, evidence for the beneficial effects of visual adjuncts on learning is mixed (Wiley, in press; Wiley & Hemmerich, in press), for the mastery of these concepts such adjuncts may be critical. We are interested in pursuing this hypothesis, and its effects on long-term learning, in future studies.

Acknowledgments

This research was made possible through the support of the Office of Naval Research and the National Science Foundation by grants to the second author. The authors thank Rebecca Schrader for her assistance in running the experiments.

References

American Geological Institute. (1991). Earth science content guidelines grades K-12. . Alexandria, Virginia: Author.

Anderson, R. C., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S., Renitskaya, A., Tilmanns, M. & Gilbert, L. (2001) The snowball phenomenon: Spread of ways of talking and ways of thinking across groups of children. *Cognition & Instruction, 19*, 1-46.

Barrow, L., & Haskins, S. (1996). Earthquake knowledge and experiences of introductory geology students. Journal of College Science Teaching, 26(2), 143-146.

Chi, M., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves learning. Cognitive Science, 18, 439-478.

Chinn, C. A., Anderson, R. C., & Waggoner, M.A. (2000) Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly, 36*, 378-411.

Foss, C. (1989). Detecting lost users: Empirical studies on browsing hypertext. (INRIA Tech Report 973). Valbonne, France: L'Institut National de Recherche en Informantique et en Automantique.

Gobert, J. D. (2000). A typology of causal models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22*, 937-977.

Goldman, S. (1997). Learning from Text: Reflections on 20 Years of Research and Suggestions for New Directions of Inquiry. Discourse Processes, 23, 357-398.

Kintsch, W. (1998). Comprehension: A paradigm for comprehension. Cambridge: Cambridge University Press.

Marques, L., & Thompson, D. (1997). Misconceptions and conceptual changes concerning continental drift and plate tectonics among portugese students aged 16-17. Research in Science & Technological Education, 15(2), 195-222.

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Cognition and Instruction, 14, 1-43.

Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. Discourse Processes, 23, 337-355.

Ross, K. E. K., & Shuell, T. J. (1993). Children's beliefs about earthquakes. Science Education, 77(2), 191-205.

Scardamalia, M., & Bereiter, C. (1992). Text-based and knowledge-based questioning by children. *Cognition and Instruction, 9(3)*,177-199.

van Oostendorp, H. (1996) Studying and annotating electronic text. In J.F. Rouet, et al (Eds.) Hypertext & Cognition. Mahwah, NJ: Erlbaum.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the Earth: A study of conceptual change in childhood. Cognitive Psychology, 24, 535-585.

Wiley, J. (2001) Supporting understanding through task and browser design. Proceedings from the Twenty-Third Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Wiley, J. (in press) Cognitive implications of visually-rich media. To appear in M. Hocks and M. Kendrick (Eds.) Eloquent images: Writing visually in new media. MIT Press.

Wiley, J. & Hemmerich, J. (in press) Literacy: Learning from Multimedia Sources. To appear in the Encyclopedia of Education.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology*, 91(2), 301-311.

Predicting Agent Spatial Information: A Comparison Between Neural Networks and Dead Reckoning Algorithms

Amy E. Henninger (amy@soartech.com)
Soar Technology, Inc. 3361 Rouse Road Suite 240
Orlando, FL 32826

Avelino J. Gonzalez (ajg@isl.engr.ucf.edu)
School of Electrical Engineering and Comupter Science
University of Central Florida, Orlando, FL 32816

Douglas A. Reece (reeced@saic.com)
SAIC 12479 Research Parkway
Orlando, Florida 32817

Abstract

In tune with the 24th Annual Cognitive Science Conference's emphasis on application, this paper presents an empirical comparison between two methods used in agent tracking. The need to predict an agent's intents or future actions has been well documented in multi-agent system's literature and has been motivated by both systematically-practical and psychologically-principled concerns. However, little effort has focused on the comparison of predictive modeling techniques. This paper compares the performance of two predictive models both developed for the same, well-defined modeling task. Specifically, this paper compares the performance of a neural network based model and dead-reckoning model, both used to predict an agent's trajectory and position. After introducing the background and motivation for the research, this paper reviews the form of the dead-reckoning algorithms, the architecture and training algorithms of the neural networks, the integration of the models into a large-scale simulation environment, and the means by which the performance measures are generated. Quantitative measures from our experiments indicate that, for the task considered, the neural network based model provides greater predictive utility, but at an increased cost in processing time. Performance measures are presented over increasing levels of error tolerance.

Introduction

Intelligent agents typically operate in an environment populated by other intelligent agents. Agents may help each other, hinder each other, get in each other's way, or ignore each other, often without directly communicating their intent. In order for an agent to achieve its goals, it is thus sometimes necessary for the agent to determine where the other agents are, what they are doing, and what their plans are. For example, an agent may want to infer what plan an opponent is executing so that the agent can select countermoves. Han and Veloso (1995), Rao (1994), Rao and Georgeff (1995), Tambe and Rosenbloom (1995), and Tambe (1996) have studied various forms of recognizing an agent's intents.

Sometimes it is necessary to infer facts that are normally observable, such as agent location, because of sensor or other limitations. For example, a pilot agent may need to predict where a threat aircraft is flying after it enters a cloud. There are many approaches to predicting agent trajectories, including Newtonian mechanics (Lin and Ng, 1993), neural networks (Kim et al, 1999), Hidden Markov Models (Washington, 1998) and others. This paper addresses a particular application of trajectory prediction in distributed simulation and compares the effectiveness of a neural network to a commonly used Newtonian approach for this application.

The remainder of this section defines the trajectory estimation problem in the distributed simulation application and describes a previous use of neural networks for estimating agent trajectory in a visual scanning application. The paper then describes a neural network approach for trajectory estimation in distributed simulation and presents results and comparisons with Newtonian dead reckoning.

Dead Reckoning in Distributed Simulation

In a Distributed Interactive Simulation (DIS) (DIS Steering Committee, 1994), simulation software for each agent runs independently of other agents and broadcasts the ground truth about the state of the agent through network packets known as protocol data units (PDUs). Each simulation in DIS uses trajectory estimation so that the state of the agents does not have to be broadcast frequently. Lin and Ng

(1993) explain how dead-reckoning can be used to maintain coherence among entities' states in a DIS environment. Each simulator uses Newtonian equations of motion such as equation 1

$$p = p_0 + (v_0 * \Delta t) + \frac{a_0 * (\Delta t)^2}{2}$$
$$v = v_0 + a_0(\Delta t) \qquad (1)$$

where p = current position
p_0 = initial position
v = current velocity
v_0 = initial velocity
a_0 = initial acceleration
Δt = elapsed time

to predict the trajectory of other agents. Each simulator also uses the same equation to model the trajectory of its own agent; the output of this equation can be compared to the output of the true dynamics model for the agent to determine when the models diverge. When, and only when, the error between models reaches a certain threshold, the simulator broadcasts new state information for its agent. Figure 1 shows this process in a DIS simulation called ModSAF (Calder et al, 1993) that we used for our experiments.

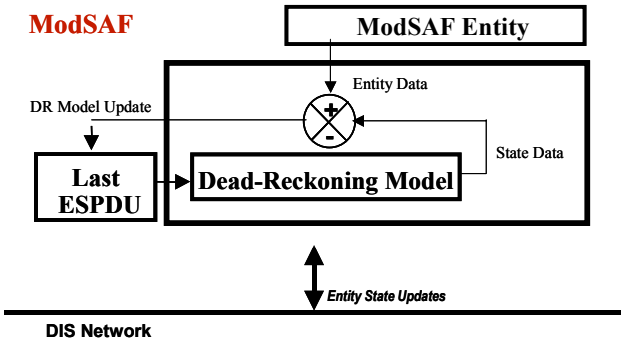


Figure 1. Dead-Reckoning Implementation in ModSAF

Figure 2 shows how at a series of time steps, the true position of an agent computed by the agent dynamics model (shown by the curve) deviates from a linear dead

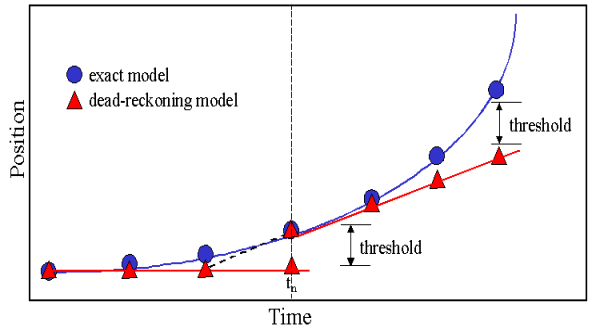


Figure 2. DIS Dead-Reckoning Process

reckoning model. When the error exceeds the threshold, the models are brought into correspondence by the issuance of an entity state PDU (ESPDU). Thus in the figure, only 3 ESPDUs are broadcast instead of one at every time step. The goal of the research presented here is to reduce the number of ESPDUs sent in by DIS simulations below the number needed using Newtonian dead reckoning.

Neural Networks for Trajectory Estimation

Kim et al (1999) developed a system to generate short-term predictions of an agent's trajectory such that it can be used to predict the agent's position at any future instance, given some window of time. They use this model as part of a helicopter agent's perceptual system to enhance the agent's ability to visiually track ground vehicles, and their motivation for this model is both psychologically and practically rooted. Psychologically, this model can be used to simulate a helicopter pilot's gaze shifting as he attempts to visually track and reaquire multiple targets. Thus, instead of operating in a state of omniscience, the agent is required to juggle the act of determining spatial information across multiple agents, as would be the human helicopter pilot. The functional ramification of this approach is that the total number of perceptual inputs to the agent is reduced at any given instance. In other words, instead of getting continuous perceptual information on all of the ground entities within the helicopter agent's field of view, by using this predictive model, the agent only requires updated information on entities when its attention is focused on those entities.

The high level architecture of this system is presented in Figure 3. The agent architecture is embedded in the

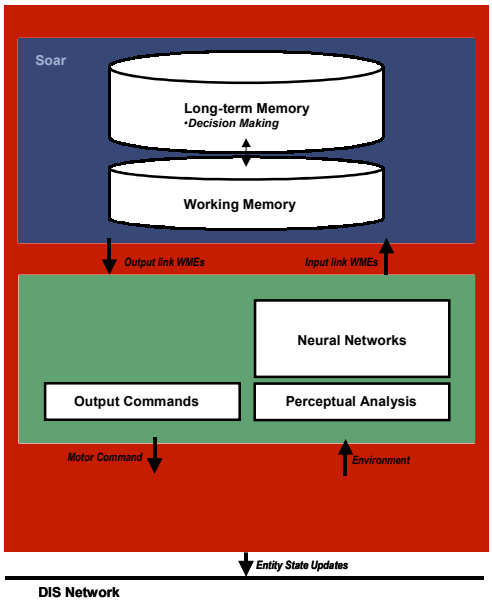


Figure 3. Visual attention for helicopter agent

Modular Semi-automated Forces (ModSAF) simulator, a system used by the military for training and research. ModSAF is elaborated on in section 2, "Methodology". The agent's intelligence is modeled in Soar (Rosenbloom et al, 1993; Newell, 1990). As a model of natural intelligence, the Soar software architecture combines the abilities to react immediately to situations, use knowledge in deliberative decision making, step back from the immediate situation to perform various forms of problem solving and planning, and learn from experience. As an indicator of the maturity and sophistication of Soar-based agents, the system has been used successfully as the production model in a number of large-scale military exercises (Hill et al, 1997; Jones et al, 1999; Nielsen et al, 2000).

The inputs to the neural networks developed for this application consist of entity data (e.g., call-sign, sim-time, position, velocity, etc.) and abstracted terrain information germane to both "on-roads" and "cross-country" travel and correlated to the entity's visual field (hill, road, lake, etc). All together, the input vector consists of 196 fields and the output vector consists of 15 output fields corresponding to discretized changes in heading ranging from –35° to 35°. The selected heading change, coupled with an assumed constant speed and "delta" time since last prediction, can be used to predict the entity's expected location at some time, **t**. With this prediction, the virtual helicopter pilot is able to look away from the ground entity for up to 7 seconds, within some error threshold.

Methodology

This paper seeks to compare the performance of a neural-network based model with the dead-reckoning model. Like both systems described in sections on dead-reckoning and neural networks for trajectory estimation, this experiment is implemented in ModSAF, a training and research system developed by the Army's Simulation, Training, and Instrumentation Command (STRICOM). ModSAF provides a set of software modules for constructing computer-generated force behaviors at the company level and below. Typically, ModSAF models are employed to represent individual soldiers or vehicles and their coordination into orderly-moving squads and platoons; but, their tactical actions as units are planned and executed by a human controller. The human behaviors represented in ModSAF include move, shoot, sense, communicate, tactics, and situation awareness. The authoritative sources of these behaviors are subject matter experts and doctrine provided by the Army Training and Doctrine Command (TRADOC). ModSAF uses state transition constructs inspired by finite state machines (FSMs) to represent the behavior and functionality of a process for a pre-defined number of states.

The scenario used for the comparison was a road-march for a tank entity 45-segment route shown in Figure 4 It is approximately 7 kilometers long and takes a tank entity about 15 minutes of simulation time to travel at a March Order speed of 8 m/s. From this 15 minute period, a total of 13760 movement updates were performed, generated at a rate of 15 HZ.

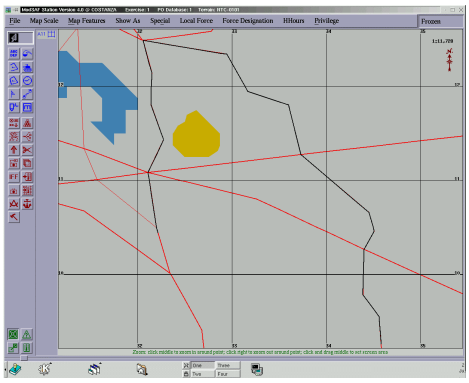


Figure 4. Route Used for Experiment

For this application, a feed-forward architecture developed with back-propagation training was used to develop the neural networks. One of these networks predicts the change in an entity's speed and the second network predicts the change in the entity's heading. Each network used a sigmoid function at the hidden nodes and a linear transformation at the output nodes. The configuration of the networks in each of the models may be seen in Table 1

Table 1. Neural Network Architecture

Model	Arch	Predictors	Resp
Speed	8-20-5-1	$Ra_{t-1}, Rb_{t-1}, Rc_{t-1}, Rp_{t-1}, Rs_{t-1}, HRab_{t-1}, HRbc_{t-1}, Hz_{t-1}$	ΔS_t
Heading	7-20-5-1	$Ra_{t-1}, Rb_{t-1}, Rc_{t-1}, Rp_{t-1}, Rs_{t-1}, HRab_{t-1}, HRbc_{t-1}$	$\Delta\theta_t$

where the inputs were normalized according to equations 2 – 19 below. Fundamentally, the inputs for each of the networks were a function of the entity's state at the last simulation clock and how this state related to the road characteristics (width, heading, length of segment, etc) and March Order parameters (speed, end-point, etc). The specific predictors are expressed in 4 – 10, and the parameters making up those inputs are explained in 11 – 19.

$$S_t = S_{t-1} + \Delta S_t \qquad (2)$$

where $\Delta S_t = f(Ra_{t-1}, Rb_{t-1}, Rc_{t-1}, Rp_{t-1}, Rs_{t-1}, HRab_{t-1}, HRbc_{t-1}, Hz_{t-1})$

$$\theta_t = \theta_{t-1} + \Delta\theta_t \quad (3)$$
where $\Delta\theta_t = f(Ra_{t-1}, Rb_{t-1}, Rc_{t-1}, Rp_{t-1}, Rs_{t-1}, HRab_{t-1}, HRbc_{t-1})$

$$Ra_t = S_t/(Da_t + M) \quad (4)$$

$$Rb_t = S_t/(Db_t + M) \quad (5)$$

$$Rc_t = S_t/(Dc_t + M) \quad (6)$$

$$Rp_t = S_t P_t / M \quad (7)$$

$$Rs_t = S_t / M \quad (8)$$

$$HRab_t = Hab_r \times Hxy_t \quad (9)$$

$$HRbc_t = Hbc_r \times Hxy_t \quad (10)$$

S_t = entity speed at t (11)

Da_t = distance to previous waypoint (12)

Db_t = distance to current waypoint (13)

Dc_t = distance to next waypoint (14)

M = march order speed (15)

P_t = perpindicular distance to road (16)

Hab_t = direction of road segment ab (17)

Hbc_t = direction of road segment bc (18)

Hxy_t = entity orientation (19)

Of these, 860 examples were used for training the speed network, 860 examples for training the heading network, and 859 examples were used for validating the training of both of these networks. The training rate was selected as 0.01 and the initial momentum parameter was .9. The momentum parameter was periodically adjusted to speed the rate of descent along the error surface. The training and validation results for each of the networks may be seen in Table 2.

Table 2. Training and Validation Errors

	Delta Speed ΔS Error(m/s)	Delta Heading $\Delta\theta$ Error(rads)
Training	0.259977±2.04558	0.004578±0.00781
Validation	0.206374±0.82532	0.014221±0.06766

Experimental Results

The neural network models were implemented in such a way that their performance for predicting entity location could be compared with the dead-reckoning model. This implementation is presented in Figure 5.

There are two ways of generating an error in our system. The first is according to the enity's location. This error is measured in terms of comparing the entity's dead-reckoned XYZ with the entity's true XYZ and is proportioned according to the width of the entity along its X, Y, and Z axes. For example, an M1A2 tank is 3.56m in width (defined along X-axis of tank), 7.34m in length (defined along Y-axis of tank), and 2.33m in height (defined along Z-axis of tank). A typical threshold for dead-reckoning error tolerance in DIS is 10% of the vehicle's dimensions. In this case then, the error tolerance for this entity's location would translate into .356m along the X-axis, .734m along the Y-axis of the tank, and .233m along the Z-axis of the tank.

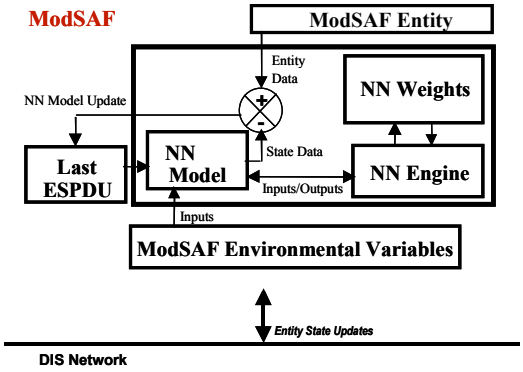


Figure 5. Neural Network Implementation used for Experiments in ModSAF

The second way of determining an update threshold is with respect to the orientation of the vehicle. In this case, the components of the dead-reckoned euler angles are compared with the components of the entity's true orientation. For tracked ground entities in ModSAF, this measure is defaulted at 3°. That is, if the dead-reckoned prediction is more than 3° off about X-axis, Y-axis, or Z-axis, an error is generated. Overall, at these error tolerances, the number of updates (ESPDUs) required by the dead-reckoning model was 351. Using these same error thresholds, the neural network models required 263 updates. Thus, the neural networks required 25% fewer updates than the dead-reckoning models. This information is presented in Figure 6 according to type of update. In the DR case, a small number of the required updates occurred simultaneously between location and rotation.

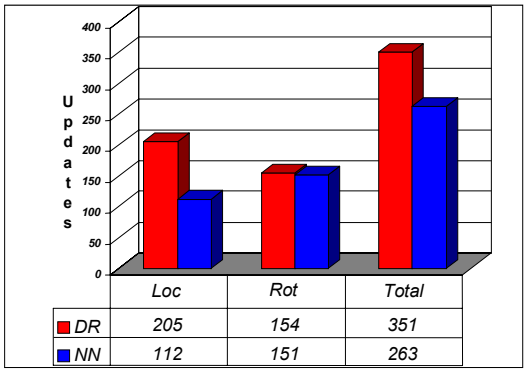


Figure 6: Number of Updates Required by DR Model versus NN Model

Although the neural network based model was able to predict the entity's path with more accuracy, as evidenced in Table 3, this increase in predictive utility comes at a cost in processing time.

Table 3. Execution Time Using Neural Networks and Dead-Reckoning Equations

	Processing Time (in 10^{-5} seconds)			
	NN Speed	NN Heading	Total NN	DR
Min	6.89029	6.19888	13.08981	2.2888
Mean	7.77692	7.47008	15.24700	2.7974
Max	20.5993	29.7999	50.3992	6.35981

Using the UNIX "gettimeofday" function, the processing speed was calculated on a Pentium III, 500 Mhz machine, running RedHat Linux 6.2. As shown in Table 3, the neural network based model required, on average, about 6 times more processing time than did the dead-reckoning based model. As stated in Section 2, the simulation time was approximately 15 minutes. On average then, the dead reckoning model produced about 23 updates per minute or rather, 1 update every 2.5 seconds (at a threshold of .356m in the X direction and .734m in the Y direction). Alternatively, the neural network based model required about 17 updates per minute, or approximately 1 update every 3.5 seconds (at the same thresholds). Coupling this information with the information on processing time tradeoffs, it becomes clear that for applications where processing time is at a premium, the use of dead reckoning-models may be preferred, in spite of their poorer predictive performance.

To further examine the relationship between the predictive power of the dead-reckoning models and this set of neural network models, we conducted experiments over a range of error tolerances. So, whereas the initial results, reported in Figure 6, were measured according to DIS default values for a tank entity (i.e., .356m, .734m, and 3°), follow-on tests incremented these error thresholds by those exact amounts. Results are presented in Table 4 and reported only by total number of required updates.

Table 4. Updates Required Over Increasing Error Thresholds

Factor of	Error Threshold			Updates Required	
	X-axis (m)	Y-axis (m)	All-axes (deg)	NN	DR
1	.356	.734	3	263	351
2	.712	1.468	6	193	237
3	1.068	2.202	9	157	188
4	1.424	2.936	12	138	156
5	1.78	3.67	15	119	137
7	2.492	5.138	21	98	109
9	3.204	6.606	27	88	92
11	3.916	8.074	33	70	79
13	4.628	9.542	39	69	75
15	5.34	11.01	45	62	73
20	7.12	14.68	60	51	60
25	8.9	18.35	75	48	55
30	10.68	22.02	90	44	48

As evidenced in Table 4, as the error tolerance increases, the predictive advantage that neural networks have over dead- reckoning models becomes less significant for this modeling task.

Summary and Conclusions

As one might expect, the choice of tool must be driven by the modeling constraints. The results reported above suggest heuristics for when to apply which modeling technique. For example, in an application where processing time is not the primary constraint e.g., multi-agent systems communicating over a wireless network, then the increased processing costs incurred from using a neural network may be defendable. Alternatively, in an application where processing time is a limiting factor, then dead-reckoning models may be the more prudent approach. It is interesting to note, also, that the differences in predictive utility of the two modeling approaches becomes less prominent as the error threshold is increased. This speaks to the power of dead-reckoning models to generalize and scale.

It is important to recognize, of course, that the modeling task in this research is limited in scope. Also, a different neural network could have yielded different results. We can not claim that this is the best network architecture or configuration for this specific modeling task. We can only claim that it was one of the more promising configurations with which we experimented. Other configurations may be better. One approach Henninger et al (1999) found particularly effective in improving the neural network based model's performance was to work with modularized models. This approach has been advocated in the control literature (Murray-Smith and Johansen, 1997; Narendra et al, 1995) and robotics literature (Brooks, 1986), and we have started exploring this approach. One of the benefits of adopting this approach is the ability to mix different modeling techniques as they best apply to the problem locally. For example, combinations of architectures and/or algorithms that can be applied to individual sub-problems, make it possible to exploit specialist capabilities. In the problem discussed in this paper, one interesting test would be to use dead-reckoning algorithms in straight parts of the road data base and then use neural networks to guide the turn, as this appears to be where the majority of updates are required.

Acknowledgements

This work was sponsored by the U.S. Army Simulation, Training, and Instrumentation Command, contract N61339-98-K-0001. That support is gratefully acknowledged.

References

Brooks, R.A. 1986. A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation, RA-2, 14-23.

Calder, R.B., Smith, J. E., Courtemanche, A.J., Mar, J.M., and Ceranowicz, A. (1993). ModSAF Behavior Simulation and Control. In Proceedings of the 3rd Conference on Computer Generated Forces and Behavioral Representation (Orlando FL), 347-356.

DIS Steering Committee 1994. "The DIS Vision: A Map to the Future of Distributed Simulation", Technical Report, IST-ST-94-01. Institute for Simulation and Training, University of Central Florida.

Han, K. and Veloso, M. 1995. Automated robot behavior recognition applied to robot soccer. Sixteenth International Joint Conference on Artificial Intelligence. Workshop on Team Behaviour and Plan Recognition, 53-64.

Henninger, A., Gonzalez, A., and Georgiopoulos, M. 1999. Modeling Semi-automated forces with neural networks: Performance improvement through a modular approach. Proceedings 9th Conference on Computer Generated Forces and Behavioral Representation, (Orlando FL), 261-268.

Hill, R. W., Chen, J., Gratch, J., Rosenbloom, P., and Tambe, M. 1997. Intelligent agents for the synthetic battlefield: A company of rotary-wing aircraft. In Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence, 1006-1012. Menlo Park, CA: AAAI Press.

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., and Koss, F. V. 1999. Automated intelligent pilots for combat flight simulation. *AI Magazine,* 20(1): 27-41.

Kim,Y., Hill, R., and Gratch, J. 1999. How long can an agent look away from a target? Proceedings 9th Conference on Computer Generated Forces and Behavioral Representation, (Orlando FL), 35-38.

Lin, K., and Ng, H. 1993. Coordinate transformations in distributed interactive simulation (DIS). Simulation, vol. 61(5):326-331.

Murray-Smith, R., and Johansen, T.A. 1997. Multiple Model Approaches to Modelling and Control. Taylor and Francis, UK.

Narendra, K. S., Balakrishnan, J., and Ciliz, K. 1995. Adaptation and learning using multiple models, switching and tuning. IEEE Control Systems Magazine June, 37-51.

Nielsen, P., Smoot, D., Martinez, R., and Dennison, J. 2000. Participation of TacAir-Soar in Road Runner and Coyote exercises at Air Force Research Lab, Mesa, AZ. Proceedings of the 9th Conference on Computer Generated Forces and Behavioral Representation, (Orlando FL), 173-180.

Newell, A. 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.

Rao, A. 1994. Means-end plan recognition. In Proceedings of KR-94, the Fourth Internatioanl Conference on Principles of Knowledge Representation and Reasoning .

Rao, A. and Georgeff, M. 1995. BDI agents: From theory to practice, In Proceedings of the First International Conference on Multi-Agent Systems, (San Francisco CA).

Rosenbloom, P., Laird, J., and Newell, A. 1993. The Soar Papers: Research on Integrated Intelligence. MIT Press, Cambridge, MA.

Tambe, M. and Rosenbloom, P. 1995. RESC: An approach for real-time, dynamic agent tracking. In Proceedings of IJCAI.95.

Tambe, M. 1996. Tracking dynamic team activity. Proceedings of AAAI-96.

Washington, R. 1998. Markov tracking for agent coordination. In Proceedings of the Second International Conference on Autonomous Agents (Minneapolis/St. Paul MN.

Anatomy is Symmetry's Best Friend: Reflections on Modeling Baylis and Driver

John Hicks (John.Hicks@ed.ac.uk)
Division of Informatics; 2 Buccleuch Place
Edinburgh, Scotland EH8 9LW

Jon Oberlander (J.Oberlander@ed.ac.uk)
Division of Informatics; 2 Buccleuch Place
Edinburgh, Scotland EH8 9LW

Abstract

An aptitude for the detection of bilateral symmetry is a fairly prominent aspect of the human visual system. Knowledge of the reasons behind this facility is not so well established, however. Some of the behavioral data indicates that processing of symmetric and non-symmetric stimuli is undertaken in two wholly different manners (i.e. *serial* versus *parallel*). However, the interpretation of this as being due to high level cognitive preferences does not exhaust the list of possible explanations. Using a split-neural network model, we show that instead of cognitive preferences, gross morphological factors may play a large role in underwriting the ability to detect symmetry as a special case of shape perception. The earlier model is consistent with behavioral data, but Occam's razor suggests that we might prefer the newer morphological explanation.

Introduction

Bi-lateral symmetry is ubiquitous in nature. Such symmetry is related to biological morphology, fitness, and behavior throughout the animal kingdom (Dakin and Herbert, 1998). Thus it is not surprising that it has also been shown to be a highly salient property of the human visual system, implicated in many phenomena.

Symmetry is both a morphological characteristic and a perceptual benchmark. From recognition of a suitable mate to apprehension of a possible predator, symmetry plays an important role, being a "non-accidental" property. That is, it is unlikely that symmetry inheres in an image by chance, or when the actual image source is asymmetric. And although increasingly there is the view that symmetry detection is not only universal, but also fundamental, emerging from very low level simple processes (Dakin and Herbert, 1998; Sally & Gurnsey, 2001), there is not yet consensus about the mechanisms that underlie the facility. On the one hand, it seems that symmetry detection is a bottom-up effect of low-level filtering in early stages of the visual process. On the other, it appears to be a top-down preference for image distillation based on its exploitability for segmentation and part decomposition (Baylis & Driver, 1994; Latecki & Lakämper, 1999). Its utility in segmentation applications has caused some to comment that "the link between symmetry and segmentation curiously seems to be more than a coincidence" (ven Tonder & Ejima, 2000).

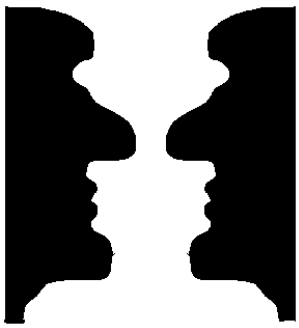


Figure 1: A familiar optical illusion whose interpretation may depend on the part-decomposition facilitated by symmetry.

Indeed, the benefit of symmetry detection for segmentation helps promote the view that it is a worthwhile thing to be good at, for segmentation is linked to part-decomposition, which in turn could be key to figure ground separation, even aiding, for example, the interchange foreground and background in a very common visual illusion (Figure 1).

This paper deals with the specific area of contour symmetry, and its effects on human visual processing, by looking at a computational model of a specific behavioral study by Baylis and Driver (1994). The field of behavioral studies on symmetry is large; it often concerns not only contour symmetry, but internal symmetry (Hicks & Monaghan, 2001) and the effects of various filtering processes. For the purposes of this paper, however, we focus on providing a computational explanation for the differences which arise in processing symmetric and repeated shapes, as seen in Figure 2.

Behavioral Studies

Part-decomposition offers a motivation for the "symmetry is special" theory, but alone it says little about the mechanisms involved. Baylis and Driver

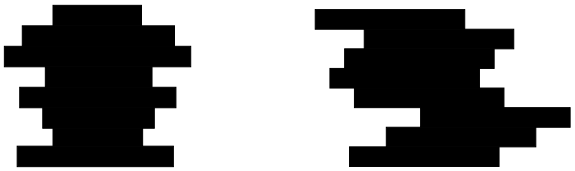


Figure 2: Stimuli: symmetric (left) and repeated (right) contours, both showing 8 discontinuities (steps) along the sides.

performed two experiments linking the perception of different shape types to distinctions in cognitive processes. In particular, the experiments aimed to elucidate the relationship between the perception of symmetry and the class of cognitive processes that are termed "parallel." In this case, "parallel" would mean that in the detection of symmetry in a two-dimensional figure, the subject *does not* engage in anything akin to a serial point-by-point comparison along the shape's contour. Baylis and Driver used a selection of perfectly symmetric shapes intermingled with shapes whose contours contained "errors" which meant a deviation from the truly symmetric form along 25% of the contour. Subjects made symmetry judgements while the experimenters varied the number of steps along the side of the shape, between 4, 8 and 16. The experimenters wanted to know whether the reaction time and error rate were significantly dependent on this variation.

It was found that symmetry was generally more quickly identified than asymmetry. This indicated directly that subjects were not involved in point-by-point search, which would always terminate earlier with erroneous examples of symmetry. Furthermore, effects of step number on subject performance were slight, and remained well within the accepted limits that define a process to be parallel.[1] Thus, the hypothesis that detection of symmetry is governed by a process impervious to increases in complexity brought about by a greater number of steps seems supported.

But what if the effects of the symmetric shapes were merely an effect of their potentially constrained nature? It might not be symmetry specifically, but redundancy in general that accounts for this semblance of parallel processing. By conducting an analogous experiment, using repetitive shapes (Fig. 2, right side), it should be possible to confirm or dismiss this confound. After all, repetitive shapes are as redundant as their symmetric counterparts, while exhibiting that redundancy via translation instead of reflection.

[1] The exception to this was when the shapes were oriented horizontally, where there was a slight effect of step number for symmetric shapes. We touch on this briefly in the discussion of our own model.

This second experiment found a significant effect for number of steps, consistent with the hypothesis that whatever process is used to judge repetition, it is effected by step count, as though it *were* a serial process. This suggests that the main difference between the two types of shapes is that in the processing of symmetry the number of discontinuities along the contour is not a significant factor, while for repeated shapes it certainly is.

Given that repetition and symmetry are equally redundant, it is clear that there must be a qualitative difference between them. The step number effect indicates a point-by-point comparison—a serial search— in the detection of repetition, which is absent from symmetry detection. But a new question arises: what is it that promotes this fast-track route for the detection of symmetry? Beneath the 'higher-level' concepts of parallel and serial processing, is there a more fundamental explanation for the fact that symmetry appears to render insignificant the relative complexity of a shape?

Modeling

This paper aims to show that this may be the case and, furthermore, that this could just as equally be the result of gross morphological aspects of anatomy as of high-level cognitive preferences. More precisely, the original assessment of the behavioral data says little about the hypothesis that what is parallel about symmetry perception is actually the 'ready-state' of the human processor to accommodate vertical symmetry.

To show this we employ a split neural network, which has previously been used as a rough correlate of the split in human visual processing, modeling reading (Shillcock & Monaghan, 2001) as well as the apprehension of the effects of symmetry in word-based stimuli (Hicks & Monaghan, 2001).

Variations in the typical architecture of neural nets often involve adding one or a number of hidden layers, vertically (Elman, 1993), or the insertion of recurrent connections. Our model employs a specific manipulation of network architecture that is not so common. Instead of devoting the entire hidden layer to the whole task upon which the network is being trained, the hidden layer can be split laterally, with each resulting half being privy to only half of the input (Shillcock et al, 2001). This split affects network performance, as shown in other modeling work. Here we apply it to the detection of symmetry in pseudo-random-block shapes.

Materials and Methods

A series of simulated neural networks, employing a back-propagation learning algorithm, were trained using sets of two-dimensional pseudo-random block shapes represented by patterns of activation. The shapes were presented to the networks through a

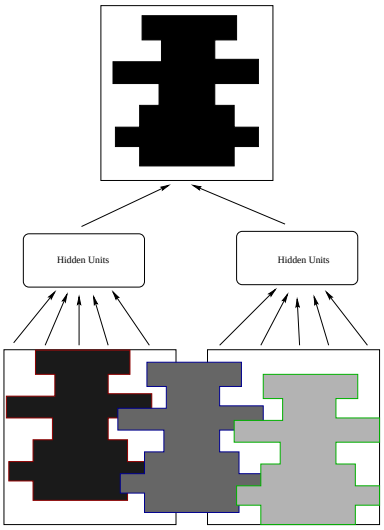


Figure 3: The split architecture network reproduces the form presented at the input, which may appear anywhere across the two visual hemi-fields.

shift invariant identity mapping (SIIM) task, maintaining the predetermined 2D block-shape of the stimuli, while moving it sequentially along the input window (Figure 3). Input nodes that fall outside the location of the block have activation zero. The vertical split in the input reflects that of the fovea and thus, as a block is repeatedly presented to the network from all possible positions across the input, it crosses from one "visual hemifield" to the other, activation being redirected to the associated hidden layer accordingly. The network is trained to recognize (represent) the shape it is being trained on.

Each stimulus set contained 60 pseudo-random block shapes, of one shape-type either all symmetric or all repetitive. For each shape type, there were three stimulus sets, with shapes having 4, 8 or 16 discontinuities along the contour (shapes with 8 discontinuities are shown in figure 2). A third class of stimulus, consisting of mixed sets, where the number of discontinuities was homogeneous, but both symmetric and repetitive shapes were represented equally, was also used in training. Due to the presentation of each pattern in all visual input positions, each stimulus accounts for 17 events in the total training set, for a total of 1020 presentation-recognition events per epoch.

After training to a predetermined number of network epochs, each net was tested with novel stimuli. The test set we focus on for the purposes of this paper contained novel blocks that were neither symmetric, nor repetitive. We were interested in seeing how the networks tended towards reproducing the type of shape they were trained on when being presented with these "random" stimuli. The metric used for gauging network performance on these test

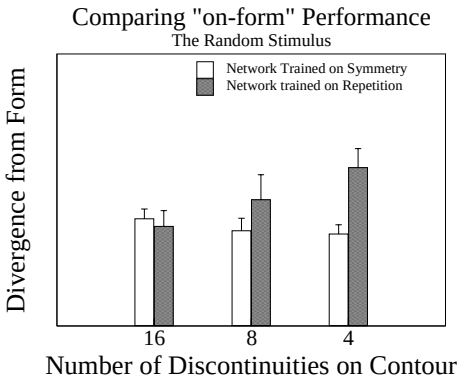


Figure 4: The compared "form" produced by trained networks, as a function of step number, under the *random* stimulus.

sets is discussed in the next section.

For all simulations the PDP++ Neural Nets software from CMU was used, running on an Ultra 5 workstation.

Results

In examining network performance on the two dimensional stimuli, we want a way of gauging the degree to which the activation at the network's output tends generally toward a given type of shape. Fortunately, the stimuli were strictly formal, in the sense that they can be defined in terms of a simple additive exemplar based function. A measure for symmetry measure is based on cancellation around the proposed axis of symmetry, while that for repetition measure but requires that activation add to a constant along arrays orthogonal to the axis of repetition (the bars that constitute the repetitive patterns are of constant length). We adopt a method of summing activations at the output so that the closer we get to the ideals, the smaller this quantity is (i.e. perfect symmetry, like perfect repetition, in the output has a form measure of zero).

Measuring form at the output

A form measure is thus available for each test shape presented to the network. I.e., for the net trained on symmetry and the test set of random (neither symmetric nor repetitive) shapes, there were 20 different weight sets for the trained net, and 20 shapes to test, giving 400 shape-weight combinations. For each we can measure both the symmetry and the repetition of the shape generated by the net's output. Note that we are not concerned with the actual error involved, but with these measures that are based on the activation levels at the output.

Thus the "on-form" measure for a net shows the tendency of its output to resemble the general shape

type with which it was trained,[2] with smaller quantities indicating greater affinity for that shape type. "Off-form" measures are also available (symmetry in the net trained on repetition, repetition in the net trained on symmetry) and were surprisingly important.

"On-Form" Measure

The general effect obtained in the model is striking. For the analysis described above, we find a significant interaction between shape type and step count when looking at the networks with respect to the type of stimulus used for training. This interaction can easily be perceived in the graphs through the much higher variance in the case of networks trained on repeated shapes (filled bars). Figure 4 shows the "on-form" analysis for the random test stimuli, comparing the degree of symmetry present in the symmetric nets, with the degree of repetition in the repetitive net. The interaction is highly significant: *F(2, 114) = 52.253, $p < .001$*).

This significant interaction between shape type and number of steps when we are using the measure appropriate for each network tells us one of two things. Either the networks, otherwise identical, have been differentially sensitized to step number by virtue of the type of shape they were trained on, or the manner in which activation is measured dictates that the quantity "tendency-toward-repetition" present in the output of the net will vary more than the the quantity "tendency-toward-symmetry," under the regime chosen to gauge it.

"Off-Form" Measure

We can clarify which of these is correct by examining the "off-form" measures. By looking at the output of the symmetrically trained net with "repetition goggles" and repetitively trained net with "symmetry goggles, " hopefully we can rule out the confound of this being a measurement effect.

In figure 5 a significant effect does obtain for the "off-form" measures of the random test stimuli ($F(2, 114) = 9.417, p < .001$). However, once more this is in the direction of repetitive nets showing more variance (by a factor of 2, upon examination of means). If the variance *was* due to the measurement of "tendency-to-repetition, " thus leading to a Type I error, we would now expect to see that variance in the symmetric when we are measuring *it* for its "tendency-to-repetition. " For the net trained on repetition, the "off-form" measure shows how symmetric its output is. The sustained effect in the net suggests a general sensitivity to step number,

[2]I.e. symmetric for the network trained on symmetry and repetition for the network trained on repeated shapes

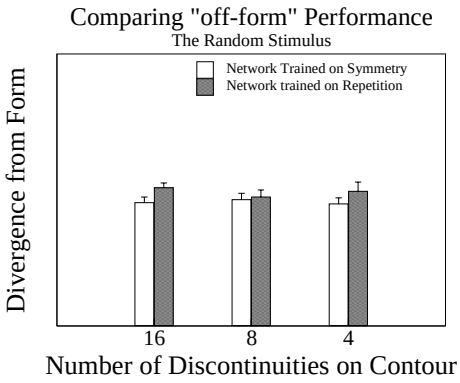


Figure 5: The compared "off-form" measures produced by trained networks, as a function of step number, under the *random* stimulus.

present even when we are examining how symmetric its output is.

Discussion

The findings of the behavioral experiment were three fold, viz.

- the processing of shapes by the human observer varies qualitatively in accordance with the characteristics of shape's contours.
- the processing of symmetric shapes is carried out in parallel, while the processing of repetitive shapes remains a serial task, with point by point comparisons
- this is ecologically consistent with a cognitive facility that maximized correct figure ground segregation in a two dimensional image

The second of these points was behind the susceptibility to changes in step number on the contours of repetitive shapes only. Their data was consistent with the hypothesis, that processing of repetitive shapes is affected by the cognitive costs associated with serial processing, unlike the processing of symmetric shapes.

In our model, there is a similar distinction in the way the network handles these two classes of stimuli. In particular, when considering the degree to which the specific regularity of shape-type is learned in the network, we find generally better performance on symmetries, in contrast with repeated shapes. In addition, the performance of the nets trained on repetitive stimuli shows that they are affected by the number of discontinuities along the side of the pseudo-random block shapes. This is true for novel stimuli of the same type as well as for non-symmetric, non-repetitive block stimuli.

Interpretations: Step Number

The main effect that we would like to address here is the differential processing of symmetry and repetition by both human subjects and the split architecture.[3] The central finding was that there was a strong interaction between shape type and step number, with the repetitive shapes taking the brunt of the effect, far outstripping the variance induced by step number in their symmetric counterparts.

It certainly seems plausible that that difference is based on fixed anatomical features; looking for the the features held in common by both the model and the human subjects, the general split in the architecture of the processor is the obvious front-runner. This anatomical consideration in particular suggests how the differentiation of symmetry and repetition may be linked to the distinction between parallel and serial processing *through* the very structure of the processor.

The formalization of what constitutes repetition and symmetry in the block-shapes, activation across the vertical access summing to a constant in the former and having a net difference of zero in the latter, was crucial to the analysis presented here. But it goes further: it implies that for a split processor, symmetry is as simple as cross-checking (or cross-generating, in the case of this task) the output from each hidden layer. Repetition, on the other hand, involves a cumbersome re-calibration, for *every* segment of the repeated pattern, because the cross-image portion of the output is not a simple reflection.

Now if this difference is that symmetrical shapes can be checked by a parallel system, then it is conceivable that that system is rooted in the recognized image being split centrally along its axis of symmetry and each half being presented to each visual cortex, which in turn provides a massively parallel "cancellation" style verification of the image. Were we to assume homotopic and inhibitory commissures, symmetry would be exactly the reciprocal cancellation of activity across the two sides of the visual cortex, an idea that finds little favor in some circles (Dakin, 1998), but which a simple model such as the one here might help to refine.

But how does this explanation differ from that drawn in the original experiment? Baylis and Driver identified an aspect of shape processing in which symmetry was distinguished from repetition by rendering null the processing costs of increased shape complexity. This was elaborated as being a case of parallel versus serial processing, in which complexity, which is directly proportional to step number along the contour, only retarded the serial process, leaving the parallel process (the detection of symmetry) unhindered. Thus, the process of recognition is essentially one that involves the comparison of the segment end-points that make up the shape, and in the case of symmetry these comparisons take place simultaneously.

In some sense, the model is not incompatible with the take on the behavioral data. Indeed, the human study leaves open the question of what the facilitating mechanism is for the parallel treatment of symmetry, and the model provides one such possibility. Nevertheless, there is an important contrast. As discussed below, the notion of what constitutes complexity is not fixed. For Baylis and Driver, complexity increased with step number, and parallel processing was where such an increase was insignificant. For the model however, it would be quite an assumption to simply associate increased step number with increased complexity, for point-by-point comparisons of the stimuli have little meaning in a model lacking a temporal dimension. However, the difference between the symmetric and repetitive shapes is just as marked. In the absence of any sequential processing, this indicates that symmetry is special not in avoiding the narrow view of complexity-as-quantity, but in generally "playing-down" any dependence on contour variations. Since parallelism and seriality have no temporal meaning in the model there must be a more fundamental retreat from complexity that symmetry offers.

Complexity

Above, we alluded to the unfixed nature of complexity. A potential shortcoming in the modeling result is that the effect of step number seems to manifest in the reverse direction. That is, an increase in the number of steps in the stimuli presented to nets trained on repetitive shapes meant an increase in how well that nets output stayed to form (form in this case being repetitive). We would expect this "accuracy" to decrease, given that more steps presumably means greater complexity and therefore a harder task (and one for which, in the original experiment, the authors saw a need for more "counting" time). If anything, the results of the model disturb the clear relation obtaining between the original interpretation of serial versus parallel and how each accommodates effective increases in complexity.

Of course, there is no claim that the model attempts to perform either serial, or parallel processing, in the sense that Baylis and Driver use those terms. And it is hoped that the previous section went some way to reducing the high-level cognitive connotations of these terms to more concrete, anatomically based concerns. The networks learn to perform an identification task, and in doing so they pick up the general trends elicited by the stimuli sets used for training. It is in this sense then that processing of shape type differs: it will depend on how

[3] Though beyond the limits of this paper, it is worth noting that an effect of orientation found in the behavioral study, marked generally by poorer performance (RT and error), also falls out of the split-network model.

the task was learned in the first place.

In this context let us ask what complexity is. For though an increase in complexity can be equated with an increase in step number; for our model it won't be. Again, complexity in terms of the number of sequential operations performed (i.e. counting the steps) has no meaning in the context of the network. So how could a contour with 16 steps appear easier to process, or be in general more likely to encourage good "on-form" output, than one with only 8 or 4 discontinuities?

The answer to this involves reviewing the nature of the task, from the network's perspective: to reproduce accurately a contour of maximal discontinuity, which, in the case of this model, is one with 16 steps, the net at least has the advantage of avoiding any cross-row constraints. That is, given that the grain of the image and the grain of the shape in question match, no additional provisos are required in order for the net to attempt reproduction of the input at the output. Thus the task is relatively unencumbered, and the result is a more stable version of the form learned during training. But reduce the number of steps along the contour and the complexity of the task the net has to solve is actually *increased* by virtue of the added constraints of aligning rows of "pixels" at the output.

It isn't that this in itself disrupts the measure of form at the output, for that is always measured at the grain of the model, but that such additional constraints divert the resources that the net has devoted to producing well-formed images. The result is a drop in the form at the output, but, and this is the key point, this whole story, in which complexity for the network is revealed to be the opposite of what one would expect, only effects repetitive shapes.

Summary

The described model presents interesting analogues to some of the main effects uncovered in behavioral studies. In particular we have the preference for the processing of symmetric shapes, which is much less susceptible to variations along the contour than is the processing of repetitive shapes.

As already noted, symmetry may initially seem more complex a phenomenon, in terms of the operations required to generate symmetric contours (translation and reflection). However, from an anatomical perspective it may fact be simpler, especially around the vertical axis. This relates to the view that homotopicality between the visual cortices promotes the recognition of vertically oriented symmetry—because information, instead of being quantized and stored, can be "mirrored and checked" directly.

Baylis and Driver present a plausible argument for symmetry preferences in terms of parallel and serial processing, in the cognitive sense. However, more parsimonious explanations may be available. Using a split architecture neural net, we have suggested that the symmetry preference may arise from gross anatomical aspects of the processor. If this is so, then the application of Occam's Razor suggests that there is a simpler story on symmetry.

References

Baylis, G. C. & Driver, J. (1994) Parallel Computation of Symmetry but Not Repetition within Single Visual Shapes, *Visual Cognition*, **1** (4), 337-400.

Dakin, S. C., & Herbert, A. M. (1998) The Spatial Region of Integration for Visual Symmetry Detection. Proceedings of the Royal Society of London, B265, pp659-664

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48, 71-99.

Hicks, J. (2002). A Computational Model of Split Processing. PhD Thesis (submitted), University of Edinburgh

Hicks, J. & Monaghan, P. (2001). Explorations of the interaction between split processing and stimulus types. In S. Wermter, J. Austin & D. Willshaw (Eds.) Emergent computational neural architectures based on neuroscience. Springer: Heidelberg.

Hicks, J., Oberlander, J. Shillcock, R.: Four letters good, six letters better: Exploring the exterior letters effect with a split architecture. Proceedings of the 22nd Annual Conference of the Cognitive Science Society. 2000. Madison, WI: Lawrence Erlbaum Associates.

Latecki, L. J. and Lakämper", Rolf (1999) Convexity Rule for Shape Decomposition Based on Discrete Contour Evolution, *Computer Vision and Image Understanding: CVIU* V. 73, No. 3, 441–454

Sally, S., and Gurnsey, R., (2001) Symmetry detection across the visual field *Spatial Vision* Volume 14, No. 2, pp. 217–234,

Shillcock, R.C.,Monaghan, P. (2001) The computational exploration of visual word recognition in a split model. Neural Computation, 13, 1171-1198

G. J. ven Tonder, Y. Ejima, (2000) The patchwork engine: image segmentation from shape symmetries *Neural Networks* 13, 291-303

Perspective-taking in Young Writer's Descriptive Writing

David R. Holliway (holliway@marshall.edu)
Department of Educational Foundations and Technology, 122 Jenkins Hall, Marshall University,
400 Hall Greer Blvd.,
Huntington, WV 25755 USA

Abstract

This paper reports a perspective-taking strategy that assisted younger writers in representing the descriptive needs of their readers. There were 154 writers (78 5th-graders and 76 9th-graders) and 52 9th-grade readers that participated in the study. Three conditions were contrasted: a feedback only condition, a "rating other" condition, and a "reading-as-the-reader" condition. Readers' correct description-to-tangram matches made for each of three sessions served as the dependent measure. Repeated measures analysis revealed both the 9th- and the 5th-grade writers showed consistent significant improvement under the "read-as-the-reader" condition when revising their essays and when drafting anew. The results suggest that when young writers engage in a process that mirrors their readers' experiences, they can more accurately revise their descriptive writing to meet their readers' informational needs.

Theoretical perspective

Writing is simultaneously an individual struggle and a social undertaking (Dyson & Freedman, 1991; Fitzgerald, 1992; Florio, 1979; Flower, 1994). Writers face the individual cognitive task of selecting what information to communicate and how they will communicate it. Inseparably, writers consider who their readers will be and the context of their reading. Writing scholars (Traxler & Gernsbacher, 1992, 1993; Fitzgerald, 1992) theorize that to meet the informational needs of readers, a writer must coordinate at least three interacting mental representations: a representation of personal communicative intent (what do I want to say?), a representation of the text produced (what have I written?), and a representation of the reader's perspective (how will the reader interpret my writing?).

Establishing reciprocity between reader, writer, and text is the hallmark of experienced writing (Witte, 1992; Olson, 1994). Considerable research (e.g., Bereiter & Scardamalia, 1987; Beal, 1996) has demonstrated that young writers are particularly challenged in learning this writer-reader-text reciprocity. Specific instructional conditions that foster "comprehension monitoring" (Beal, 1996; Fitzgerald & Markham, 1987) and "knowledge-transforming" (Bereiter & Scardamalia, 1987) can help young writers discriminate their intended message from the actual text they have composed, thus influencing *the textual quality* of their writing. Fewer studies, however, have outlined conditions that may help improve younger writers' awareness of their readers' possible interpretations. For example, Frank (1992) found that subtle manipulation of "audience specification" in writing prompts led fifth-grade writers to compose their newspaper advertisements differently for two separate audiences. The research literature (e.g., Bonk, 1990) remains unclear about instructional conditions that can help young writers view their text from the perspective of their readers, and thus improve *the communicative quality* of their writing.

To investigate how older students might become more sensitive to their readers needs, Traxler and Gernsbacher (1992, 1993) asked college students to compose and revise descriptions of tangrams for anonymous readers. The readers' task was to read each description and then select the matching tangram from a group of similar-looking tangrams. The writers who went through a revision process identical to that of their readers consistently wrote the most effective texts. Traxler and Gernsbacher (1993) concluded that the reciprocity between readers' needs, text, and writer could be successfully accomplished when writers read-as-their-readers, that is, when writers learn to take the informational perspective of their readers. If consideration of the readers' needs is critical to "good thinking during composition" (Fitzgerald, 1992, p. 345), then "reading-as-the-reader" may enable young writers to consider the descriptive needs of their readers. Reading-as-the-reader may be one strategy whereby young writers can coordinate "what do I want to say?" and "what have I written?" with "how will the reader interpret my writing?"

This paper reports on-going research (Holliway, 2000; Holliway & McCutchen, in press) that suggests "reading-as-the-reader" can improve fifth- and ninth-grade writers' ability to compose descriptive writing consistent with their readers' informational needs. Three questions guide the present paper: 1) Can "reading-as-the-reader" assist young writers in composing and revising descriptive writing that meets their readers'

informational needs? 2) What do the writers' post-experiment reflections reveal about three contrasting perspective-taking conditions? 3) How are the "readers' informational needs" reflected in the descriptive strategies used by these writers?

Methods

Participants

All participants came from four elementary schools and three high schools located in a large metropolitan area. There were 154 writers (78 5th-graders and 76 9th-graders) that came from regular language arts classes. The readers were a separate group of 52 9th-grade readers in advanced placement English classes.

Design

A written referential communicative paradigm was adapted from Traxler and Gernsbacher (1993). The writer communicated the details of Tangram to a reader who chose the "target-gram" from a group of similar-looking tangrams. There were three writing sessions each separated by one-week intervals. Each writing session was followed on a separate day of the same week by a reading session.

Materials

The tangrams that the writers described came from a collection of 72 figures (similar to those used by Traxler and Gernsbacher, 1992, 1993; Clark and Wilkes-Gibbs, 1986). Tangrams were counterbalanced across sessions and conditions.

Procedures for Writers

Session one

All writers were given a notebook with three tangram figures to be described. Writers had 30 – 35 minutes to complete their descriptions.

Session two

Writers in each classroom were randomly assigned to one of three perspective-taking conditions. The three conditions differed in how closely the writers' task mirrored that of their readers.

Feedback-only condition

Writers received a sentence for each description indicating whether their reader had successfully matched the description with the associated target-gram. Writers then revised their original descriptions.

Feedback + rating-other condition

Writers received a feedback sentence for each description indicating whether their reader had successfully matched the description with the associated target-gram, plus three descriptions written by another student. Writers rated the descriptions by considering the informational adequacy of each description (e.g., which description creates a clearer picture in your mind?). After finishing they revised their original descriptions.

Feedback + reading-as-the-reader condition

Writers received a feedback sentence for each description indicating whether their reader had successfully matched the description with the associated target-gram, plus three descriptions written by another student, and then they matched each description with tangrams, exactly as their readers had done. After they finished their matching, they revised their original descriptions.

Session three

After finishing their task-specific activity, all writers received a new set of three tangrams to describe.

Procedures for Readers

The readers received a notebook that contained typed versions of the tangram descriptions, and a scorebook wherein they made their description-to-targetgram matches. For the entirety of the experiment, the same reader scored the same three writers, each writer representing one of the three experimental conditions.

Results

Quantitative analysis

The dependent measure for the 2 (grades) x 3 (tasks) x 3 (sessions) repeated measures analysis was the number of correct description-to-"target-gram" matches that each reader made for each description they read (For Mean differences see Table 1 below).

Table 1: Means and standard deviations by Session, condition, and grade.

	Session 1			Session 2			Session 3		
Condition	**N**	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**
Feedback									
9th-grade	18	2.17	.92	18	2.39	.85	18	2.28	.96
5th-grade	25	1.80	.76	25	2.20	.76	25	1.68	.95
Rate-Other									
9th-grade	26	2.23	.77	26	2.42	.76	26	2.42	.70
5th-grade	30	1.87	1.04	30	2.00	.95	30	2.27	.87
Read-as-the Reader									
9th-grade	32	1.75	.88	32	2.25	.84	32	2.47	.67
5th-grade	23	1.57	.59	23	2.13	1.01	23	2.26	.69

The between subject effect revealed a main effect of Writer, $F(1,148)=11.00$, $p=.001$. On average, the ninth-graders scored higher than the fifth-graders throughout all sessions and in all tasks. The within subjects effects revealed a significant main effect of Session, $F(2,296)=8.76$, $p<.001$, with session 1 ($M = 1.88$, $SD =$

.86) yielding fewer matches than session 2 (M = 2.22, SD = .87) and session 3 (M = 2.24, SD = .83). However, the session main effect was compromised by a significant interaction between session and condition, F (4,296) = 2.96, p = .019. Post hoc analyses (Tukey's Honestly Significant Difference (HSD) approach) established that differences between session 1 and sessions 2 and 3 were significant only for the read-as-the-reader group (critical value = .375, p = .05). No other interactions reached significance (F <1). These results suggest that reading-as-the-reader helped both the 5th- and the 9th-graders in meeting their readers' informational needs more than the other two conditions.

Qualitative analysis

An analysis of the writers' open-ended free-writes about their "reading-as-the-reader" experiences revealed that the task was very useful for these writers. Students portrayed their writing experiences on a variety of levels, usually characterizing the task in some way as fun or boring, insightful or uninspiring. The student free-write responses were used to generate a general coding scheme that categorized their experiences as positive or negative, useful or not useful.

Based on the coding scheme, the percentage of students in each condition who characterized their writing experience positively was calculated. Not all students provided a free-write. Table 2 presents the number of students responding in each condition, as well as the percentage. The percentage of positive responses from students in the Feedback condition was compared to those in the Rate-Other and the Read-as-Reader conditions. Students in the Feedback condition were significantly less likely to characterize their writing experiences as positive, compared with students in the other two groups (Fisher's Exact = 6.787, p =.005).

Table 2: Actual number and percentage of students who responded positively to their experimental condition.

	Positive Responses	
	# Responding in each condition	% Who responded from each condition
Feedback	20	46.5%
Rate-Other	38	67.9%
Read-as-Reader	39	70.9%

To investigate the "readers' needs" a profile was compiled based on the readers' open-ended comments made at the end of each reading session. An analysis of their comments about what they needed from their writers revealed that *a global conceptual image* created by an analogy with a balance of *local shape and spatial elaborations* helped them discriminate and chose the "target-gram" from the group of similar-looking tangrams. For example, one reader commented: "The descriptions that were the best were very detailed in the shapes and what the figure looks like it's doing." The readers' profile revealed that the readers' informational needs were met more efficiently by writers who elaborated on the analogical referent with a balance of shape names (e.g., triangle, parallelogram, square), geometrical qualifications (e.g., zigzaggy, diagonal, pointy) and location descriptors (e.g., to the right, on its left, the left one).

A text analysis of the descriptions revealed that many writers, regardless of condition and grade, began their descriptions with analogies (e.g., "It looks like a running fox," "This tangram looks like a ghost flying.") These "spontaneous analogies" (English, 1997, p. 15) may be one way writers are attempting to establish a common perceptual ground with their readers. Writers varied, however, in the way they elaborated on the spatial and geometric qualities of the tangrams they described. Many writers used an "object centered" strategy that focused on the intrinsic details of each tangram. For example this writer's description represents a common strategy: "It looks like a goose. It has a long zigzagging neck. It has a small head and a pointed beak. Its body is kinda [sic] long and it has two feet on top of each other." At this point in the analysis it is not easy to identify changes in writing strategies and textual features due to enhanced reader perspective. Initial text analyses of the descriptive essays generated in this study reveal few structural differences that can be associated with condition.

Discussion

Theoretical Implications

All three groups of writers received feedback indicating the accuracy of their reader's choice. The rate-other group also read and evaluated descriptive tangram texts written by other students. However, only the read-as-the-reader group was asked to take their readers' perspective in the actual task of matching descriptions to tangrams. Although the mean scores improved significantly from session 1 to session 2, the cognitive potency of the read-as-the-reader condition emerged most strikingly in the "transfer comparison" between sessions 1 and 3. It may be the case that the intervention duration was not sufficient in the second session to show a significant revision in the original descriptions. Perhaps more than one experience with reading-as-the-reader is necessary for younger writers to show the benefits.

An alternative explanation involves the nature of the writing task in session 3 compared to session 2. In the second session (revision session), writers may have been under the influence of the text they had already created; the actual physical text that they composed in the first session may have constrained the creation of a new text fresh with detail. Bereiter and Scardamalia (1987) suggest, "the original version of text, *because it is perceptually present* [emphasis added], has a direct claim on conscious attention. Unless the writer can deliberately bring alternatives to mind, the original text will win for lack of competition" (p. 87). The reading-as-the-reader condition had the greatest impact when students were given a chance to apply and recontextualize what they had learned from composing one set of texts to the composition of a similar, but new texts. That is, when writers drafted anew in the third session, unconstrained by an existing less-effective text, they were able to demonstrate what they had learned from "reading-as-the-reader."

The positive responses that students made suggests that "reading-as-the-reader" gave these writers a perspective on the effects of their writing that they otherwise might not have considered. One writer reflected "I like to read other kids' descriptions because sometimes if I read other kids [sic] descriptions I can get more ideas . . . because when I look back into the pictures I can't see the pictures they see." The analysis of the writer's comments from the read-as-the-reader group suggests that actually doing the task their readers did revealed to them the necessary information they needed to include and the unnecessary information they needed to exclude in their descriptions.

Further research might "directly probe the ways in which individuals cope with the items or task, in an effort to illuminate the processes that underlay item response and task performance" (Messick, 1989, p. 6). By conducting protocol analysis students' thoughts could be assessed to reveal the kinds of decisions that writers make and the kinds of information they chose to include and exclude, and ultimately, the kinds of discourse strategies that they chose to use in an attempt to meet their readers' informational needs. This is one approach we might take to better understand how reading-as-the-reader can assist younger writers in accomplishing the writer-reader-text reciprocity.

Educational Implications

This research contributes to a body of literature (e.g., Beal, 1996; Cameron, Hunt, & Linton, 1996; Frank, 1992; Oliver, 1995) that clarifies some of the instructional conditions that can help young writers envision how their readers' interpret the text they have written. Specifically, it contributes to our understanding of how younger writers can learn of the reciprocity between writing, reading, and text (see Witte, 1992). The study offers empirical support for the widespread classroom practice of peer editing and peer response. This study suggests, however, that peer response may be more effective when peers actually use the text in some way, because they are forced to confront the text's strengths and weaknesses in a concrete context, rather than the more abstract context of giving literary feedback.

Although the "referential communication design" has been traditionally associated with experimental psychology and spoken communication, similar activities and communicative processes are found in writing instruction literature. For example, an adaptation of the writing/reading exercise "reading-as-the-reader" might be added to a teacher's repertoire of "optimal environmental activities" (Daniles, 1990, pp. 118-121). Daniles suggests, "lessons about effective descriptive writing emerge from experiencing strategies in use" (p. 119). "Reading as the reader" is a perspective-taking strategy experienced when the writer attempts to create a specific description their readers can "see."

Another application of "reading as the reader" would be a perspective-taking strategy that can be added to a "writer's tool box" (Harper, 1997). Harper describes five revision tools that she suggests have worked for her as a practicing middle school writing teacher. One such tool is the "snapshot." Students compose written snapshots similar to a detailed photographic snapshot. Snapshots are writing activities that compel students to concentrate on the physical properties and descriptive qualities of various "objects." "Reading-as-the-reader" could be an instructional tool that writing teachers incorporate into his/her repertoire of classroom activities to help students become more efficient descriptive writers.

Finally, "reading-as-the-reader" may help students to make details explicit and assist students in recognizing other text creating approaches that could be used with other functions of writing. Composing concrete poems and descriptive essays and then "reading as the reader" are classroom experiences that may facilitate students going beyond their immediate personal and social circumstance (Cameron, Hunt, & Linton, 1996; Elasasser & John-Steiner, 1977; Florio, 1979). If "reading-as-the-reader" is a learning strategy that worked for younger writers in helping them develop a readers' perspective in transactional writing, it might also be a strategy transferable to other writing purposes. Not only does "reading-as-the-reader" assist writers in asking, "what do I want to write?" and "what have I written?" , more importantly, it may assist in addressing the more challenging task of, "How will the reader interpret my writing?"

Acknowledgments

I would like to thank Professor Deborah McCutchen at the University of Washington, Department of Educational Psychology for her support and advice throughout this project.

References

Beal, C. (1996). The role of comprehension monitoring in children's revision. *Educational Psychology Review, 8* (3), 219 - 238.

Bereiter C., & Scardamalia, M. (1987). *The psychology of written composition.* Hilsdale, NJ: Erlbaum.

Bonk, C. (1990). A synthesis of social cognition and writing research. *Written Communication, 7,* (1), 136 - 163.

Cameron, C., Hunt, K. A., & Linton, M. (1996). Written expression as recontextualization: Children write in social time. *Educational Psychology Review, 8* (2), 125 - 150.

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1 - 39.

Daniles, H. (1990). Young writers and readers read out: Developing a sense of audience. In Timothy Shanahan (Ed.), *Reading and writing together: New perspectives for the classroom.* Norwood, MA: Chirstopher-Gordon Publishers.

Dyson, A & Freedman, S. (1991). Writing. In J. Flood, J. M. Jensen, D. Lapp, and J. R. Squire (Eds.), *Handbook on teaching the English language Arts* . New York: Macmillan.

Elasasser, N. & John-Steiner, V. P. (1977). An interactionist approach to advancing literacy. *Harvard Educational Review 47* (3), 355 - 369.

English, L. (Ed.). (1997). Mathematical reasoning: Analogies, metaphors, and images. Mahawah, NJ. Lawrence Erlbaum.

Fitzgerald, J. & Markham, L. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4* (1), 3 - 24.

Fitzgerald, J. (1992). Variant views about good thinking during composing: Focus on revision. In, M. Pressley, K. Harris, and J. Guthrie. (Eds.). *Promoting academic competence and literacy in school.* San Diego, CA: Academic Press.

Florio, S. (1979). The problem of dead letters: Perspective on the teaching of writing. *The Elementary School Journal 80*, (1), 1 - 7.

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing.* Carbondale: Southern Illinois University Press.

Frank, L. (1992). Writing to be read: young writer's ability to demonstrate audience awareness when evaluated by their readers. *Research in the Teaching of English, 26*, 277 - 298.

Harper, L. (1997). The writer's toolbox: Five tools for active revision instruction. *Language Arts, 74,* 193 - 200.

Holliway, D. (2000). It looks like a goose: Composing for the informational needs of readers. American Educational Research Association (AERA), Paper presentation for the writing and literacies special interest group, New Orleans, LA.

Holliway, D. & McCutchen, D. (in press). Audience perspective in children's descriptive writing: Reading as the reader. In Linda Allal, Lucile Chanquoy, & Pierre Largy (Eds.), *Revision of written language: Cognitive and instructional processes.* Amsterdam: Kluwer Academic Press.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18* (2), 5 – 11.

Oliver, E. (1995). The writing quality of seventh, ninth, and eleventh graders, and college freshman: Does rhetorical specification in writing prompts make a difference? *Research in the Teaching of English, 29* (4), 422 - 450.

Olson, D. (1994). *The world on paper.* Cambridge, MA: Cambridge University Press.

Traxler, M. & Gernsbacher, M. (1992). Improving written communication through minimal feedback. *Language and cognitive process, 7,* 1 - 22.

Traxler, M. & Gernsbacher, M. (1993). Improving written communication through perspective-taking. *Language and cognitive process, 8* (3), 311 - 334.

Witte, S. (1992). Context, text, intertext: Toward a constructionist semiotic of writing. *Written communication, 9,* (2), 237 – 308.

An Instance-based Model of the Effect of Previous Choices on the Control of Interactive Search

Andrew Howes (HowesA@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff, CF10 3YG, Wales, UK

Stephen J. Payne (PayneS@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff, CF10 3YG, Wales, UK

Juliet Richardson
School of Psychology, Cardiff University, Cardiff, CF10 3YG, Wales, UK

Abstract

How do people control interactive search? One type of decision that is made when performing a task such as searching the web is whether to continue to explore unattractive but immediately available links or to backup to previously experienced links. It has recently been suggested that this choice may be governed by a preset threshold. We report empirical evidence that in fact the choice is governed by memory for the quality of the unselected alternatives on previous pages. Further, we report a computational model that combines an instance-based memory for previous evaluations with display-driven action to control interactive search.

Introduction

Tasks such as web browsing and menu search are examples of what we call *interactive search* tasks. They differ from other problem solving tasks in that the effect of an operator is unknown until the operator has been implemented in the world. In these circumstances a problem solver cannot use mental lookahead in order to constrain search, rather search is constrained by two mutually dependent cognitive activities.

First, people estimate the relative likelihoods that operators will lead to the goal and trade these off against estimates of cost (Pirolli and Card, 1999). In interactive search, estimates of likelihood are typically based on an interpretation of the relationship between the goal and the semantics of the word(s) and icon(s) used to represent the operator. Estimates of cost are often based on the time that operators are expected to take to implement in the world.

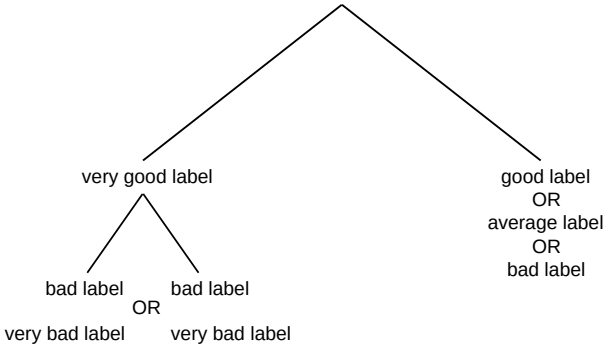
Second, the process of likelihood estimation must be embedded within a strategy for controlling search. A strategy typically defines policies for determining which operators are included in the set of those considered and policies for reducing or eliminating the probability that operators that have been tried before are redundantly reselected. Typically a strategy for interactive search will be supported by memory for which operators have already been tried on the current search (so that reselection can be avoided) and by memory for information about which operators lead to success or failure on previous trials (Howes, 1994). Howes and Payne (2001) have argued that these needs are not easily captured by current architectural theories of human cognition.

A search strategy is required in addition to an ability to follow label semantics because in real menus and on real web pages, label semantics (and therefore estimates of likelihood) are rarely sufficient to guarantee that users will navigate directly to the location of a goal without some fruitless exploration of other parts of the search space. An understanding of the process by which people rank order operators must be complemented by a model of which they consider, how they remember which they have tried, and how they remember where to find them.

Previous research on the strategies that people use to control exploration in these circumstances has emphasised that performance is often display-based. I.e. problem solving is constrained by the set of operators currently available on the display of the device (Howes and Young, 1996). The device display imposes constraints on the problem solving process that limit the cognitive costs of device use. Other research has indicated that judgements of what choices have been taken before are sometimes made on the basis of long-term representations of familiarity (Payne, Richardson, Howes, 2000).

Other research has directly addressed the question of when users choose to 'backup' from a choice point. A 'backup' operator is a special kind of operator in interactive search in that the user knows that its effect is to return the user to the previous, higher, node in the tree (although the user may not know the content of this node). There are a number of types of backup in web browsing. The 'Back' button on a browser takes the user to the previous page in the recent history stack. A link labelled 'back' on a web page will typically take a user to a page one level higher in the site hierarchy (though this is not guaranteed). The differences between these operations are potentially interesting but for here we consider a generic backup where the two definitions are aligned.

One answer to the question of when users choose to backup is that backups are selected when evaluations of all available forward moves fall below some threshold (Miller and Remington, 2001). In their model, Miller



and Remington assumed that users selected a link whenever the perceived likelihood of success exceeded some predetermined threshold. They point out that a feature of this model is that it places little demand on memory. Search is controlled without a memory for previous alternatives to the current search path. Miller and Remington also describe an elaboration of their threshold model in which "improbable links at a lower tier" are selected opportunistically. This is achieved by temporarily reducing the threshold for selection. Their model was motivated by examination of web usage logs which suggested that users selected less probable links before backtracking to other possibilities.

An alternative model is that people moderate their willingness to backup according, in part, to their memory for the quality of previously unselected operators at higher levels in the menu tree. In this model a single fixed threshold would not be used, rather backup would be selected according to a computation of the relative values of an extended set of operators that included, but was not limited to, those derived from the currently displayed labels. It is possible that people backup if the previous label was significantly better than the current labels, or perhaps only if both the current labels were below some threshold (as in Miller and Remington's model) and the previous label was above some other threshold.

In addition it seems likely that the perceived time cost of successfully returning to a remembered option will moderate people's willingness to select a backup option. Such behaviour would be consistent with recent findings in Human-Computer Interaction (Gray & Fu, 2001) and with the conflict resolution mechanism of ACT-R (Anderson and Lebiere, 1998).

In this paper we first report an experiment designed to test the hypothesis that when engaged in interactive search tasks people take into account the value and cost of options other than those that are immediately available on the computer display. We also report a model of interactive search, that is consistent with the results of the experiment, and which is based on an instance-based framework of memory for problem solving. The model was developed in response to the current findings and in part to our previous work on interactive search (Payne, Richardson and Howes, 2000; Howes and Payne, 2001).

Experiment

The aim of the experiment was to test the hypothesis that people take memory for the value and cost of unselected menu options from previous choice points into account when deciding whether to backup.

Participants were asked to search for a different target in each of a series of identically structured binary menu trees. Selection of an option resulted in the displayed menu being replaced by a submenu. The trees differed in three respects: (a) the quality of the weaker option at the top level choice-point in the menu (either a good, average or bad choice); (b) the quality of the two options at the lower level, or critical choice-point (either both bad or both very bad choices for the goal); (c) the cost of backup.

The quality of labels was determined by previous studies in which participants were asked to fill in a questionnaire indicating the likelihood that they would select a label for a particular goal. For example, they were asked to rate "Cowboy movies" and "Facts and Figures" for the goal, "find when John Wayne died." Answers were given on a five point scale, from 1 (very likely), to 5 (very unlikely), the levels of which are referred to in this paper as very good, good, average, bad, and very bad.

In accordance with (a) and (b) above, at the top level of the menu tree there was always one very good option and one that was either good, average or bad. On the menu underneath the top level very good option there was a node with labels that had both been rated bad or both very bad. (Of their own volition, participants were expected to mostly select the best choice at the top level.)

The number of times that participants chose to backup as the first move made from the critical choice-point was recorded. The design of the experiment allowed us to determine whether the number of backups was dependent only on the quality of the labels at the critical choice-point (bad vs. very bad) or also on the quality of remembered but untried labels at higher levels of the tree (good vs. average vs. bad) and/or on the cost of backup.

A "give up" option was available so that participants did not need to follow paths under bad or very bad options in order to find the goal. This would otherwise be the case in the menu trees where the alternative option at the first choice-point was a label that had been rated as bad. This was designed to ensure that participants experienced minimal positive feedback for the bad menu labels.

Method

Participants. Thirty-six undergraduate students (30 females and 6 males) participated in this study for course credits. The mean age of participants was 19 years 11 months.

Materials. Twenty-seven binary menu trees were used. Participants were required to find a single goal in each of the menu trees. The goals were all to find general information on different topics.

The twenty-seven menu trees consisted of eighteen test menu trees and nine filler trees. In each tree the first choice was between a label that had previously been rated as a very good label for the goal, and one that had been rated as either good, average or bad. The very good choice led to a choice-point where the two labels had been rated as either both bad or both very bad.

Nine different topics were used for the test trees, with each topic occurring twice. Each topic was used for a menu tree with a very bad critical choice-point and for a menu tree with a bad critical choice-point giving nine of each in total. Within each of these sets of nine menu trees, three had a good alternative option at the first choice-point, three had an average option and three had a bad option. Across participants, each topic was presented equally often as each of the six different types of test menu.

There were two locations for the goal in each of the test menu trees. The goal information could be found either by moving forwards from the critical choice-point, or by backing up from it and searching the other half of the structure.

Procedure. All participants carried out a simple menu search training tasks before taking part in this experiment. After reading the instructions, participants worked through three practice search tasks and then through the twenty-seven menu search tasks presented in a different partially-randomised order for each participant. Presentation of tasks was self-paced in that participants had to find the goal (or give up) before the next task could be started. Selections were automatically recorded by the program.

Participants were randomly assigned to either a low-cost, immediate backup group or to a high-cost, slow backup group. Participants in the high-cost group had to carry out two intermediate steps between clicking on backup and getting to the previous choice-point. Both of these steps required participants to click buttons in windows to confirm that they wanted to backup. As before, participants in these two groups were matched for frequency and length of Internet use.

Results and Discusssion

The tendency to select the give-up option rather than continuing until the goal was found was very low. It was used on only 5% of tasks on average. As expected, it was used most often on those tasks where the untried option at the first choice-point was bad rather than good or average.

The mean percentage of backups made as the first move from the two types of critical choice-point (bad and very bad) in each of the three types of menu tree (good, average or bad untried previous option) was calculated for each participant. These data are summarised in Table 1 and were subjected to an Anova to test for effects of critical choice-point type, preceding untried option type and cost of backup.

		Fast backups		Slow backups	
Critical choice-point	Previous untried option	M	S.D.	M	S.D.
Very bad	Good	73%	31%	55%	37%
Very bad	Average	61%	26%	52%	25%
Very bad	Bad	50%	26%	50%	30%
Bad	Good	70%	28%	61%	26%
Bad	Average	52%	38%	45%	22%
Bad	Bad	44%	30%	44%	25%

Table 1. The mean backups made as the first move from the critical choice-points in each menu type.

There was a significant effect of the goodness of the preceding untried option on the number of backups made, $F(2, 68) = 8.45$, $p < 0.01$. Significantly more backups were made from the critical choice-point when the preceding untried option was good than when it was bad or average. There was no main effect of the quality of labels at the critical choice-point, $F(1, 34) = 1.14$, $p = 0.29$. Equal numbers of backups were made whether the options were bad or very bad. This was not due to floor or ceiling performance: the average percentage of backups made from the critical choice-points was 55%.

There was not a significant main effect of cost of backup, $F(1, 34) = 1.77$, $p = 0.19$, nor were there any significant interactions.

Finally, there are two ways of looking at this data, either in terms of the assessment of the preceding untried option (as above), or in terms of the difference between the untried option and the options at the current choice-point. However, it is hard to quantify differences in assessments. The fact that there was no significant difference between bad and very bad critical choice-points is evidence against the difference in assessments being a factor.

An instance based model

The model is a computationally implemented model of the strategy underlying the direction of the effects observed in the experiment. It consists of an algorithm

implemented in a simple but novel production system framework developed by the authors. A brief description of the framework is given before the details of the strategy.

Framework

The basic assumptions in the framework (though not the model) were motivated by previous research rather than by the current findings. The primary motivation was the problem of discriminating which trial a memory was from, and in particular whether a memory was from the current trial or from a previous trial. Howes and Payne (2001) have argued that the way in which information is represented in ACT-R's declarative memory (Anderson and Lebiere, 1998) makes it difficult to model the control of search over multiple trials within the same search space. One problem is that the combination of frequency and recency information in base level activation makes it difficult to distinguish whether an activation is high because a representation was used on the current trial (recency), or high because it has been used many times before (frequency). While ACT-R models are sometimes built so that they do encode episodic chunks, it is not clear from the theory when a new chunk should be encoded and when the activation of an old chunk should be increased.

The instance-based framework that we describe here is a response to these problems. Where in ACT-R, repeated exposure to a goal or aspect of the environment results in an incremental increase in the base-level activation of the chunk, in the framework described here, repeated exposure to patterns results in the encoding of separate instances (where an instance is a structure consisting of a collection of attribue/value pairs). Effects of frequency can be captured in the framework by a race between instances that match to the current goal and state. The approach has been inspired by Logan's (1988) instance-model of practice and by Altmann and John's (1999) episodic account of how people control search during program comprehension.

In brief, the main assumptions behind the framework are given below. Many of the assumptions are derived from ACT-R and Soar but the framework differs substantially from both in the structure of its declarative/working memory. While we believe that these assumptions have the potential to offer a novel approach to modeling the control of cognitive behavior, they should not be taken in their current form as a competitor to the established architectures. ACT-R for example consists of a sophisticated set of mechanisms that have been shown to be useful in modeling a broad range of behavior. In contrast, we have focused on just those mechanisms required to capture a handful of experiments on a specific but important issue. The assumptions are:

1. The framework includes two types of data structure: (1) production rules, and (2) instance structures. Production rules match to instance structures to produce more instances and/or action.
2. Instance structures consist of (Identifier, Attribute, Value) triples. So for example, (o1, isa, operator),(o1, name, press),(o1, target, "tools"),(o1, state, s1) represents an operator o1 with four features. Similarly, (s1, isa, state) might be part of the representation of the state to which o1 has been applied. An instance cannot be modified or deleted. New instances may refer back to old instances.
3. The identifier of the most recent instance is held in a buffer. Another buffer holds a specification of the input (information from perception).
4. Whenever a production rule fires it adds new instances to instance memory. So for example, if the production that created o1 was to fire again it might add the triples, (o2, isa, operator),(o2, name, press),(o2, target, "tools"), (o2, state, s5). Both o1 and o2 would then be in instance memory, but note that only o2 would be linked to s5.
5. Conflict resolution. Serial control is imposed at the level of production firing. A production only fires once on the same data. Production matches are selected at random, though behavior may be moderated by high frequency matches.
6. Productions propose operators. Operators can carry preferences, e.g. "high", but are otherwise selected at random after a certain number of cycles have passed since the previous choice.

Unlike in ACT-R, frequency and recency information are not merged and it is not therefore difficult to distinguish the current trial from previous trials. The framework is suitable for modeling the findings of Howes and Payne (2001). It is also suitable for modeling the results of the experiment reported here.

Strategy

The results of the experiment indicate that much of the time participants preferred higher value operators regardless of whether they were available on the current menu. The strategy for the model therefore considered not only choices available on the current display (i.e. those that are cued by the environment) but also choices that it had previously experienced. The strategy was encoded in the instance-based framework in terms of a set of production rules. These rules proposed operators determined by the currently displayed menu items and by instance-based memory for previously displayed untried operators. Importantly, as we will see, the model did not need to remember the previous label, merely the fact that there was a previous highly rated choice.

The experiment was also suggestive of some effect of the cost of backup on participants' decision making. While this effect was not significant, it would be surprising and counter to much previous work if people did not take cost into account in this kind of decision and we have therefore chosen to include a sensitivity to the cost of backup in the model.

Even for this simple experimental task, the production rules also need to be sensitive to whether a memory was from the current trial or from a previous trial. Participants in the experiment experienced a whole sequence of tasks, and would have had to be able to determine whether a memory for a previous, highly rated menu option was for the current task. This is achieved by taking advantage of the discrimination made available by the instance-based encoding.

Behavior of the model

To illustrate the behavior we offer a trace for a typical experimental scenario. The model was given the goal of finding the target "John Wayne". The first choice was between "Films" and "Celebrities" for both of which the model had been given a "high" likelihood rating (based on a collection of human ratings). The model retrieved these ratings (lines 2 and 4) and also asserted that neither label had been recognized as tried for this trial (lines 1 and 3). On the basis of the gathered evidence the model then proposed the selection of each button (lines 5 and 6) and then selected "Celebrities" at random (line 8). (note that "…" indicates where there was a sequence of "wait" operators (e.g. line 7).)

1. recognise_no i13 label="Films"
2. retrieve_likelihood i14 label="Films" value=high
3. recognise_no i15 label="Celebrities"
4. retrieve_likelihood i16 label="Celebrities" value=high
5. propose_forward i17 label="Films" pref=high
6. propose_forward i18 label="Celebrities" pref=high
7. …
8. Select: i18
9. apply_forward i24 ACTION (press"Celebrities")

The model was then presented with a choice between "Comedy Films" and "Companies" both of which had been given a "low" rating (lines 10 and 13). Two "low" rated forward operators were then proposed on the basis of the gathered evidence (lines 16 and 17). In addition, a "high" rated alternative was retrieved (line 11). This retrieval was made from a previously encoded instance of a highly rated proposal, but importantly, retrieval for the actual label was not required. The retrieval lead to the proposal of a "medium" rated backup operator (line 12). The model chose the backup operator (line 19) over the "low" rated forward operators. NB. backup was only given a "medium" rating in this circumstance because of the additional cost to be expected prior to the selection of the forward move to which the model was returning.

10. retrieve_likelihood i30 label="Comedy Films" value=low
11. retrieve_alternative i31 target=i17 pref=high
12. propose_backup i32 label=backup target=i17 pref=medium
13. retrieve_likelihood i33 label="Companies" value=low
14. recognise_no i34 label="Comedy Films"
15. recognise_no i35 label="Companies"
16. propose_forward i36 label="Comedy Films" pref=low
17. propose_forward i37 label="Companies" pref=low
18. propose_backup i38 label=backup target=i17 pref=medium
19. Select: i38
20. apply_backup i40 ACTION (press backup)

At this stage the model has returned to the top-level choice point and immediately recognized that it has tried the "Celebrities" label before on this trial (line 21). However, as "Films" is not recognized as tried and is highly rated it selects it (line 27).

21. recognise_yes i46 label="Celebrities"
22. retrieve_likelihood i47 label="Films" value=high
23. recognise_no i48 label="Films"
24. propose_forward i49 label="Films" pref=high
25. retrieve_likelihood i50 label="Celebrities" value=high
26. …
27. Select: i49
28. apply_forward i56 ACTION (press"Films")

The model is now given a choice between two "low" rated labels. This time, no retrieval of a previous and highly rated operator occurs so one of the "low" operators is selected. (The 5% of trials on which participants chose to "give up" the search at points like this are not modeled.)

29. retrieve_likelihood i62 label= "Careers" value=low
30. recognise_no i63 label= "Education"
31. retrieve_likelihood i64 label= "Education" value=low
32. propose_forward i65 label= "Education" pref=low
33. recognise_no i66 label= "Careers"
34. propose_forward i67 label= "Careers" pref=low
35. …
36. Select: i65
37. apply_forward i72 ACTION (press "Education")

In addition, to the above, the model was run on the range of label rating combinations used in the experiment and produced behavior consistent with the findings in each circumstance. We have not reported aggregated statistics of the models performance here, as the participant responses to which such an analysis would be compared were probably dependent on finer grain label ratings than were provided to the model. What is important for our current purposes is that the model captures the qualitative distinctions observed in the experiment.

Discussion

We have presented an integrated model of interactive search that is based on an instance-based account of human memory. The model captures findings from an experiment reported in the current paper and is consistent with previous findings (Howes and Payne, 2001). Specifically, while operator proposal is primarily display-based, operators are also proposed on the basis of memory for previous untried same-trial operators. We have claimed that this instance-based approach provides the fine discrimination for the source of memories that is required in order to model the data.

While we have empirically demonstrated that people moderate their willingness to select backup operators on the basis of memory for previous unselected alternatives, a threshold account may still relevant to performance. For example, a threshold may be required to determine 'give up' decisions, and also to determine, at the first choice point, whether to select an item or scan for another. How this threshold is determined is an issue that requires further research.

There are many aspects of the interactive search data that we have not attempted to capture. Miller and Remington (2001), for example, describe a thorough analysis of how their model captures aspects of the depth/breadth trade-off in human performance with menu systems. It is possible that our model is consistent with Miller and Remington's threshold model in this respect but the analysis remains to be done.

The model that we have described can be contrasted to a method of search control known as operator subgoaling (Laird, Newell and Rosenbloom, 1987). With operator subgoaling, the best operator that has been proposed is selected even though it cannot be implemented directly in the current state. The operator is posted on the goal stack and the preconditions for operator application are posted as the current goal. In contrast, the search strategy that we have described here is relatively lean in the demands that it places on memory. When a decision was made that there was an attractive, previously experienced operator, this operator was not posted as the goal, rather the problem solver chose the single operator required to achieve the required preconditions. Once these have been met, the choices on the new menu are considered afresh and a choice made. In general, it is possible, that the greater power of the operator subgoaling mechanism is required to model human interactive search. It is often the case, for example, that establishing the preconditions for an operator requires more than one step. In this circumstance operator selection needs to be guided by a consistent focus on the desired preconditions. We see no reason why the instance-based framework that we have described should not be capable of supporting this more sophisticated strategy.

Lastly, it is worth considering the fact that we have not chosen to include mechanisms of decay and interference in the model reported here. The reason for this is that these mechanisms were not required to capture the findings of the experiment. However control strategies often do not degrade gracefully as memory becomes unreliable. Implausible perseveration is, for example, a frequent consequence of the loss of critical information from the memory of a model. It is likely therefore that this issue will need to be revisited.

References

Altmann, E. M. & John, B. E. (1999). Episodic indexing: A model of memory for attention events. *Cognitive Science, 23*,117-156.

Anderson, J. R. & Lebière, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Gray, W. D., & Fu, W.-t. (2001). Ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head: Implications of rational analysis for interface design. *CHI Letters, 3*(1), 112-119.

Howes, A. (1994). A model of the acquisition of menu knowledge by exploration. In W. Kellog & T. Hewett (Eds.) *Proceedings of the ACM Conference on Human Factors CHI'94*. New York: ACM.

Howes, A. & Young, R.M. (1996). Learning consistent, interactive and meaningful device methods: A computational model. *Cognitive Science, 20*, 301-356.

Howes, A & Payne, S.J. (2001). The strategic use of memory for frequency and recency in search control. *Proceedings of the Annual Conference of the Cognitive Science Society*, Edinburgh.

Laird, J., Newell, A., Rosenbloom, P. (1987). Soar: an architecture for general intelligence. *Artificial Intelligence, 33*, 1-64.

Logan, G.D. (1988) Toward an instance theory of automatization. *Psychological Review, 95*, 492-527

Miller, C.S. & Remington, R.W. (2001). Modelling an opportunistic strategy for information navigation. In *Proceedings of the Cognitive Science Society Annual Conference*, Edinburgh, 2001.

Payne S.J. (1991) Display-based action at the user interface. *International Journal of Man-Machine Studies, 35*, 275-289.

Payne, S.J., Richardson, J. and Howes, A (2000) . Strategic use of familiarity in display-based problem solving. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26*, 1685-1701.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*, 643-675.

Modeling Capabilities and Workload in Intelligent Agents for Simulating Teamwork

Thomas R. Ioerger, Linli He, Deborah Lord
Department of Computer Science, Texas A&M University
College Station, TX 77843-3112
{ioerger,linli,drl4468}@cs.tamu.edu
Pamela Tsang
Department of Psychology, Wright State University
Dayton, OH 45435-0001
(pamela.tsang@wright.edu)

Abstract

The ability of members on a team to reason about each others' capabilities and workload is important for effective teamwork. This is required for proper task allocation and load balancing, as well as many other team processes such as adaptiveness, proactive assistance, and backing-up behavior. The present work proposes to incorporate capability reasoning into intelligent agents to produce better teamwork simulations, to work better with humans as virtual team members, and to facilitate team training. However, classical models of capabilities in computational systems and intelligent agents are inadequate for representing the more complex aspects of human performance, such as the ability to perform multiple tasks in parallel, interference among these tasks, effects of limits on attention and other cognitive resources, and the ability of humans to dynamically adjust their level of effort on tasks. In this paper, we present a formal mathematical model of capabilities that accounts for these effects. The model posits finite pools of internal resources, for which tasks compete; quality of performance depends on the amount of resources allocated. Capabilities are defined according to whether a feasible schedule can be found that allows a set of tasks to be completed within given constraints (e.g. deadlines) while not exceeding the capacity of any internal resource. An extension of the model is then proposed to incorporate multiple resources.

Introduction

Many studies have suggested that the ability to distribute tasks appropriately and to adaptively balance the workload within teams is essential for producing effective teamwork (Kleinman et al., 1992; Kozlowski, 1997). In order to do this, team members must be able to reason about their own and each others' capabilities. For example, they must be able to know when to accept or reject new tasks, based on how they might interfere with current on-going tasks, to delegate sub-tasks to the (best) team members who are not overloaded, and to offer assistance to those who are. Even in intra-team communication and coordination, assessment of capabilities and workload have an impact; in one study, it was found that communications among team members in the best-performing teams actually decreased in high-tempo situations (Serfaty et al., 1997), presumably due to a recognition that excessive communication activities place a demand for attention on both the sender and receiver that competes with processing of intense taskwork. Therefore reasoning about capabilities, including knowledge of task demands, skill levels of individual team members, and momentary workload across the team in a given situation, must be considered an essential component of team cognition.

Recently there has been a rise of interest in incorporating intelligent agents into automated team-training systems (Rickel and Johnson, 1997). These agents could be used in a variety of ways, from automated assistants (decision aids), to virtual role players, to coaches. In order for agents to monitor, understand, critique, or participate in teamwork with human trainees, the agents must also be endowed with the ability to reason about capabilities and workload of individuals on the team. Agents in the simulation must be able to assess the workload of humans with whom they interact in order to make decisions about when and how to interact in a way that is not disruptive or unnatural. (This is an additional constraint that purely agent-based systems do not have to be concerned with.) However, most existing formal models of capability reasoning in agents do not adequately address the kind of reasoning that is required in these agent-based team-training systems. Typically, these prior models treat capabilities as a simple association between actors (agents or humans) and "executable" actions, though the actors must also be aware that they can do these things, i.e. have sufficient "know-how" (Moore, 1985; Singh, 1991; van der Hoek et al., 1994).

These computational models allow agent-based systems to be designed where the agents can reason about each others' capabilities, and even perform task distribution and load balancing. However, these models generally assume task completion is binary (success or failure) and do not take into account graded senses of capability, which are more meaningful to human performance. Humans can often achieve better results by working "harder" (applying more effort or attention), they can dynamically reduce their effort on one task to accomodate per-

forming other tasks in parallel, and they are often limited by pragmatic upper-bounds on performance (e.g. due to finite skills or attention). What is needed is a formal system that will enable an agent to understand when a human is too busy doing certain activities (e.g. flying an aircraft in combat or engaging an enemy) to do other things (e.g. monitor for new visual contacts, listen to background radio traffic). The agent needs to be able to compute the relative impact of new tasks on the accuracy of performing existing tasks, and the potential for delay in completion of individual tasks by their deadlines. This is different from just asking whether an operator is capable of doing the additional tasks "in principle."

Humans are capable of performing multiple tasks in parallel, and there is a great deal of literature on analyzing time-sharing performance (Wickens and Holland, 2000). Yet humans ultimately have limits on their processing capacity, exemplified by the notion of finite limits on attention, which has been rigorously documented. Furthermore, there is clear evidence that some task combinations are time-shared more efficiently than others, such as the difference between drawing a sketch while listening to the radio versus reading while listening to the radio. Some models, such as the multiple-resource model (e.g., Wickens, 1984) have postulated distinct and separate cognitive resources for different types of cognitive processing to explain the wide range of observed task interactions. Another important issue that makes human capabilities difficult to reason about is that performance is not a binary quantity, but rather a graded value (e.g. accuracy, reaction time), and humans can intentionally adjust task performance in a number of controllable ways, such as increasing quality by focusing attention and applying more cognitive "effort," or by reducing effort by spreading the task processing out over a longer interval of time (Hendy et al., 1997), such as multiplying multi-digit numbers together in one's head more slowly for greater accuracy. Therefore, whether or not a human member of a team is "capable" of doing something depends on a great many things, including what other tasks he or she is doing (their current workload), the degree to which the new task might interfere with them, the individual's skill level(s), attention management skill (Gopher, 1993), and the adaptability of the task performance with respect to the tightness of the constraints on completion (e.g. deadlines, quality criteria). This is a more situation-based or context-dependent perspective on capability.

Reasoning about capabilities at this quantitative level is important for modeling and understanding teamwork. To date, very little research has addressed the relationship between individual cognition and quality of teamwork, though the connection is discussed in (Huey and Wickens, 1993). An understanding of individuals' capabilities and workload are clearly important to the efficient operation of a team, such as for distributing tasks to the most appropriate/skillful members, balancing the load (to maintain flexibility), and proactively assisting or backing each other up. With regard to training, we hypothesize that each team member him/herself must develop enough reserve capacity on top of their individual taskwork to devote some attention to participating in the teamwork, such as tracking status or progress of team goals, sharing information relevant to others, or building distributed situation awareness.

In this paper, we present a formal, mathematical model for reasoning about capabilities, especially for agents to reason about and interact with their human teammates. The approach sythesizes ideas from a number of previous descriptions of workload, attention, and performance into a computational model that can be concretely implemented as a decision-making procedure in a multi-agent system. After establishing some terms and assumptions, a definition of capability will be presented in terms of whether a human could adapt his or her performance (i.e. to find a schedule and select effort levels) that would accomodate a given set of tasks with a set of specified constraints. We conclude by discussing the implications of this computational model of capabilities for modeling and understanding team performance, and for designing new approaches to team training.

A Formal Model

In this section, we present a quantitative model for reasoning about capabilities. For simplicity, we start with description of a single-resource model as a basis. Later we extend it to show how it can accomodate the assumption of multiple cognitive resources.

Single-Resource Foundation

Preliminarily, assume there is a single common cognitive resource for which tasks compete. Perhaps it might be labelled generically as "attentional resources." At any given time, a person might be using some amount of this resource, $u(t)$, but the resource is bounded, $\forall t\ 0 \leq u(t) \leq u_{max}$. Since the scale is arbitrary, we normalize resource utilization so that $u_{max} = 1.0$, putting it on a uniform scale of 0 to 1.

We assume this common resource can be allocated to, or divided among, several concurrent tasks. The amount of resource being applied to a given task i at a given moment t is referred to as "effort," and is denoted $e_i(t)$. We view the sum of resources being applied to all tasks at a given moment as a reflection or internal measure of workload. Let the set of tasks be called $\tau_1 ... \tau_i$. The "workload" is defined as:

$$w(t) \equiv \Sigma_i\ e_i(t)$$

and it is constrained not to exceed the limit, $0 \leq w(t) \leq u_{max}$.

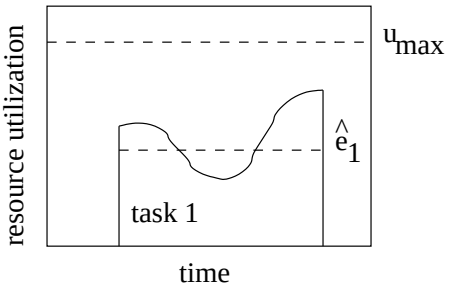


Figure 1: Resource utilization over time (with limit).

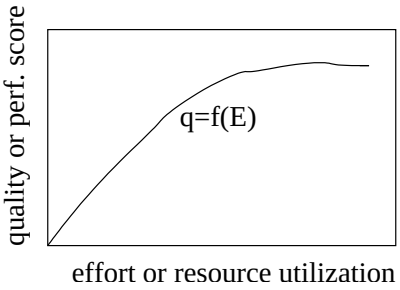


Figure 2: Illustration of a Performance Resource Function.

Whereas this notion of effort is defined on a moment-by-moment basis, the total effort expended on a task, or total resources applied E_i, is the sum of the effort allocated to that task over the duration of its performance (illustrated in Figure 1). In other words, it is the sum, or integral, of the moment-by-moment resources utilized in performing that task:

$$E_i \equiv \int_{t=start(i)}^{t=end(i)} e_i(t)\ dt$$

While the amount of resources applied to a task is not necessarily constant, we assume there is an average effort value $\overline{e}_i$, and our model is based on this approximation.

The amount of resources required for an individual to perform a given task depends on a number of internal and external determinants. Externally, the difficulty of the task, as well as constraints on accuracy or speed (i.e. deadlines), can influence the processing resources required (e.g. it is harder to do a task better or faster, and some tasks have parameters related to difficulty, such as number of items to remember, and so on). Internally, a specific individual's response to task demand can be affected by their innate ability and executive management skill, prior training, degree of automation, etc. We quantify the relationship between amount of resources applied to a task and quality of performance using a function for quality-effort tradeoff: $q_i = f(E_i)$ (also known as a Performance Resource Function (Norman and Bobrow, 1976); see Figure 2).[1] Quality can represent any number of performance measures specific to the task, such as accuracy, inverse of reaction time, etc. We do not place many restrictions on the form of this function, but typically, we assume it is monotonic: increasing effort on most tasks increases quality (Wickens, 1984). (Often, they reach a plateau where greater effort does not improve quality, in which case they are said to be "data-limited.")

[1]Quality of performance might also depend on some measure of task difficulty, which can be treated either by adding an argument for whatever variable parameterizes the degree of difficulty, or by simply viewing them as separate tasks.

Humans can apply the same amount of total effort to a task in a range of different ways. In particular, they might choose to work as hard as possible on the task, completing it in a short amount of time, or they might decide to spread the processing out over time, reducing the moment-by-moment effort, for example to have some reserve capacity left over to apply to other tasks in parallel. Given that we model effort applied as the integral of resource utilization over time, and we assume there is an average level of effort dedicated to a task, the relationship among momentary effort, total effort applied, and duration may be expressed as a simple formula:

$$E_i = \overline{e}_i \cdot d_i$$

where $d_i = end(i) - start(i)$ is the duration of the task. Therefore, the effort-duration tradeoff may be represented as a (presumably) hyperbolic function, and different levels of total effort appear as iso-curves (see Figure 3). Each point on a given curve bounds a box of constant area, representing the common degree of total effort. Harder versions of the same task correspond to curves farther out (dashed line in Figure 3), and improvements in ability, e.g. due to training, appear as curves closer to the origin.

We assume there are range bounds on both duration and effort. Of course, a task can utilize no more than 100% of a resource, and this puts a bound on the minimum execution time (speed), as a result of the hyperbolic function. Similarly, we assume there is a minimum amount of resource required, and a corresponding limit on the slowest effective rate of performance. We represent these ranges as $(d_{i,min}, d_{i,max})$ and $(u_{i,min}, u_{i,max})$.

The performance-resource function is not only a function of task, but also of the individual. We model the difference among individuals by assuming the form of the equation is the same, and applying a multiplier that represents their degree of skill s of individual j (relative to the average of the population, for which we set $\overline{s} = 1$):

$$q_{j,i} = f(E_i) \cdot s_{j,i}$$

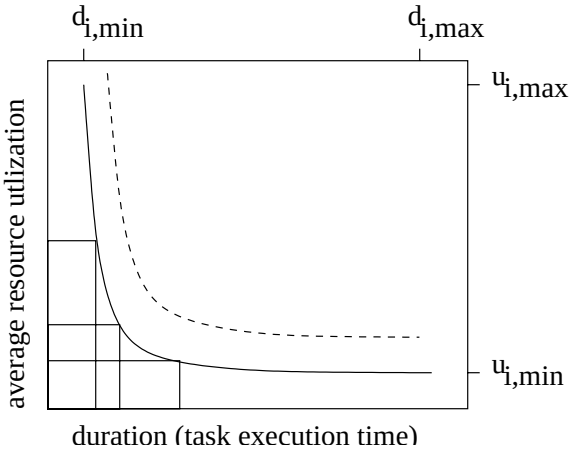


Figure 3: Effort iso-curves.

Hence the greater the skill, the greater the quality of performance for a fixed amount of effort (can be visualized as higher curves in Figure 2). This allows us to model the differences between novices and experts in a simple way.

Capability Assessment as Scheduling

Given this quantitative model of task performance, we can now offer an initial formal definition of "capability." Recall that we are interested not just in whether an individual is capable of doing something "in principle," but also whether it can be carried out effectively in the time allotted and to the level of quality or accuracy required, all within the context of other on-going activities. We view this as a kind of "scheduling" problem, where capability is determined by whether or not the individual can find an arrangement of processing so that all the tasks can be completed without violating any internal capacity limitations.

Definition 1: A *schedule* for a set of tasks $\tau_1 ... \tau_n$ being processed or executed by an individual is a set of parameter vectors $\{\langle start(i), end(i), \overline{e}_i \rangle\}$ defining the start and end times of each task, along with planned average resource utilization to be applied to each.

Definition 2: An individual j who is currently performing a set of tasks $\tau_1 ... \tau_n$, with quality constraints $q_i ... q_n$ and deadlines $dl_1 ... dl_n$ is said to be *capable* of performing a new task τ_{n+1} (with constraints q_{n+1} and d_{n+1}), if there exists a schedule $\mathcal{S}$ over $\tau_1 ... \tau_{n+1}$ defining the start and end times along with average resource utilization of each task $\langle start(i), end(i), \overline{e}_i \rangle$, such that all constraints remain satisfied. Specifically:

1. $q_i \leq f_j(E_i) = f(\overline{e}_i \cdot (end(i) - start(i)))$,
2. $end(i) < dl_i$, and

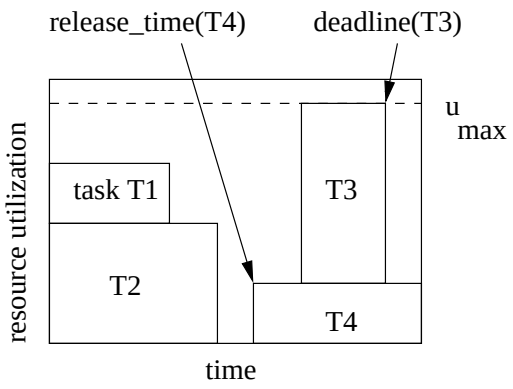


Figure 4: Example of task schedule.

3. $\forall t \; w(t) = \Sigma_i \; \overline{e}_i \leq u_{max}$, where the sum runs over all tasks i s.t. $start(i) \leq t \leq end(i)$.

We note that, among the existing tasks, some may currently be being processed, while others may be scheduled to start after some delay (i.e. pending tasks). If processing of certain currently executing tasks is considered uninterruptible, then we require $start(i) = t_{now}$ for those tasks in the revised schedule to maintain continuity (though the effort level may be modified). Figure 4 shows an example of a task schedule, with selected effort levels and durations, and some representative constraints.

The point of this definition of capabilities is that determination of capability in context must be done flexibly, since there are a wide variety of ways in which performance of tasks can be rearranged to accomodate multiple on-going activities. One primary mechanism is delaying processing of tasks that are not as time-critical. This naturally leads to a scheduling metaphor (Tulga and Sheridan, 1980). Various scheduling algorithms can be drawn from other fields, such as real-time systems. While exact solutions to these problems are often provably intractable, reasonably efficient approximation algorithms often exist (e.g. greedy, earliest deadline first, most-difficult task first, etc.). A major open question is: which approximations seem to correspond to the kinds of heuristics humans use in deciding how to carry out multiple tasks in complex environments?

One unique characteristic of this application of scheduling is that, in addition to manipulating start and end times, another dimension taken into account is the level of effort. In other words, individuals have the option of reducing or increasing their resources allocated to a given task, which can result in a correponding increase or decrease in duration required to produce equivalent performance. Hence, one may decide to defer processing of a new task until the existing ones are complete, or, if there is insufficient time, may decide to begin processing the new task right away by shifting some of their emphasis or at-

tention away from the current tasks, as long as it will not threaten their successful completion.

Using our scheduling-based definition of task performance under resource constraints (both internal and external), we can implement a concrete, computational method for agents to estimate workload of humans and use it to simulate decision-making about when they are likely to accept or reject new tasks in a dynamic environment. Specifically, the model would predict task acceptance if and only if a feasible schedule can be found (at least by a reasonably plausible heuristic method) that would accomodate the new task along with all existing ones, where they would all be completed in time to meet their respective deadlines, and the effort requirements (workload) would not exceed the limits (maximum capacity) of the internal cognitive resource.

A pragmatic issue in developing such a computational method is that the performance characteristics for each individual would need to be derived. We believe that these parameters can be inferred from empirical observations under various controlled conditions by using data-mining techniques, but a detailed description of the methodology is outside the scope of this paper.

Extension to Multiple Resources

The problem with the model as we have presented up to this point is that it is based on a single-resource assumption; thus it cannot account for variable degrees of interaction among tasks of different types. To extend our model to incorporate multiple resouces, we start by assuming that there is a fixed set of resource pools, $r_1 ... r_n$. For example, these might represent the eight components in Wickens' (2000) model, with resources for: auditory input processing, visual input processing, perceptual/central processing, response processing, spatial processing, verbal processing, manual response processing, and speech response processing. Each of these is postulated to be used to different degrees (possibly zero) by any given task.

Thus, instead of a univariate curve for the performace-resource function, we have in principle a function of n dimensions, representing allocation of each resource independently (Tsang and Velasquez, 1996). However, to keep the model manageable (and for parsimony), we instead use a "profile" for each task to represent, under single-task, full-attention conditions, the relative amounts of each resource required: $\langle u(r_1), ..., u(r_n) \rangle$ (as illustrated in Figure 5). Then each resource level is modified proportionally based on what fraction of 100% attention, $att(\tau)$, is allocated to a given task τ in a specific situation, effectively parameterizing the resource demands. Hence the task demand, distributed over the various resource components, becomes:

$$\langle att(\tau) \cdot u(r_1, \tau), ..., att(\tau) \cdot u(r_n, \tau) \rangle$$

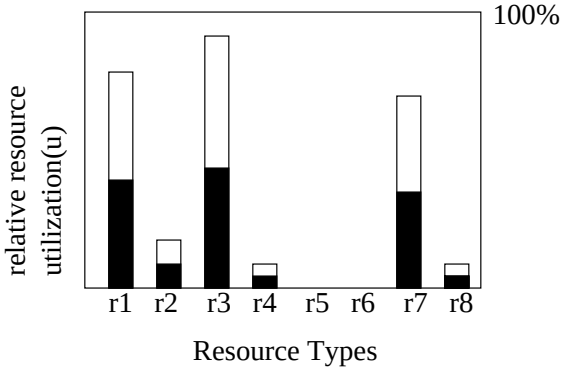


Figure 5: Resource-demand profile for a hypothetical task. The open bars represent the relative amount of demand by a task on each type of internal (cognitive) resource when it is being given full attention. The solid bars show resource allocation levels scaled down proportionally for a case where an individual is only able to focus half of the requisite attention on this particular task.

This approach allows us to treat task performance as intrinsically multi-dimensional. Now the rest of the model (including scheduling) can be applied as before, with the condition that, regardless of how the tasks combine or overlap, no individual resource may exceed its capacity at any point in time:

$$\forall t \; w_k(t) = \Sigma_i \; att(\tau_i, t) \cdot u(r_k, \tau_i) \leq u_{max}(r_k)$$

for all resource pools k, where i is summed over all tasks being executed at time t. The task difficulty and quality (i.e. accuracy) requirements set the level of effort required for the individual, the individual chooses a suitable duration and corresponding level of average emphasis to apply, and then this is used to compute utilization of each resource based on a scaled version of the single-task utilization profile. To determine whether an individual is capable of accepting a new task in the context of existing ones, a schedule must be sought that allows all tasks to complete within the time and quality constraints, while violating no limits on internal resources.

The primary benefit of this multi-dimensional model is that it can be used to simulate different degrees of interference among tasks depending on their type. For example, even though tasks A, B, and C are considered equally demanding, it might be more efficient to process A and C in parallel than A and B. This effect could be captured by saying that the profiles for A and B both share high demand for the same underlying resource, while the components utilized by A and C are relatively distinct. The phenomenon of differential interference based on task type has been called "structural similarity" in the literature (Wickens and Holland, 2000). Our work

is intended to form a preliminary basis for theoretical and empirical modeling of this effect.

Discussion

Capabilities and workload are one part of the "shared mental model" that must be computed, along with others' beliefs, goals, situations, etc., to generate believable simulations of teamwork. This model could be applied to enhancing the simulation and generation of teamwork by influencing role selection, delegation, negotiation, and pro-active behavior. For example, responsibilities could be re-defined to take into account the degree to which one is capable of doing something, delegation policies and task allocation strategies could be modified to reflect an awareness of individuals' workload (i.e. to select a member for whom it would least interfere), and agents could adjust their initiative in offering to help team members with tasks based on an assessment of how over-loaded they are versus how much of a distraction it would be.

An important application of this computational model of capabilities could be for designing intelligent agents for use in team-training systems. Specifically, this model would allow agents to monitor, exercise, and evaluate individuals' ability on human teams to appropriately and effectively participate in the teamwork, as a function of their own skills, workload response, and attention management strategies. The goal would be the development of novel training interventions that could promote the balance of the cognitive demands of taskwork versus teamwork (i.e., spending time reasoning about each other).

Acknowledgments

This work was supported in part by MURI grant #F49620-00-1-0326 from DoD and AFOSR.

References

Gopher, D. (1993). The skill of attention control: Acquisition and execution of attention strategies. In Meyer, D. and Kornblum, S., editors, *Attention and Performance XIV*. Cambridge, MA: MIT Press.

Hendy, K., Liao, K., and Milgram, P. (1997). Combining time and intensity effects in assessing operator information-processing load. *Human Factors*, 39:30–47.

Huey, M. and Wickens, C. (1993). *Workload transition: Implications for individual and team performance*. Washington, D.C.: National Academy Press.

Kleinman, D., Luh, P., Pattipati, K., and Serfaty, D. (1992). Mathematical models of team performance: A distributed decision-making approach. In Sweezy, R. and Salas, E., editors, *Teams: Their Training and Performance*. New York: Ablex.

Kozlowski, S. (1997). Training and development of adaptive teams. In Cannon-Bowers, J. and Salas, E., editors, *Making Decisions Under Stress*, pages 115–153. Washington, D.C.: American Psychological Association.

Moore, R. (1985). A formal theory of knowledge and action. In Hobbs, J. and Moore, R., editors, *Formal Theories of the Commonsense World*. Norwood, NJ: Ablex.

Norman, D. and Bobrow, D. (1976). On the analysis of performance operating characteristics. *Psychological Review*, 83:508–510.

Rickel, J. and Johnson, W. (1997). Intelligent tutoring in virtual reality: A preliminary report. In *Proc. of the International Conference on Artificial Intelligence in Education*, pages 294–301.

Serfaty, D., Entin, E., and Johnston, J. (1997). Team coordination training. In Cannon-Bowers, J. and Salas, E., editors, *Making Decisions Under Stress*, pages 221–245. Washington, D.C.: American Psychological Association.

Singh, M. (1991). A logic of situated know-how. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 343–348.

Tsang, P. and Velasquez, V. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39:358–381.

Tulga, M. and Sheridan, T. (1980). Dynamic decisions and workload in multitask supervisory control. *IEEE Transactions on Systems, Man, and Cybernetics*, 10(5):217–232.

van der Hoek, W., van Linder, B., and Meyer, J. C. (1994). A logic of capabilities. In *Lecture Notes in Computer Science*, volume 813, pages 366–378. Springer-Verlag.

Wickens, C. (1984). Processing resources in attention. In Parasuraman, R. and Davies, D., editors, *Varieties of Attention*. New York: Academic Press.

Wickens, C. and Holland, J. (2000). Attention, timesharing, and workload. In Wickens, C. and Holland, J., editors, *Engineering Psychology and Human Performance*, chapter 11. Upper Saddle Reiver, NJ: Prentice Hall, 3rd edition.

Self-Organizing Recognition and Classification of Relational Structures

Brijnesh J. Jain (bjj@cs.tu-berlin.de)
Department of Computer Science; Technical University Berlin
Germany

Fritz Wysotzki (wysotzki@cs.tu-berlin.de)
Department of Computer Science; Technical University Berlin
Germany

Abstract

We present a novel self-organizing structure recognition (SOSR) network for classification and recognition of relational structures represented by graphs. The system consists of several subnets each comparing an input structure with a given model structure. The subnets are indirectly coupled via a winner-take-all (WTA) classifier. During classification the SOSR system deactivates subnets which indicate large dissimilarities between the input structure and the corresponding models. First experiments show that this mechanism significantly reduces the computational effort in comparison to traditional classification systems using a comparative maximum selector as a classifier.

Introduction

We describe a hierarchical neural net for the recognition and classification of relational structures by matching with class prototypes which was primarily developed from a theoretical point of view and for practical applications in Artificial Intelligence. Classification by means of prototypes is well known in the psychological literature (e.g. Rosch, 1975; Rosch and Lloyd, 1978) but usually is modeled using feature vectors as description of objects and prototypes, respectively. In the context of modeling semantic memory and discussion of the binding problem in Cognitive Neuroscience relational descriptions and representations of structured objects play nowadays a major role (e.g. Hinton, 1994; Taylor, 1993; Taylor, 1996; Singer, 2000). Seen from the point of view of modeling the dynamics of neural structures in connection with psychologically observed behavior we are not primarily interested in the neural (population or assembly) code of representing relations (e.g. Singer, 2000) but in studying the processing strategies using symbolic descriptions of objects and prototypes by graphs and a hierarchical organized *winner-takes-all* (WTA) net. This net will classify objects by competitive matching with a set of predefined prototypes in a *self-organizing manner*, i.e. without a *homunculus* acting as a supervisor. The investigation of the WTA-processing strategies might also shed light on principles of functioning of the Short-Term-Memory (e.g. Grossberg, 1987a; Grossberg, 1987b), on the role of attention (Lee et al., 1999), and on a trade-off between accuracy vs. speed of recognition depending on the strength of inhibition as shown in our first experimental results given below.

In Artificial Intelligence and Image Recognition graphs are a well suited representation of relational structures like molecular structures, data structures, or semantic networks. In any case, a relational structure consists of elementary objects and binary relations between these objects. In a graph of a relational structure the elementary objects are represented by vertices and their relations by directed or undirected edges. For example, in chemistry, graphs model molecular structures where the vertices represent atoms and the edges represent bonds. In Computer Vision vertices of a graph are objects within a scene and edges are structural relationships between those objects.

A fundamental problem in many application domains of processing relational structures is the identification and recognition of common structural parts between two relational structures. For example in classification, recognition or clustering tasks, information about structural overlaps between two structures is required in order to determine a similarity or distance of these structures. Here we call the computation of a similarity or distance between two relational structures *graph matching*.

In general graph matching problems are well-known NP-complete problems (Garey & Johnson, 1979). Due to the high computational complexity much effort has been directed toward devising efficient heuristics to find optimal or approximate solutions for graph matching problems. Among other heuristics artificial neural networks have been proposed as a promising model of computation for solving graph matching problems (Schädler & Wysotzki, 1999).

The high computational complexity is even more inconvenient if the solution of a problem requires several graph matching procedures. In distance-based classification using neural networks an input graph G is matched against a given set of N model graphs $M_1, \ldots, M_N$ representing prototypes of category $C_1, \ldots, C_N$, respectively. The matching is performed by recurrent neural networks $S_1, \ldots, S_N$. In

the following we will call these networks S_k *subnets*.

One fundamental approach in distance-based classification of structures is the classification by means of a *discriminant function* and a *comparative maximum selector* (CMS) classifier. A classification task is solved by a CMS approach in the following chronological order (Schädler & Wysotzki, 1999): (1) A feature extractor computes the discriminants ϱ_k of input G and each model M_k. The discriminant ϱ_k is computed by the k-th subnet S_k and serves as a measure for the similarity between G and M_k. (2) The discriminants are passed to a CMS classifier. (3) The CMS classifier sequentially compares the calculated discriminants and assigns the input graph to the category for which the corresponding discriminant is largest. Thus a CMS classifier processes its incoming data like a supervising *homunculus*.

In the traditional CMS classifier approach each match between G and M_k has equal priority in the sense, that each subnet S_k evolves until it has computed a discriminant ϱ_k for G and M_k. Thus a CMS classifier completes the evolution of all N subnets although the internal states of some subnets may indicate high dissimilarity of the graphs to be compared at an *early stage* during the matching process.

In order to improve the computational performance of a classifier for relational structures and to investigate the effects of self-organization in hierarchical networks we present a *self-organizing structure recognition* (SOSR) network. To accomplish a better computational performance than the CMS classifier the SOSR model identifies dissimilar pairs G and M_k at an early stage of the matching process and aborts the computation of the corresponding subnets S_k. Thus the SOSR network focuses on promising subnets and neglects subnets indicating high dissimilarity between the input and the corresponding model. This mechanism is realized by replacing the CMS by an inhibitory WTA network for the maximum selection and intertwining the tasks of step (1)-(3). Further improvements to reduce the computational effort can be made by parallel processing which is not possible with a sequential CMS classifier.

Note, that the SOSR model has much in common with the *competitive relational mind model* proposed by Taylor (1996). Furthermore, Grossberg (1987a) uses in a similar way a network for competitive learning with a reset mechanism which deactivates subnets announcing high dissimilarity to allow them to rebuild the model, i.e. the expectation.

The rest of this paper is organized as follows: We conclude this introductory section with some basic notations and definitions. In the next Section we formulate the graph matching problem in terms of the maximum clique problem. Subsequently we describe the SOSR network architecture. In an empirical study we investigate the behavior of the SOSR system. Finally, the last section summarizes this contribution.

Notations and Definitions: A *graph* is a pair $G = (V, E)$ consisting of a finite set $V \neq \emptyset$ of *nodes* and a binary relation $E \subseteq V^2 := \{(i,j) \mid i,j \in V,\ i \neq j\}$. The elements $(i,j) \in E$ are called *edges*. A *subgraph* $H = (V_H, E_H)$ of G is a graph with $V_H \subseteq V$ and $E_H \subseteq V_H^2 \cap E$. An *induced subgraph* H of G is a subgraph with $E_H = V_H^2 \cap E$. A graph G is called *complete*, if $E = V^2$. A complete subgraph $C \subseteq G$ is called *clique* of G. A *maximum clique* $C \subseteq G$ is a clique with maximum number of vertices. A *maximal clique* is a clique which is not contained in any larger clique. The *clique number* $\omega(G)$ of a graph G is the number of vertices of a maximum clique in G. The *size* n_V of a graph G is the number $|V|$ of its vertices.

Graph Matching and Maximal Cliques

The graph matching problem is the problem to find the best partial mapping between two graphs where the quality of the mapping is estimated in terms of a problem dependent objective function. Our SOSR approach can also be applied to inexact graph matching problems of colored graphs, where vertex colors represent elementary objects and edges colors represent the type of relation between theses objects. But for convenience we only consider the common maximum induced subgraph problem which comprises the graph isomorphism and subgraph isomorphism problem as special cases.

A common approach to solve these classes of graph matching problems is based on maximum clique detection in an *association graph* (Ballard & Brown, 1982)). The association graph is formed by creating vertices $\boldsymbol{i}$ from each pair of vertices $(i_1, i_2) \in V_1 \times V_2$ and inserting edges between vertices $\boldsymbol{i} = (i_1, i_2)$ and $\boldsymbol{j} = (j_1, j_2)$ if (i_1, j_1) and (i_2, j_2) are edges in G_1 and G_2, respectively.

By definition of an association graph the cliques of $A = A(G_1, G_2)$ are in 1-1 correspondence to common isomorphic induced subgraphs of G_1 and G_2 and the maximum cliques uniquely correspond to the common maximum induced subgraphs of G_1 and G_2. This maps the graph matching problem to the optimization problem of finding a maximum clique C in A where the discriminant $\varrho(A)$ is a function on the number of vertices of C.

The SOSR Model

In the following let $A_k = (V_k, E_k)$ be the association graph of input G and model M_k $(1 \leq k \leq N)$, where M_k represents category C_k. With ϱ_k we denote the discriminant of A_k.

The SOSR model consists of two interconnected layers, a feature extractor layer and a classifier layer (see Figure 1). The feature layer contains N subnets S_k each comparing input G with model M_k. The

classifier layer is a competitive WTA network for the maximum selection consisting of N inhibitory connected units where unit c_k represents category C_k. Each subnet S_k of the feature extractor is connected to unit c_k of the WTA classifier via an *inter-unit* i_k.

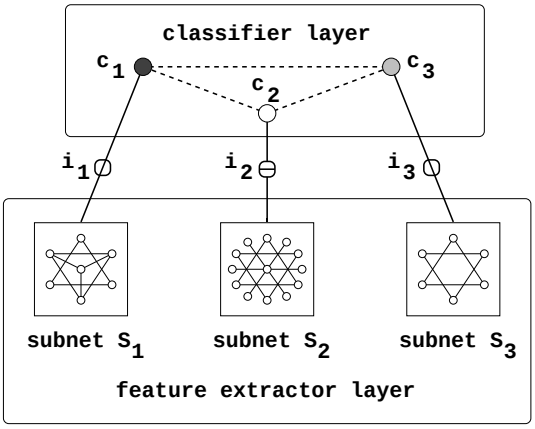


Figure 1: Architecture of a SOSR network.

During classification the subnets S_k evolve synchronously and continuously pass their current internal states to the inter-units i_k. The inter-units compute interim values $\varrho_k(t)$ of the discriminants ϱ_k and transfer them to the WTA classifier. The WTA classifier evaluates the evidence presented for a decision at an early stage. If the activation $z_k(t)$ of unit c_k in the classifier layer falls below zero, the WTA network disconnects unit c_k from inter-unit i_k, such that subnet S_k is excluded from the competition.

Figure 1 depicts a functional diagram of the SOSR model for the $N = 3$ category problem. The shading of units c_k in the classifier layer indicates their output level where darker shading means a higher output. Thus unit c_1 is dominating while unit c_2 has the lowest activation. Inter-units i_k are depicted as switches. A subnet S_k participates in the competition if the connection between S_k and c_k is switched on. Otherwise, S_k is considered to be irrelevant for the decision making process and inter-unit i_k switches off the connection to exclude S_k from the competition. In figure 1 inter-unit i_2 disconnects subnet S_2 from the classifier layer.

In mathematical terms the following equations describe the behavior of the SOSR model:

$$\dot{x}_i^k(t) = -dx_i^k(t) + \sum_j w_{ij} f\left(x_j^k(t)\right) \quad (1)$$

$$y_k(t) = \left[\left[\varrho\left(\boldsymbol{x_k}(t)\right) \right]_0^1 + \varepsilon - s_\varepsilon\left(-z_k(t)\right) \right]_0^1 \quad (2)$$

$$\dot{z}_k(t) = -d' z_k(t) - w \sum_{j \neq k} \left[z_j(t) \right]_0 + y_k(t) \quad (3)$$

where $[x]_\theta^\vartheta := \max\{\theta, \min\{x, \vartheta\}\}$ is the limiter function with lower and upper bound $\theta < \vartheta$, $[x]_\theta := \max\{\theta, x\}$ is the linear threshold function with lower bound θ, and s_ε is a trigger function of the form

$$s_\varepsilon(x) = \begin{cases} 1+\varepsilon & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

Equation (1) describes the dynamics of an additive recurrent subnet S_k in the feature extractor, equation (2) describes the behavior of inter-unit i_k connecting subnet S_k with unit c_k of the WTA net, and equation (3) specifies the WTA dynamics. The system terminates if only a single unit c_{k_*} in the WTA classifier is left with an activation $z_{k_*}(t) > 0$ while all other units c_k are inhibited, i.e. $z_k(t) \leq 0$. Under the assumption of converging subnets termination follows from (Jain & Wysotzki, 2001a; Wersing & Beyn & Ritter, 2001).

Equation (1): A Maximal Clique Solver

Let index k refer to subnet S_k solving the maximum clique problem for graph A_k.

Many different neural network approaches and techniques have been proposed to solve the maximum clique problem (Bomze et al., 1999). As a representative model we consider a general additive recurrent network (1) where $x_i^k(t)$ denotes the activity of unit i of subnet k at time t and the constant $d \in [0, 1]$ describes the self-inhibition. The strength of the connection between unit i and unit j is determined by the synaptic weight $w_{ij} = w_{ji}$. The output of each unit is computed by a non-decreasing bounded function f.

In order to solve the maximum clique problem of the k-th association graph $A_k = (V_k, E_k)$ the network consists of $|V_k| = n_k$ units which are connected with weight $w_{ij}^k = w_E^k > 0$ if $(i, j) \in E_k$ is an edge in A_k and with weight $w_{ij}^k = -w_I^k < 0$ if $(i, j) \notin E_k$. Self weights w_{ii}^k are set to zero.

Given appropriate parameter settings the maximal clique solver operates as follows (Schädler & Wysotzki, 1998; Schädler & Wysotzki, 1999): An initial activation is imposed on the network. Finding a maximum clique then proceeds in accordance with equation (1) until the system reaches a stable state. The stable states correspond to the maximal cliques of A_k. In the ideal case a maximum clique is found. The clique size can be read out by counting the number of units with activation $x_i^k(t) \geq 1$.

For our experiments we used the time-discrete approximation of (1) given by

$$x_i(t+1) = (1-d)x_i(t) + \sum_{j \neq i} w_{ij} f_T\left(x_j(t)\right) + \eta(t) \quad (4)$$

where $\eta(t)$ is a small random noise to dissolve ambiguities and f_T is a controllable limiter function of the form

$$f_T(x) = \begin{cases} 1 & : \quad x \geq T \\ 0 & : \quad x \leq 0 \\ x/T & : \quad \text{otherwise} \end{cases}$$

with control parameter T which we call the *pseudo-temperature*. Starting with a high inititial value for $T = T_0$ the pseudo-temperature T is decreased during the evolution of the network according to an annealing schedule until it reaches a sufficient low final value $T = T_f$. Applying an annealing scheme avoids that the system gets stuck in spurious minima. The annealing schedule is of the following form:

1. Intialize $T \leftarrow T_0$.
2. Let the network iterate τ steps according to the dynamical rule given in (4).
3. Decrease the pseudo-temperature by $T \leftarrow a \cdot T$ where $0 < a < 1$ is an annealing constant.
4. If $T > T_f$ continue with step 2. Otherwise terminate the algorithm.

The general parameter settings of the subnets S_k follows a theoretical analysis given in Jain & Wysotzki (2002). For the weights we set

$$\begin{aligned} w_E^k &= \frac{2}{\deg_I^k \cdot (n_k - \deg_E^k)} \\ w_I^k &= \deg_E^k \cdot w_E^k \end{aligned}$$

where n_k is the number of units of subnet S_k and $\deg_E^k$ ($\deg_I^k$) is the maximum number of excitatory (inhibitory) connections of an unit in S_k.

Equation (2): Inter-Units:

An inter-unit i_k connects subnet $S_k = (V_k, E_k)$ with unit c_k in the WTA classifier and controls the external input $y_k(t)$ of c_k.

Let $\boldsymbol{x_k}(t)$ be the current state vector of subnet S_k. Inter-unit i_k receives $\boldsymbol{x_k}(t)$ as its input and computes an interim value $\varrho_k(t) = \varrho(\boldsymbol{x_k}(t))$ of the discriminant ϱ_k. The computation of interim values $\varrho_k(t)$ is constrained to

$$\varrho_k(t) \leq \varrho_k \tag{5}$$

where equality holds if and only if S_k is in a stable state corresponding to a maximum clique. The discrimination value ϱ_k measures the resemblance between input G and model M_k. Thus the interim values $\varrho_k(t)$ reflect a preliminary estimate of the discriminant ϱ_k where the degree of resemblance of G and M_k gradually emerges with the time spent on the matching problem. This is indicated by (5). At beginning an interim value $\varrho_k(t)$ is at a low level and with increasing time $\varrho_k(t)$ approaches ϱ_k. During evolution of S_k it is not required that $\varrho_k(t)$ is monotonously increasing with the time. We call local maxima and minima of $\varrho_k(t)$ *deceptions*. Deceptions may leed to misclassifications. A local minimum of $\varrho_k(t)$ may result in an exclusion of S_k from the competition. In this case the resemblance of G and M_k is underestimated during the comparision of the structures G and M_k. Similarly, a local maximum of $\varrho_k(t)$ may result in a premature decision which assigns G to category C_k. Here the match of input G and model M_k is overestimated during computation. Another source of misclassifications arise from insufficient synchronization among the evolving neural maximal clique solvers S_k. Here, too fast (too slow) convergence of $\varrho_k(t)$ to a low (high) discriminant ϱ_k can lead to an erroneous decision. Thus it is an important objective to design the computation of $\varrho_k(t)$, such that deceptions are avoided and the matching procedures of the subnets S_k are synchronized.

Depending on its current state $y_k(t)$ an inter-unit i_k transfers the interim value $\varrho_k(t)$ to unit c_k of the WTA classifier. An inter-unit i_k is in state ON if $y_k(t) > 0$ and in state OFF if $y_k(t) \leq 0$. Only inter-units in state ON pass interim values to the WTA classifier. Initially, the state of each inter-unit is ON where $\varepsilon > 0$ is a small constant to avoid a deactivation of i_k at the beginning. According to (3) let $z_k(t)$ be the activation of unit c_k in the WTA classifier. Unit c_k switches OFF inter-unit i_k if $z_k(t) \leq 0$ and does not effect i_k otherwise. This mechanism is realized by the trigger function s_ε. If $z_k(t) \leq 0$ then $-z_k(t) \geq 0$ and thus the value $s_\varepsilon(-z_k(t)) = 1 + \varepsilon$ is subtracted from $\phi_k(t) + \varepsilon \leq 1 + \varepsilon$. This sets the activation level $y_k(t)$ of inter-unit i_k equal to 0. Similarly, if $z_k(t) > 0$ the trigger function s_ε has no influence on the activation $y_k(t)$ of inter-unit i_k. Once in state OFF an inter-unit can not be reactivated. In practical applications the corresponding subnets S_k can be switched off.

Next we give an example how to compute the interim values $\varrho_k(t)$ on the basis of a family of discriminant functions

$$\varrho_k = \frac{\alpha|V_{C_k}| + \beta|E_{C_k}|}{\mu_k}$$

where $\alpha, \beta \geq 0$ are problem specific constants which weight the vertex and edge matches, $C_k = (V_{C_k}, E_{C_k})$ is a maximum clique of A_k, and $\mu_k > 0$ is a normalization constant to ensure an upper bound of 1. Note, that $1 - \varrho_k$ defines a family of distance metrics (Schädler & Wysotzki, 1999). We define $\varrho_k(t)$ to be a function on the number of θ-*active units*:

1. Let $V_k^\theta \subseteq V_k$ be the set of θ-active units with activation $x_i^k(t) \geq \theta$ where $0 \leq \theta \leq 1$ is a threshold. Compute the current θ-*intensity*

$$\sigma_V^\theta(t) = \sum_{i \in V_k^\theta} \left[x_i^k(t)\right]_\theta^1 \tag{6}$$

2. Let E_k^θ be the set of all excitatory connections (i, j) between θ-active units $i, j \in V_k^\theta$. Compute

the current θ-*connective-intensity*

$$\sigma_E^\theta(t) = \sum_{(i,j)\in E_k^\theta} \left[x_i^k(t)\right]_\theta^1 + \left[x_j^k(t)\right]_\theta^1 \qquad (7)$$

3. Let I_k^θ be the set of all inhibitory connections (i,j) between θ-active units $i,j \in V_k^\theta$. Compute the current θ-*incompatibility*

$$\sigma_I^\theta(t) = \sum_{(i,j)\in I_k^\theta} \pi_{ij}^k(t) \qquad (8)$$

where $\pi_{ij}^k(t) \geq 0$ is a penalty term for the presence of inhibitory connected active units.

4. Compute the current interim value

$$\varrho_k(t) = \frac{\alpha\sigma_V^\theta(t) + \beta\sigma_E^\theta(t)}{\mu_k} - \sigma_I^\theta(t) \qquad (9)$$

as a function of the current θ-intensity, the current θ-connective-intensity, and the current θ-incompatibility.

The appropriate choice of $\pi_{ij}^k(t)$ is crucial to synchronize the subnets and to avoid deceptions. Too high or too low values of $\pi_{ij}^k(t)$ result in a higher percentage of misclassifications.

Note, that in a stable state of S_k the interim values $\varrho_k(t)$ are identical to the discriminant ϱ_k defined by a corresponding maximal clique C_k.

Equation (3): A WTA Classifier:

The dynamical system given by equation (3) is an inhibitory WTA network for the maximum selection from a set of external input signals where unit k represents category C_k, $z_k(t)$ is the activation of unit c_k, $-w < 0$ represents the inhibitory strength of the synapses connecting any pair of units, $d' > 0$ is the self-inhibition, and $y_k(t)$ is the external input from inter-unit i_k.

Given an initial input vector the operation of these networks is a mode of contrast adjustment and pattern normalization where only the unit with the highest activation fires and all other units in the network are inhibited after some setting time (Jain & Wysotzki, 2001a; Jain & Wysotzki, 2001b).

In the context of the relational mind model of Taylor (1996) the WTA classifier may be seen as the analogue of the thalamic NRT-C complex acting as a central executive for global competition.

Experiments

In first experiments we focused on the performance of the proposed SOSR model and on the role of the inhibition $-w$.

To keep the experiments simple we considered the more general task of identifying a graph A_{k_0} with maximal clique number ω_{k_0} among a set of $N = 10$ graphs $A_1, \ldots, A_N$ of identical size $n_V := |V_1| = \cdots = |V_N|$. For a single run of the algorithm with fixed size n_V the clique numbers $\omega_1, \ldots, \omega_N$ of the N graphs were drawn from a Gaussian distribution restricted to the interval $[3, n_V - 1]$ with identically distributed random mean $3 \leq E[\omega] < n_V$ and variance $0 < Var[\omega] < n_V/2$. The chosen sizes are $n_V = \{25, 50, 100, 250\}$. To assess the effects of the inhibition $-w$ we varied the weight $w = \{20^{-1}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. We performed 100 runs of the SOSR algorithm for each chosen n_V and w.

n_v / t_{MS}	25 / 180.1		50 / 341.3	
w	κ	t_{SO}	κ	t_{SO}
0.2	1.00	18.8	0.95	33.0
0.1	1.00	19.2	0.95	44.0
0.01	1.00	75.2	0.99	97.5
0.001	0.99	167.3	0.94	251.1
0.0001	0.99	178.7	0.85	323.7

Table 1: Classification accuracies κ and number of iterations t_{SO} for size $n_V = 25$ and $n_V = 50$.

Parameter Settings: We have chosen $T_0 = w_E \cdot \deg_E^2 + w_I \cdot \deg_I^2$, $T_f = 0.1$, $a = 0.75$, and $\tau = 0.6 \cdot n_V$ as parameters for the annealing schedule of the subnets S_k. The self-inhibitions d and d' of the subnets S_k and the WTA classifier are set to 0.

To compute the discriminants we set $\alpha = 0$, $\beta = 1$, $\theta = 0$, $\varepsilon = 0.1$, and $\mu_k = \varepsilon^{-1} \cdot \omega_{k_0}^2$ for all k where ω_{k_0} is the maximal clique number of a sample $\omega_1, \ldots, \omega_N$. The θ-incompatibility $\sigma_I^\theta(t)$ is defined by the heuristically chosen penalties $\pi_{ij}^k(t) = \frac{4}{n_V}\left([x_i^k(t)]_\theta^1 + [x_j^k(t)]_\theta^1\right)$.

Results: Table 1 and 2 summarizes the results. The first row shows the size n_V and the average iterations t_{MS} of a subnet *without switching* OFF, averaged over all N subnets and all 500 runs for size n_V. The other entries show for each size n_V the rate of correct classifications κ and the average iterations t_{SO} of a subnet *with switching* OFF, averaged over all N subnets and all 100 runs for a given weight $-w$ and size n_V. For example a subnet S_k consisting of 150 units in a 10-category SOSR system with $w = 0.01$ averages 230.8 iterations until it terminates. In a CMS classifier system S_k averages 993.7 iterations. The classification accuracy of the SOSR system for this configuration is 0.92.

Discussion: If the inhibition $-w$ in the WTA classifier is on a low level an increase will tend to more accurate and faster generated responses. But when

n_v / t_{MS}	100 / 571.0		150 / 993.7	
w	κ	t_{SO}	κ	t_{SO}
0.2	0.89	45.5	0.83	140.0
0.1	0.94	49.9	0.84	130.3
0.01	0.94	126.4	0.92	230.8
0.001	0.87	361.9	0.92	604.3
0.0001	0.74	543.7	0.67	882.0

Table 2: Classification accuracies κ and number of iterations t_{SO} for size $n_V = 100$ and $n_V = 150$.

inhibition is at a high level, the increased competition may interfere with correct decisions by trapping into deceptions or by immobility in finding any response. Thus there will be an optimal level of inhibition for effective behavior. The optimal setting of w compromises the conflicting tasks of significantly improving the computational efforts and gaining high classification accuracy.

Conclusion

We introduced a new self-organizing structure recognition system. The architecture couples the subnets in the feature extractor to a WTA classifier such that the subnets are in an indirect competition during the matching process. The system consecutively switches off subnets if their interim values indicate a *worse* match than the remaining active subnets and shifts its attention to more promising subnets. In first experiments we showed that the switching-off mechanism of the SOSR system significantly reduces the computational effort without suffering in substantial losses of classification accuracy. The inhibition $-w$ of the WTA classifier controls the conflicting interests of high classification accuracy and fast decision making.

Ballard, D.H. & Brown, C.M. (1982). *Computer Vision.* Englewood Cliffs, NJ: Prentice Hall.

Bomze, I.M. & Budinich, M. & Pardalos, P.M. & Pelillo, M. (1999). The maximum clique problem. In D.-Z. Du & P.M. Pardalos (Eds.), *Handbook of Combinatorial Optimization.* Boston, MA: Kluwer Academic Publishers.

Garey, M. & Johnson, D. (1979).*Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman and Company.

Grossberg, S. (1987a). Competitive Learning: From interactive activation to adaptive resonance. *Cognitive Science, 11*, 23-63.

Grossberg, S. (1987b). *The Adaptive Brain.* Elsevier: Amsterdam.

Hinton, G. (1994). Wie Neuronale Netze aus Erfahrung lernen. In W. Singer (Ed.), *Gehirn und Bewusstsein.* Spektrum.

Jain, B.J. & Wysotzki, F. (2001a). On the short-term-memory of WTA nets. In M. Verleysen(Ed.), *Proceedings of the Ninth European Symposium on Artificial Neural Networks* (pp. 289-294). Brussels, Belgium: D-Facto.

Jain, B.J. & Wysotzki, F. (2001b). Efficient Pattern Discrimination with Inhibitory WTA Nets. In G. Dorffner, H. Bischof, K. Hornik (Eds.), *Artificial Neural Networks - ICANN 2001* (pp. 827-834). Springer.

Jain, B.J. & Wysotzki, F. (2002). Parameter Settings of a Neural Network for the Maximum Clique Problem. *To appear.*

Lee, D. & Itti, L. & Koch, C. & Braun, J. (1999). Attention activates winner-take-all competition among visual filter. *Neural Neuroscience, 2*, 375-381.

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General, 104,* 192-233.

Rosch, E. & Lloyd, B. (1978). *Cognition and Categorization.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schädler, K. & Wysotzki, F. (1998). Application of a neural net in classification & knowledge discovery. In M. Verleysen(Ed.), *Proceedings of the sixth European Symposium on Artificial Neural Networks* (pp. 117-122). Brussels, Belgium: D-Facto.

Schädler, K. & Wysotzki, F. (1999). Comparing Structures Using a Hopfield-Style Neural Network. *Applied Intelligence, 11*, 15-30.

Singer, W. (2000). Response Synchronization: A Universal Coding Strategy for the Definition of Relations. In M.S. Gazzaniga (Ed), *The New Cognitive Neurosciences.* MIT Press.

Taylor, J.G. (1993). A Global Gating Model of Attention and Consciousness. In M. Oaksford & G. Brown (Eds.), *Neuro-Dynamics and Psychology.* New York: Academic Press.

Taylor, J.G. (1996). Modelling what is like to be. In S.R. Hameroff et al. (Eds.), *Toward a Science of Consciousness.* MIT Press.

Wersing, H. & Beyn, W.-J. & Ritter, H. (2001). Dynamical Stability Conditions for Recurrent Neural Networks with Unsaturating Piecewise Linear Transfer Functions. *Neural Computation, 13*, 1811-1825.

Integrating Perceptual Organization and Attention: A New Model For Object-Based Attention

Jerzy P. Jarmasz (jjarmasz@ccs.carleton.ca)
Cognitive Science Program and Centre for Applied Cognitive Research,
Carleton University, Ottawa, Canada, K1R 7W9

Abstract

Recent research shows that, under certain conditions, visual attention is object-based. That is, attention preferentially selects objects in the visual field. These objects are processed, culminating in object recognition. On this formulation, the objects selected by attention are perceptual groups determined by the principles of perceptual organization of Gestalt psychology. These groups are formed independently of attentional processes and conceptual knowledge. This view is not consistent with available data about the visual system, which shows that perceptual organization is sensitive to conceptual information, depends on attentional processes, and infers representations that best explain the visual stimulus. Here, I propose a new account of visual attention that aims to correct these limitations of the Gestalt-based formulation. The nature of the object representations underlying perceptual and attentional mechanisms is discussed. It is proposed that attention and perception interact in an iterative process wherein constraints imposed both by the visual stimulus and an observer's cognitive set determine the "objects" to which attention is allocated. Thus, visual attention is object based precisely because it is intricately involved in perceptual organization, and not because it selects the output of perceptual organization, as is generally claimed. Experimental results that support the claim that attention influences perceptual organization are reviewed. Finally, the implications for human factors research and the metaphysics of everyday objects are discussed.

Introduction

Vision is generally assumed to have the functions of identifying, locating, and directing action towards objects (Solso, 1996). It is also assumed that the visual system requires attentional mechanisms to limit the amount of sensory information it processes (Fernadez-Duque & Johnson, 1999). Thus, awareness of objects in the environment is supposed to result from a series of processing stages that select sensory information and then construct representations of objects by extracting regularities from the visual stimulus and matching them to patterns in memory (Palmer, 1999).

It was first assumed that visual attention selects certain regions of the visual field, much the way a spotlight illuminates part of a stage and leaves the rest in the dark. Accordingly, this model is known as the spotlight model of visual attention (Fernandez-Duque & Johnson, 1999). On this model, attention is first directed to a region of the visual field, and only the information within that region is processed for object identification. This assumption was questioned when researchers observed that people respond to visual features that belong to a single object more quickly and accurately than when the features belong to two objects (Duncan, 1984; Treisman, Kahneman & Burkell, 1983). Subsequent research confirmed that it is usually easier to process information within a single object than across objects (Lavie & Driver, 1996). These findings have lead to the object-based model of visual attention (Duncan, 1984; Lavie & Driver, 1996). It is now generally recognized that the spotlight and the object-based models capture complementary aspects of visual attention (Driver & Baylis, 1998).

It is undeniable that information can be processed more readily within one object than across many (Lavie & Driver, 1996; Driver & Baylis, 1998). However, the object-based explanation for this difference in processing efficiency is problematic. Cognitive psychologists generally distinguish between spatio-temporally bounded physical objects and the mental representations of these objects. Physical objects correspond to what philosophers call concrete particulars (Loux, 1998), and will subsequently be referred to as c-objects. Similarly, the mental representation of visual objects will be henceforth referred to as p-objects (for "phenomenological" objects). The generally accepted story about object perception is that the visual system constructs p-objects, which represent c-objects via various perceptual and cognitive processes. Researchers who accept the object-based model contend that attention selects "objects" for further processing. Which objects are these – p-objects or c-objects? P-objects are supposed to be the end product of visual processing (Solso, 1996), so attention must presumably be engaged prior to the construction of p-objects. However, the alternate claim that attention directly processes c-objects themselves instead of sensory input is nonsense. Most researchers assume the visual system first constructs low-level representations of c-objects, based on the physical properties of the stimulus. These representations are then elaborated into p-objects by higher-order visual and conceptual processes (Hoffman, 1998; Palmer, 1999). These low-level representations will be referred to as a-objects (for "attentional objects"). The object-based model can be restated thus: Visual attention selects a-objects, which are passed on to higher visual processes for elaboration into p-objects, which are representations of c-objects.

Philosophers are actively studying the nature of c-objects (see Loux, 1998) and perceptual psychologists are

researching p-objects (e.g. see Biederman, 1995, and Kosslyn, 1995). But the notion of a-object implicit in object-based attention is still poorly defined. Most researchers take a-objects to be perceptual groupings based on the Gestalt principles of perceptual organization (Driver & Baylis, 1998), according to which observers perceive the details of a scene only as parts of global patterns. Perceptual organization was thought to conform to the general principle of figural goodness, or *Prägnanz* (Koffka, 1935). Figural goodness was exemplified in a number of specific principles (e.g., figure-ground, grouping by similarity, good continuation, closure, and common fate; Palmer, 1999). However, this view of a-objects is inadequate.

Cognitive scientists tend to assume that cognitive processing occurs in discrete stages, as first proposed by Sternberg (1969), until evidence forces them to think otherwise. Accordingly, researchers studying object-based attention have typically assumed that perceptual grouping, in the form of Gestalt grouping, occurs at a processing stage that is independent of, but feeds into, attentional processes, and that the product of attentional selection are independent of, but feed into, object recognition processes (Feldman, 1999). This view is problematic on two counts: first, the evidence that perceptual organization occurs prior to, and independently of, visual attention is not definitive. Second, the Gestalt account of perceptual organization itself has many shortcomings. Let us examine these two issues in turn.

Attention and Perceptual Organization

Mack, Tang, Tuma, Kahn and Rock (1992) and Rock, Linnett, Grant and Mack (1992) have presented results that suggest that perceptual organization does not occur without attention. They had participants perform a task that engaged their attention (typically, judging the relative length of the branches of a cross) while varying the background on which the main stimulus was displayed. Most trials had either a blank or a random background, but each participant also saw three (non-consecutive) trials where the background contained a 'critical stimulus', either a single shape or formed some Gestalt grouping. On the first critical trial, most participants reported not seeing the Gestalt grouping or not perceiving the shape or size of the lone object. On the second critical stimulus trial, participants were generally more successful in detecting the Gestalt group or the object. On third critical stimulus trial, participants were asked to report on the background stimulus only, generally with nearly perfect results.

Mack et al. (1992) and Rock et al. (1992) assumed that on the first critical stimulus trial, participants were not expecting to see anything of import in the background, and thus focused all of their attention on their primary task, whereas on the second and third critical stimulus trials, they implicitly allocated some or all of their attention to the background pattern. Accordingly, they interpreted their results to mean that perceptual organization cannot occur without attention. Ben-Av, Sagi and Braun (1992) reported a similar study where participants had difficulty identifying background Gestalt groupings as their primary visual task became more demanding.

The results just discussed are not conclusive, however, as they cannot rule out that participants were merely unable to remember or encode the 'unattended' stimuli rather than failing to perceive them at all. Evidence along these lines is provided by Moore and Egeth (1997)[1]. In a series of experiments, they had subjects judge the relative length of two parallel horizontal lines while varying the background. On half the trials the background consisted of random black and white dots, while on the other half, the background together with the two lines (now identical in length) formed the well-known Ponzo and Müller-Lyer illusions through the Gestalt principle of grouping by similarity. Participants reliably responded in a manner consistent with the illusions (i.e. reporting that the appropriate line was longer), even though the vast majority of them were not aware of seeing the background pattern. What is crucial here is that processing the background pattern is necessary for the illusion to influence participants' responses. The conclusion is that participants perceived the right background grouping even though they had no awareness of having done so.

Do the results of Moore and Egeth (1997) establish that perceptual organization does not require attention? Not necessarily. Moore and Egeth (1997) had each participant view the displays with the illusions 16 times while they performed the line-comparison task, whereas Mack et al. and Rock et al. tested their participants' awareness of the background patterns after only the first time each participant saw the pattern. It is possible that Moore and Egeth's participants learned implicitly and unconsciously that the background was informative and divided their attention between the primary and the secondary stimuli. Furthermore, they report that once the participants were aware of the Ponzo illusion, their performance of the line-judgment task dropped to chance levels, suggesting that the illusion was no longer effective (this pattern did not obtain with the Müller-Lyer illusion). Thus, for the Ponzo illusion at least, attention *does* play a role in perceptual organization.

Beyond Gestalt Grouping Principles

Taken together, the results from Mack et al. (1992), Moore and Egeth (1997) and Rock et al. (1992) suggest that attention can influence perceptual organization. Gestalt theory offers no way of accounting for this, as on this view perceptual grouping is largely determined by stimulus properties. Palmer (1999) and Pomeranz and Kubovy (1986) have pointed out further problems for the Gestalt view. First, the Gestalt principles don't distinguish between objects and groups of objects. Also, Gestalt principles ignore the role of top-down, general-purpose knowledge in perceptual organization. For instance, the Gestalt principles cannot explain why people who don't know that Figure 1 is

[1] I would like to thank the anonymous reviewer who brought this to my attention.

a picture of a Dalmatian usually fail to see any meaningful pattern in the image, whereas people who are aware that the image represents a dog not only find the dog easily but also organize the stimulus so that otherwise indistinguishable black dots become a dog, a sidewalk, and the shade beneath a tree.

Figure 1: Spot the Dalmatian!

In order to address these problems, an account of perceptual organization must do two things: it must show how perception and attention interact to form a-objects, and it must show how general-purpose conceptual knowledge can participate in the formation of a-objects without requiring full object recognition. Both of these objectives could be facilitated by construing perceptual organization as Inference to the Best Explanation (IBE), whereby where the visual system infers three-dimensional structures which best explain the retinal image (Hoffman, 1998; Leyton, 1992; Feldman, 1999) [2]. IBE is an appealing account of human explanatory practice in general, but it suffers from the defect that explanatory 'goodness' has not yet been properly defined (Lipton, 1991). Nevertheless, vision researchers provide some candidates for goodness criteria in perceptual inference. Albert and Hoffman (1995), Feldman (1999), and Pomerantz and Kubovy (1986) have suggested that visual inference is overarchingly guided by the principle of genericity. That is, the visual system assumes, in the absence of other data, that the retinal image is a generic view of three-dimensional objects, rather than a very specific and "accidental" view of some other set of three-dimensional objects. A generic view is a two-dimensional projection of a three-dimensional structure that does not entail special or accidental circumstances in the projection. For instance, a straight line in the retinal image is a generic view of a straight edge in the environment, but would be a non-generic view of a curved edge that just happens to be seen head-on. The genericity principle can account for a large number of phenomena of perceptual organization. Furthermore, Pomerantz and Kubovy (1986) show that the Gestalt principles can and should be reinterpreted as instances of the genericity principle.

The notion of genericity can be extended to explain the role of conceptual knowledge in perceptual organization. Assuming an observer expects to see a Dalmatian in Figure 1, the splotches obviously form a generic view of the dog. Whereas, if the splotches corresponded to a horse, it would have to be either a typical horse seen under very specific shading conditions, or a strangely Dalmatian-like horse seen under normal viewing conditions. "Explaining" the image as that of a horse would require invoking a number of special circumstances, which interpreting the image as that of a Dalmatian does not. The visual system might thus limit the conceptual information involved in constructing a-objects to knowledge of generic views and expectations about which objects are present in a scene.

A-objects can now be re-defined: An a-object is a representation of the three-dimensional structure that best explains the two-dimensional retinal stimulus according to the genericity principle, which takes into account both physical stimulus properties and general-purpose conceptual knowledge.

Although the Dalmatian example demonstrates the role of semantic information in perceptual organization, it remains to be shown that a-objects as defined above actually play a role in object-based attention. The following section presents some recent experimental evidence bearing on this issue from our laboratory.

Recent Evidence for the Involvement of Attention in Perceptual Organization

A first line of evidence for the role of attention in perceptual organization comes from recent studies on object-based attention using moving stimuli (Jarmasz, 2001; Jarmasz, Herdman & Johannsdottir, in preparation). In these experiments, participants were shown a display consisting of two groups of identical dots. One set of dots was static, while the dots in the second group moved in unison in an elliptical trajectory that overlapped the location of the static dots. During each trial, two of the dots in the display changed color from light gray to one of two colors (either red and green, or blue and yellow). The target dots were located both in the static group, both in the moving group, or one in each group. Participants were required to determine whether the target dots were the same color. On some trials participants had to focus their attention on only one group of dots, while in other trials they had to spread their attention to the display as a whole, and avoid focusing on a specific group. When participants attended to the whole display, they responded significantly faster to target dots displayed within a single group than to targets appearing across both groups. The results were consistent with those found in the object-based attention literature using static displays (e.g. see Lavie & Driver, 1996; Treisman et al., 1983). However, when participants focused on only one of the groups, their responses were faster when both targets appeared in the attended group, and slowest when targets appeared either in the unattended group only or across groups. A comparison of response latencies across attentional focus conditions suggests that focused attention inhibits information processing of unattended stimuli rather than enhancing processing of attended stimuli relative to

[2]This does not necessarily imply deliberate, conscious inference.

situations where attention is deployed across the whole display. In all attentional focus conditions, both groups of dots overlapped in the display, so foveal limitation on visual acuity likely do not account for these results. Rather, they suggest that deliberate attentional allocation strategies influence the degree to which multiple perceptual groups are perceived as being separate from the others, or conversely as being part of a larger whole. This is consistent with the proposal above that a-objects depend on attentional factors as well as bottom-up stimuli properties.

A second line of evidence implicating attention in perceptual organization comes from a study using static visual objects (Jarmasz, 2002). In this study, participants were shown a display based on one used by Lavie and Driver (1996) consisting of two intersecting dashed lines. On each trial, two of dashes were replaced by two target elements, either a shorter dash or a gap (absence of a dash). The target elements could appear either near to each other on separate lines ('near' condition), far from each other on separate lines ('far' condition) or far from each other but on the same line ('object' condition). On half the trials, both dashed lines were the same color (either pink or yellow), whereas in the other trials each line was a different color. In all cases, participants responded most quickly to targets in the object condition, even though they were approximately seven times further apart in the object condition than in the near condition, thus replicating the effect found by Lavie and Driver (1996). When the dashed lines were of different colors, participants were equally slow in responding to targets in the near and far conditions relative to the object condition. However, when both lines shared a common color, participants responded more quickly to targets in the far condition than in the near condition (responses to targets in the object condition were still fastest overall). This suggests that in the same color condition, participants parsed the display either as two separate lines or as one large figure, depending on the location of the targets. These results further suggest that observers can implicitly acquire top-down attention allocation strategies that affect the perceptual organization of a stimulus from bottom-up cues (color & target location).

The two studies described above show that attention can be deployed, either implicitly or explicitly, so as to influence perceptual organization. The results of these two studies are inconsistent with the spotlight model of visual attention, in which spatial separation, but not shape, affects how quickly two stimuli are compared. However, the "standard" object-based attention model cannot account for these results either, as on this model perceptual grouping is assumed to be preattentive, and thus impervious to changes in attentional allocation strategies. An adequate account of visual attention and object perception will have to explain both how perceptual organization can affect attention (i.e. object-based attention) and how attention can affect perceptual organization. The next section presents such a model.

Towards a New Theory of Object-Based Attention

Current accounts of object-based attention do not reflect the reciprocal influences between attention and perceptual organization. Consequently, I propose a new account, the Inferential Attentional Allocation Model (IAAM; Jarmasz, 2001). Briefly, on this model attention and perceptual organization interact to incrementally build up representations of c-objects. Potential a-objects are constructed from regions of uniform color, luminance, and texture, and edge-bound surfaces in the visual stimulus (Palmer, 1999). These potential a-objects represent rival hypotheses as to the 3-D structures in the environment, and lack detail. A-objects are refined through cycles of grouping and selection (Grossberg, Mingolla & Ross, 1994). At each cycle, attention selects a-objects that best satisfy both genericity and cognitive set (i.e., an observer's perceptual expectations and general-purpose knowledge). This progressively liberates resources for generating more detailed a-objects that better "explain" the retinal image. If this process is interrupted before it culminates in stable a-objects, an observer may fail to perceive a c-object (Di Lollo, Enns, and Rensink, 2000, have found evidence for such a phenomenon). Figure 2 depicts these iterative processes.

The IAAM is a heuristic (i.e., exploratory) model. Even so, it allows for the following predictions: (1) stimulus-dependent, bottom-up information constrains possible a-objects in the scene, and as well as how much attentional resources the stabilization of particular a-objects will require; and (2) conceptual, top-down knowledge acts to determine which a-objects will eventually become stable and become p-objects. Thus, bottom-up properties of a stimulus should determine how "easy" it is to perceive certain objects, i.e., how much attention perceiving those objects will require and how efficient the processing will be. Top-down factors will sometimes push the visual system to organize a stimulus into more attentionally demanding configurations which will result in less efficient processing (reflected either in slower processing or more interference from other stimuli). The experiments reported above are largely consistent with these hypotheses. Common motion (a bottom-up factor) makes the segregation of a stimulus into distinct objects possible, but the intention to pick out one of these objects (top-down) enhances the processing of that object at the expense of processing information from other objects (Jarmasz, 2001; Jarmasz, Herdman & Johannsdottir, in preparation). Similarly, collinearity and common color (bottom-up) facilitate the segregation of two dashed lines into two objects, but implicit task demands (top-down) seem to determine whether the two lines are actually parsed as one large figure or two lines (Jarmasz, 2002)

Further work is needed to elaborate and test the IAAM. Namely, the notions of 'ease' of perceptual grouping and of efficiency of visual processing need to be operationally defined. Nevertheless, one can imagine how the IAAM

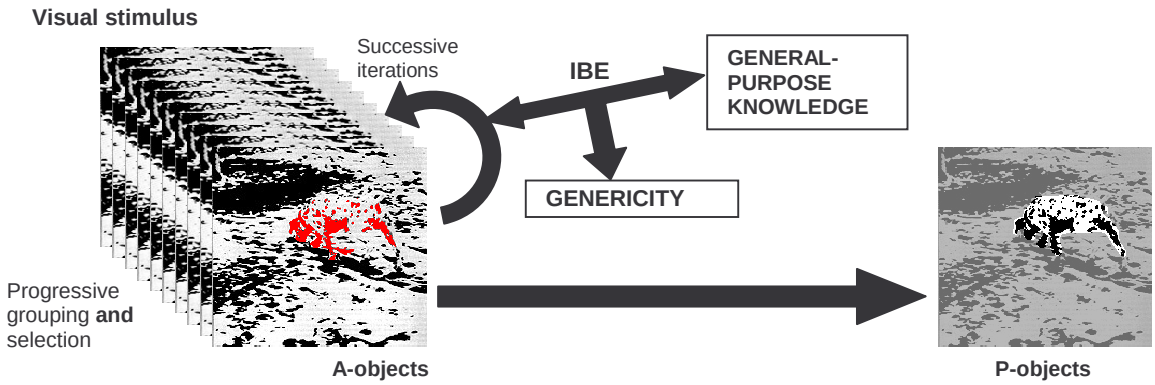


Figure 1: Grouping and selection processes in the IAAM

might apply to "real world" stimuli. For instance, if someone intends to move a box with a lid from a table to the top shelf of a bookcase, they will form a single a-object corresponding to the box and its lid. If, however, that person wants to open the box, they will form two a-objects, one for the lid and one for the box. On this view, a-objects are interest-relative; that is, a-objects depend on an observer's goals and general-purpose conceptual knowledge, in addition to bottom-up stimuli. This is in contrast to the standard view, where a-objects are defined purely by stimulus properties.

Conclusion: Some Implications of the IAAM

The object-based model of attention is currently based on the assumption that visual attention selects perceptual groups that are formed preattentively according to the Gestalt grouping principles. This conceptualization of visual attention does not reflect the reciprocal influences between perceptual organization and attention, and further ignores the role of top-down information in perceptual organization. Moreover, this formulation is limited in its ability to guide human factors research, where broad principles are often lacking and problems often require ad-hoc solutions. A growing body of experimental evidence supports the notion that while attention is influenced by perceptual organization, it in turn influences whether and how perceptual organization occurs as well. Consequently, a new model of object-based attention, the Inferential Attentional Allocation Model, is proposed which attempts to capture the interaction between attention and perceptual organization. On this model, visual attention is object based not because attention selects objects, but rather because attention itself is indispensable to perceptual organization.

The IAAM is a heuristic model. In addition to providing a framework for developing a comprehensive account of visual attention, the IAAM also has potentially significant implications both for human factors research and for the metaphysics of concrete particulars. Regarding research on human factors, the IAAM shows that strategies for deploying attention interact both with visual stimuli and with task demands. Thus, the design of graphical user interfaces such as desktop computer applications and head-up displays in aircraft and automobiles should take into account how a user's cognitive and attentional sets might interact with the display.

The IAAM shows that what counts as an object for the visual system depends intimately on an observer's goals and expectations. This reminds us that in a larger sense, everyday objects are embedded in a complex web of human activities and conventions. Standard metaphysical theories generally attempt to define concrete particulars without any reference to the agents that use and perceive them (e.g., see Loux, 1998). However, assuming that at least the broad lines of the IAAM are a valid account of object-based attention, what counts as an object for us as agents also depends on our expectations, intentions and general background knowledge. The criteria of "objecthood" might ultimately depend as much on epistemic issues as on metaphysical ones, as suggested by Smith (1996). Attention is object-based not only because attention and perceptual organization are mutually dependent, but also because objects would not be objects if we did not perceive them as such, but merely relatively coherent portions of the spatiotemporal flux we call the universe.

Acknowledgments

I am grateful to C. Herdman, A. Brook, A. Vellino, J. LeFevre, R. West, S. Scott and A. Pyke for their helpful feedback. L. Jerzykiewicz and two anonymous reviewers provided valuable comments on a previous version of this paper that was presented at the Graduate Student Conference on Philosophy of Mind, Philosophy of Language, and Cognitive Science at Carleton University, Ottawa, Canada on September 29, 2001. L. Stelmach of the Communications Research Centre in Ottawa, Canada was most helpful with technical support in the early stages of this research. Thanks go to K. Johannsdottir and J. Shaw for their help with conducting various experiments. This research is funded by the Centre for Research in Earth and Space Technology, the Natural Sciences and Engineering

Research Council of Canada, CMC Electronics, the HFE Group, the Aviation and Cognitive Engineering Laboratory at Carleton University, and Neptec.

References

Albert, M. K., & Hoffman, D. D. (1995). "Genericitiy in spatial vision." In R. D. Luce, M. D'Zmura, D. D. Hoffman, G. J. Iverson, & A. K. Romney (Eds.), *Geometric representations of perceptual phenomena.* Mahwah, NJ: Erlbaum.

Ben-Av, M. B., Sagi, D., & Braun, J. (1992). "Visual attention and perceptual grouping." *Perception & Psychophysics 52:* 277-294.

Biederman, I. 1995. "Visual object recognition." In S. M. Kosslyn and D. N. Osherson (Eds.), *An Invitation to Cognitive Science (Second Edition), Volume 2.* Cambridge, MA: MIT Press.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). "Competition for consciousness among visual events: The psychophysics of reentrant visual processes." *Journal of Experimental Psychology: General 12:* 481-507.

Driver, J., & Baylis, G. C. (1989). "Movement and visual attention: The spotlight metaphor brakes down." *Journal of Experimental Psychology: Human Perception and Performance 15:* 448-456.

Driver, J., & Baylis, G.C. (1998). "Attention and visual object segmentation." In R. Parasuraman (Ed.), *The Attentive Brain.* Cambridge, MA: MIT Press.

Duncan, J., (1984). "Selective attention and the organization of visual information." *Journal of Experimental Psychology: General 113:* 501-517.

Feldman, J. (1999). "The role of objects in perceptual grouping." *Acta Psychologica 102:* 137-163.

Fernandez-Duque, D., & Johnson, M. L. (1999). "Attention metaphors: How metaphors guide the cognitive psychology of attention." *Cognitive Science 23:* 83-116.

Grossberg, S., Mingolla, E., & Ross, W. D. (1994). "A neural theory of attentive visual search: interactions of boundary, surface, spatial and object representations." *Psychological Review 101:* 470-489.

Hoffman, D. D. (1998). *Visual Intelligence.* New York: W. W. Norton & Company.

Jarmasz, J. (2001). Towards the Integration of Perceptual Organization and Visual Attention: The Inferential Attentional Allocation Model. Carleton University Cognitive Science Technical Report 2001-08. URL http://www.carleton.ca/iis/TechReports.

Jarmasz, J (2002). Brief report on the effects of color and target location in the Lavie and Driver object-based attention paradigm. Cognitive Science Technical Report 2002-01. URL http://www.carleton.ca/iis/TechReports.

Jarmasz, J. Herdman, C. M., Johannsdottir, K. R. (in preparation). Object-based attention and cognitive tunneling with heads-up displays in aircraft.

Koffka, K. 1935. *Principles of Gestalt Psychology*. New York: Harcourt, Brace & World.

Kosslyn, S. M. (1995). "Mental Imagery." In S. M. Kosslyn and D. N. Osherson (Eds.), *An Invitation to Cognitive Science (Second Edition), Volume 2*. Cambridge, MA: MIT Press.

Lavie, N. & Driver, J. (1996). "On the spatial extent of attention in object-based visual selection." *Perception & Psychophysics 58:* 1238-1251.

Leyton, M. (1992). *Symmetry, Causality, Mind.* Cambridge, MA: MIT Press.

Lipton, P. (1991). *Inference to the Best Explanation.* London: Routledge.

Loux, M. J. (1998). *Metaphysics: A Contemporary Introduction.* London: Routledge.

Mack, A., Tang, B., Tuma, R., Kahn, S., & Rock, I. (1992). "Perceptual organization and attention." *Cognitive Psychology 24:* 475-501.

McLeod, P., Driver, J., Dienes, Z., & Crisp, J. (1991). "Filtering by movement in visual search." *Journal of Experimental Psychology: Human Perception and Performance 17:* 55-64.

Moore, C. M., & Egeth, H. (1997). "Perception Without Attention: Evidence of Grouping Under Conditions of Inattention." *Journal of Experimental Psychology: Human Perception and Performance 23:* 339-352.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press.

Pomerantz, J. R., & Kubovy, M. (1986). "Theoretical approaches to perceptual organization." In , K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of Perception and Human Performance, Volume II*. New York, NY: John Wiley & Sons, Inc.

Rock, I., Linnett, C. M., Grant, P., & Mack, A. (1992). "Perception without attention: Results of a new method." *Cognitive Psychology 24:* 502-534.

Smith, B. C. (1996). *On the Origin of Objects.* Cambridge, MA: MIT Press.

Sternberg, S. (1969). Memory-scanning : Mental processes revealed by reaction-time experiments. In *American Scientist, 57,* 421-457.

Solso, R. L. (1996). *Cognition and the Visual Arts.* Cambridge, MA: MIT Press.

Treisman, A., Kahneman, D., & Burkell, J. (1983). "Perceptual objects and the cost of filtering." *Perception & Psychophysics 33:* 527-532.

Children's Acceptance and Use of Unexpected Category Labels to Draw Non-Obvious Inferences

Vikram K. Jaswal (jaswal@psych.stanford.edu)
Department of Psychology, Jordan Hall, Bldg. 420, Stanford University
Stanford, CA 94305-2130 USA

Ellen M. Markman (markman@psych.stanford.edu)
Department of Psychology, Jordan Hall, Bldg. 420, Stanford University
Stanford, CA 94305-2130 USA

Abstract

Language provides an efficient, uniquely human way of transmitting non-obvious category information between individuals and across generations. To explore whether it can serve this purpose even for very young children, we conducted two experiments: one with 24-month-olds, and the other with preschoolers. Children made non-obvious inferences about perceptually misleading animals. Those who heard the animals called by counter-intuitive category labels made inferences different from those who did not hear the labels, demonstrating an important influence of language on thought, even at 24 months of age. However, preschoolers appear to have been less influenced than toddlers, suggesting that there are limits on children's willingness to accept anomalous category labels.

Introduction

Categorization is fundamental to human cognition, enabling communication and serving as the basis for the representation of objects and for predicting and explaining their behavior (Anglin, 1977; Markman, 1989). Children as young as 3.5 months (Eimas & Quinn, 1994), adults (Rosch et al., 1976), and non-human animals (e.g., Freedman et al., 2001) alike readily use perceptual similarity to determine the category to which something belongs. But when reasoning about an object, or explaining or predicting its behavior, perceptual appearance is not always criterial of category membership. For example, even though eels look like snakes, in order to more accurately characterize their ancestry, behavior and physiological processes, experts categorize them as fish.

Given sufficient experience, children as young as 30 months can form non-obvious categories by noting causal (Gopnik & Sobel, 2000) or functional (Kemler Nelson et al., 2000) regularities between objects. Under some circumstances, even non-human animals can learn to categorize objects in perceptually non-obvious ways (Herrnstein & DeVilliers, 1980). However, recognizing non-obvious similarities can be a slow and laborious process, often requiring experience that is difficult to obtain. Moreover, it requires every individual in every generation to have the experience for him or herself (Tomasello, 1999). Another, arguably more reliable and efficient way to obtain non-obvious category information is through language (Gelman et al., 2000): When a trusted source uses an unexpected category label for an object, it reflects a particular perspective that others have found useful when thinking and reasoning about that object in the past, and it can cause us to revise a classification immediately. For language to have this effect, however, listeners may have to give up a compelling, perceptually based classification in favor of a classification they do not immediately understand.

As adults, we can accept linguistically provided non-obvious classifications (e.g., a whale is a mammal, not a fish) because we implicitly assume that something deeper than surface similarity binds category members together (Medin & Ortony, 1989). Such an essentialist assumption would facilitate the rapid uptake of non-obvious category terms. However, it is not clear whether very young children also expect categories and category terms to encode more than surface similarities. For example, it has been argued that children begin with or quickly develop an expectation that category labels encode similarly shaped objects (e.g., Imai, Gentner, & Uchida, 1994; Smith, 1999). Furthermore, whereas children readily learn words for basic-level categories (e.g., "table"), they have difficulty learning words for superordinate categories (e.g., "furniture"), which have fewer perceptual features in common than basic-level ones (e.g., Horton & Markman, 1980; Mervis & Rosch, 1981; Rosch et al., 1976). Finally, young children sometimes object to parental attempts to correct perceptually based categorization errors (Mervis, 1987).

In two studies, we investigated children's willingness to accept perceptually counter-intuitive classifications on the basis of linguistically provided information alone. In our first study, we asked this question of 24-month-olds. Gelman and her colleagues have found that children as young as 32 months can make inferences on the basis of linguistic information rather than perceptual appearance (Gelman & Coley, 1990; Gelman & Markman, 1986), but their procedure involves memory demands, verbal responses, and linguistic comprehension abilities beyond those of younger children. Our procedure uses an imitation paradigm that minimizes these task demands (see also Baldwin, Markman, & Melartin, 1993; Mandler & McDonough, 1996; Welder & Graham, 2001).

Although preschoolers have been shown to defer to adult labels in the face of perceptually inconsistent information (Gelman & Coley, 1990; Gelman & Markman, 1986), there must be limits on this process. Surprisingly, no one has considered situations under which children reject linguistically provided category information. Using the same paradigm and materials as in the first study, our second study takes up this question with preschoolers.

Experiment 1

Experiment 1 was designed to test whether 24-month-olds would use an experimenter-provided, sometimes counter-intuitive, category label when making non-obvious inferences. In particular, we were interested in whether they would use the label to make an inference that was different from the one they would make without a label.

Method

Participants Participants were 32 toddlers, ranging in age from 23 months, 6 days to 25 months, 28 days (M = 24 months, 10 days). Five additional toddlers were tested, but their data are not included due to extreme fussiness resulting in an inability to complete at least half of the session (4), or due to experimenter error (1).

Stimuli Eight animals from familiar categories were grouped into four sets based on similar sizes and body shapes: Cat-dog, horse-cow, bear-pig, and squirrel-rabbit. Category familiarity was confirmed by consulting the MacArthur Communicative Development Inventory for toddlers (Dale & Fenson, 1996), which indicated that at least 44% of 24-month-old children could produce the word associated with each animal (range: 43.9% for "squirrel" to 91.6% for "dog").

Realistic, color drawings of a typical exemplar of each animal were obtained from commercially available picturebooks, and were digitized for computer manipulation (hereafter, "standard animals"). In addition, three test animals for each set were generated on Adobe PhotoShop from the standard animals of that set: Two were typical exemplars (hereafter "typical test animals"), and were created by manipulating the coloration of the standards; the third test animal looked more like one standard animal than the other (hereafter, "misleading test animal"). Misleading test animals were created by using one of the standards (or an additional typical image) as a base, and adding features from the other standard image of that pair. Examples of standard and misleading animals from the cat-dog set are shown in Figure 1.

Sixteen graduate students rated each animal, including the standards, from each set on a 7-point continuum. At one end of the continuum was the category label for one of the animals in a given set (e.g., "cat"), and at the other end was the category label for the other animal in that set (e.g., "dog"). Subjects indicated on the continuum the position of each animal image, with 4 being ambiguous—exactly halfway between the two. The average ratings of the standards and typical test animals were always at the ends of the continuum (between 1 and 2 or between 6 and 7), and those of the misleading test animals were always slightly closer to the center, but leaning in one direction (between 2 and 3, or between 5 and 6) and significantly different from 4 and from each other. A full display of the stimuli and details of their adult ratings can be obtained at http://www-psych.stanford.edu/~jaswal/Stimuli/index.html.

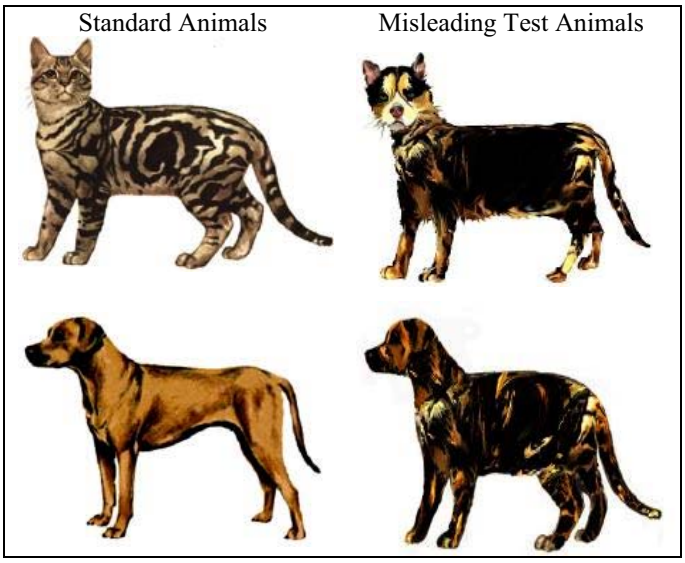


Figure 1: Sample stimuli from the cat-dog set. Each stimulus set was made up of two standard animals, two typical animals (not shown, but very similar to the standards), and one of two misleading animals.

All images were sized to approximately 3 to 4 inches wide, with their heights constrained by their widths. Each image (and its left-right reverse) was printed on a 600 dpi color printer, cut out, laminated, and mounted into a small stand that allowed it to remain upright.

Each set of animals was associated with a particular activity, and members of the same pair were associated with contrasting props or dioramas (hereafter, "props"). Activities were chosen that could be easily and clearly demonstrated with a 3-dimensional prop, and that might (though not necessarily) be familiar to children: The cat played with a ball of yarn, the dog played with a stick; the horse slept on the hay, the cow slept on the grass; the bear lived in the forest, the pig lived in the mud; the squirrel ate a nut, the rabbit ate a carrot. Props were purchased or specially constructed.

Design The 32 children were randomly assigned to a label or a no-label condition, resulting in 16 children per condition, balanced for sex. Each child in the label condition was yoked to a child in the no-label condition so that both saw exactly the same animals in exactly the same configuration and order. The only difference was that one heard a label during the presentation of each test animal, and the other did not.

Procedure Toddlers were seated on their parent's lap at a table in a testing room, or at a small table on their own, with the experimenter sitting across the table from them. Each session began with a warm-up trial to familiarize them with the procedure: The experimenter demonstrated that a fish lived in an aquatic scene ("water"), and a bird lived in a tree. They were then shown additional typical fish and birds in alternating order and asked to show where each lived, until they succeeded in putting a fish in the water and a bird in the tree consecutively, or until their attention waned. Correct selections were praised, and incorrect selections were corrected.

During the testing phase, children watched as the experimenter demonstrated and explained aloud that one standard animal engaged in an activity with one prop (for example, that a cat played with a ball of yarn), and that the second standard animal of that set engaged in an activity with another prop (for example, that a dog played with a stick). Following the introduction and labeling of each standard animal, children were encouraged to imitate the activity with that same animal.

Children were then shown the two typical test animals and one of the misleading test animals from that set, one at a time and in a random order. They were asked to show the activity in which each engaged. The children in the label condition heard the experimenter use a category label to introduce each of the test animals (for example, "Look at this! Look at this dog! Can you show me what this dog plays with?"). The typical test animals were always called by labels that matched their appearance, and the misleading test animals were always called by labels that were the opposite of their appearance (i.e., if adult raters had indicated that a misleading test animal looked more like a dog than a cat, it was referred to as "this cat"). To establish baseline levels, children in the no-label condition heard the experimenter use the phrase "this one" to introduce each test animal (for example, "Look at this! Look at this one! Can you show me what this one plays with?"). Regardless of their selection, children were given neutral feedback in a positive tone ("OK!"), and the experimenter then proceeded to the next test animal or animal set.

Most children were presented with all four sets of animals; however, due to fussiness, five children were presented with three sets rather than four. The order in which the animal sets were presented was counterbalanced across pairs of children according to a Latin Square design. The prop or diorama matching a misleading test animal's perceptual appearance appeared twice on the left and twice on the right for each child, and this was counterbalanced across pairs of children.

Coding was conducted off-line, via videotape, and involved noting which of the two possible props was selected for each test animal. Two coders, blind to condition, each coded one-half of the sessions. To assess reliability, each coder also independently coded 1/4 of the other coder's sessions (selected randomly); agreement on which prop was selected was 99% (Cohen's kappa=.98).

Results and Discussion

As shown in Figure 2, children in both the label and the no-label conditions inferred that the typical test animals engaged with the props associated with their perceptual appearance significantly more frequently than would be expected by chance (50%) [label condition: $t(15)=4.01$, $p<0.01$; no-label condition: $t(15)=2.93$, $p<0.05$], and at levels not different from each other [$t(30)<1$, *n.s.*]. This is consistent with other work showing that, in the absence of labels, even 9-month-olds can make non-obvious inferences based on an object's appearance (Baldwin et al., 1993).

When a test animal was perceptually misleading, however, labeling had a dramatic effect, as Figure 2 shows. Children in the label condition made inferences based on the perceptual appearance of the misleading test animals less frequently than by chance [$t(15)=-2.93$, $p<0.05$], whereas those in the no-label condition did so at chance levels [$t(15)<1$, *n.s.*]. A 2-way mixed ANOVA on the mean percentage of perceptual inferences (label/no-label x typical/misleading) yielded a main effect of stimulus type [$F(1,30)=13.41$, $p<0.01$], and a significant interaction [$F(1,30)=6.62$, $p<0.05$]. Children in the label condition made significantly fewer inferences in line with the perceptual appearance of the misleading test animals than those in the no-label condition [$t(30)=-2.32$, $p<0.05$].

In short, children in the label condition inferred that the misleading animal engaged with the prop associated with its label, while those in the no-label condition were significantly more likely to infer that it engaged with the prop associated with its perceptual appearance. For example, when shown the misleading animal that adult subjects rated as dog-like, children who heard it referred to as "this cat" inferred that it played with the ball of yarn, while those who heard it referred to as "this one" were more likely to infer that it played with a stick.

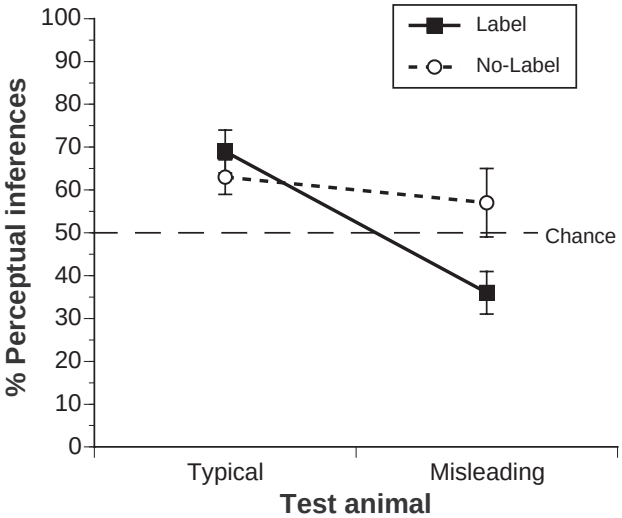


Figure 2: Mean percentage of inferences by toddlers that were consistent with perceptual appearance. Error bars show standard error.

Using a non-verbal response measure, we found that even 24-month-old children weigh the category label that a speaker applies to an animal more heavily than the animal's perceptual appearance when the two are in conflict. Although the vast majority of category labels in a typical two-year-old's productive vocabulary represent categories whose members cohere perceptually (Samuelson & Smith, 1999), these children nonetheless expect category members to share deeper similarities, and they can use language to think about objects in non-obvious—even counter-intuitive—ways. The power of a simple category label to convey non-obvious category information may well develop with age and experience with language. Certainly by 24 months, however, this process is in place, and can serve as an important mechanism for the cultural transmission of knowledge and information.

This raises interesting questions about the limits of this process, including circumstances under which children might weigh an object's perceptual appearance more heavily than its category label. In our second experiment, we explored one possibility, namely that older children might be less willing to accept a perceptually counter-intuitive category label than younger children.

Experiment 2

Experiment 2 was designed to explore whether preschoolers would be as willing to accept perceptually counter-intuitive category labels as the toddlers in Experiment 1. Presumably, older children have been exposed both to more exemplars and to a wider range of exemplars from the familiar animal categories used in our experiment. Because of this experience, they may be less willing to accept anomalous category information. For example, when they see an animal clearly possessing the perceptual features of a dog (a "misleading test animal"), they may construe it as a dog that shares properties with other dogs—regardless of what the experimenter calls it.

Method

Participants Twelve 3-year-old (M = 3 years, 3 months; range = 2-10 to 3-6) of middle-class and upper-middle-class backgrounds participated at a university-affiliated preschool. Six were girls and six were boys.

Stimuli The stimuli were the same as those used in Experiment 1.

Design Experiment 2 used a within-subject design, with the same child hearing some test animals labeled and other test animals referred to as "this one."

Procedure Children were tested individually in a quiet room at their preschool. The procedure was generally the same as that used in Experiment 1: Children watched as the experimenter demonstrated and explained aloud that one standard animal engaged with one prop, and the second standard animal of that pair engaged with another. They were then shown the three test animals one at a time and in a random order, and asked to show the activity in which each engaged. All children were shown four sets of animals.[1]

The major difference from Experiment 1 was that children heard labels for test animals from two of the sets, and did not hear labels for test animals from the other two. As in Experiment 1, children were yoked in pairs so that both saw exactly the same stimuli in exactly the same configuration and order. When one child heard test animals from a particular set labeled, his or her yoked partner did not. The order in which the sets were presented was random, as was the order of the test-baseline sets. The prop matching the misleading animal's perceptual appearance appeared equally often on the left and right for label and no-label sets.

Coding was conducted off-line, via videotape, and involved noting which of the two possible props was selected for each test animal.

Results and Discussion

As shown in Figure 3, like the toddlers in Experiment 1, preschoolers inferred that the typical test animals engaged with the props associated with their perceptual appearance significantly more frequently than would be expected by chance, regardless of whether the animals were labeled or not [label condition: $t(11)=13.40$, $p<0.01$; no-label condition: $t(11)=6.09$, $p<0.01$], and at levels not different from each other [$t(11)=1.00$, *n.s.*].

For perceptually misleading animals that were not labeled, preschoolers had them engage with the prop that matched their appearance more frequently than would be expected by chance [$t(11)=11.00$, $p<0.01$]. By contrast, when they were labeled, children had them engage with the prop that matched their appearance at chance levels [$t(11)<1$, *n.s.*]. A 2-way repeated measures ANOVA on the

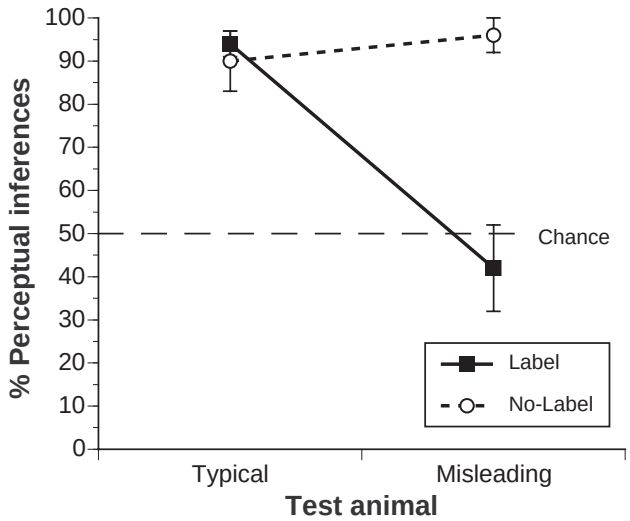


Figure 3: Mean percentage of inferences by preschoolers that were consistent with perceptual appearance. Error bars show standard error.

[1] Additionally, children were shown two sets of artifacts; those data will not be presented here.

mean percentage of perceptual inferences (label/no-label x typical/misleading) yielded a main effect of condition [$F(1,11)=33.00$, $p<0.01$], a main effect of stimulus type [$F(1,11)=7.44$, $p<0.05$], and a significant interaction [$F(1,11)=23.44$, $p<0.01$]. When they heard misleading animals labeled, children made significantly fewer inferences in line with their perceptual appearance than when they did not hear them labeled [$t(11)=-5.61$, $p<0.01$].

As in Experiment 1 with toddlers, then, labeling clearly had an effect on preschoolers' inferences about the perceptually misleading animals. However, as will be discussed below, once performance on the typical animals was equated, preschoolers appear to have been less likely to accept and use counter-intuitive category labels than toddlers.

General Discussion

These two experiments investigated children's willingness to accept perceptually counter-intuitive classifications on the basis of linguistically provided information. In Experiment 1, 24-month-olds used category labels provided by the experimenter to make non-obvious inferences about animals that were the opposite of those they would make without a label. These results converge with those of Welder and Graham (2001), who found that 16- to 21-month-olds could use a label to make a non-obvious inferences about a novel object. In that study, infants who heard two moderately dissimilar novel objects called by the same novel name inferred that they shared a non-obvious property (Study 2). When the objects were not labeled, infants did not spontaneously assume that they shared the same property (Study 1).

Our design differs from that of Welder and Graham (2001) in an important way: In particular, whereas Welder and Graham used novel objects and novel labels (except in Study 3, where they used novel objects and familiar labels), we used instead familiar objects and labels. By using familiar categories, we were able to show that despite compelling, contradictory perceptual information indicating membership in a different category, 24-month-olds nonetheless could accept and use the experimenter's label in making a non-obvious inference.

In Experiment 2, we found that preschoolers could also use labels to make inferences different from those they would make without a label. This is consistent with work by Gelman and her colleagues (Gelman & Coley, 1990; Gelman & Markman, 1986), who have found similar results with preschoolers using different procedures. However, we also found that preschoolers were less willing to accept and use an unexpected category label than the toddlers in Experiment 1. The first piece of evidence for this conclusion comes from the chance analyses, which showed that hearing a label for a misleading animal made toddlers in Experiment 1 less likely than chance to make an inference in line with the animal's perceptual appearance, whereas preschoolers in Experiment 2 were merely at chance—meaning that the older children were as likely to use the animal's perceptual appearance as the experimenter-provided label.

The second piece of evidence comes from an additional analysis of the data from the label conditions. So far, we have been considering the proportion of inferences children made in line with each test animal's perceptual appearance. Another way to consider the data from the label conditions only is in terms of the proportion of inferences children made in line with the experimenter-provided labels. These data are shown in Figure 4.

One might reasonably expect the proportion of label inferences for typical animals to be higher than the proportion of label inferences for the misleading animals, because in the case of a typical animal, there is no conflict between the label and the animal's appearance. And, in fact, as can be seen in Figure 4, for both toddlers and preschoolers, this is the case. However, whereas toddlers made only slightly fewer label inferences for misleading animals than typical ones, preschoolers were much less likely to make label inferences for misleading animals than typical ones. Indeed, a 2-way mixed ANOVA on the average percentage of label inferences (toddler/preschool x typical/misleading) yielded a main effect of stimulus type [$F(1,26)=9.67$, $p<0.01$], and a significant interaction [$F(1,26)=7.01$, $p<0.05$]. In other words, toddlers were equally likely to make an inference in line with misleading animals' labels as typical animals' labels [$t(15)<1$, *n.s.*], while preschoolers were significantly less likely to do so [$t(11)=-3.74$, $p<.01$].

Although differences in the design of these two studies make the cross-experiment comparisons somewhat tentative, this difference in children's willingness to accept anomalous category labels is extremely provocative: It suggests that 24-month-olds may be more "open-minded" about category information than 3-year-olds, perhaps because the toddlers assumed that they had not yet encountered all possible exemplars of even common

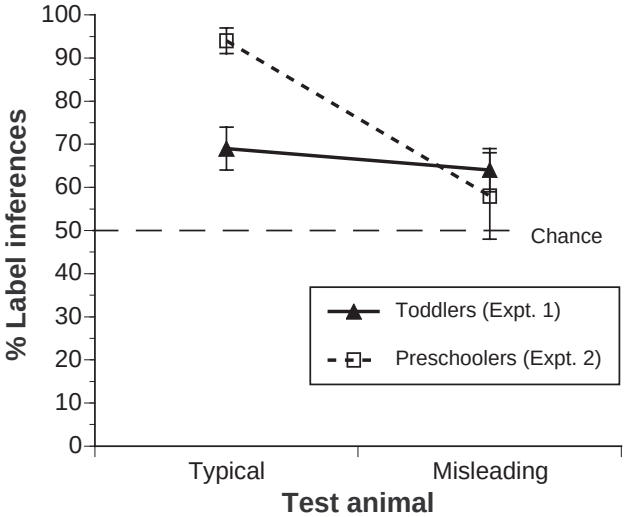


Figure 4: For toddlers and preschoolers who heard test animals labeled, mean percentage of inferences that were consistent with the labels. Error bars show standard error.

animals, whereas the preschoolers assumed they had (see Naigles, Gleitman, and Gleitman, 1992, for an analogous argument in verb learning). On-going work is considering this possibility.

Putnam (1977) has argued that a key element of adult communication is that we frequently use terms to refer to things without necessarily knowing the criteria for those terms. However, we assume that there are experts in our community who could provide the criteria and perform a test, if necessary, to fix the extension of those terms. In his words, language use requires a "division of linguistic labor." We have shown that children as young as 24 months also have something like a division of linguistic labor operating: They can accept and use what might be considered baffling category labels in order to make non-obvious inferences about animals. Using language in this way allows them to stretch the boundaries of their own spontaneously generated, often perceptually based, categories (but see Mandler & McDonough, 1996), and to take advantage of the richer and more conceptual frameworks that their cultures have evolved. A fruitful area for further investigation will be exploring the limits of this process.

Acknowledgments

We thank the children, teachers, staff, and parents who participated in this research, including those at the Arboretum Child Care Center and Bing Nursery School. Additionally, we thank A. Fernald for her assistance in subject recruitment and for providing testing space, and A. Sud for assistance in data collection. This research was supported by an NRSA predoctoral award from NIH and a Stanford Graduate Research Opportunity grant to VKJ.

References

Anglin, J. M. (1977). *Word, object, and conceptual development*. New York: Norton.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64*, 711-728.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. Behavior Research Methods, Instruments & Computers, *28*, 125-127.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development, 65*, 903-917.

Freedman, D. J., Risenhuber, M., Poggio, T., Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*, 312-316

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796-804.

Gelman, S. A., Hollander, M., Star, J., & Heyman, G. D. (2000). The role of language in the construction of kinds. *The Psychology of Learning and Motivation, 39*, 201-263.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-208.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*, 1205-1222.

Herrnstein, R. J., & de Villiers, P. A. (1980). Fish as a natural category for people and pigeons. *The Psychology of Learning and Motivation, 14*, 59-95.

Horton, M. S., & Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child Development, 51*, 708-719.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*, 45-75.

Kemler Nelson, D. G., Russel, R., Duke, N., & Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development, 71*, 1271-1288.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition, 59*, 307-335.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge: Cambridge UP.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge: Cambridge UP.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Naigles, L. G., Gleitman, H., & Gleitman, L. R. (1992). Children acquire word meaning components from syntactic evidence. In E. Dromi (Ed.), *Language and cognition: A developmental perspective*. Norwood, NJ: Ablex.

Putnam, H. (1977). Meaning and reference. In S. P. Schwartz (Ed.), *Naming, necessity, and natural kinds*. Ithaca: Cornell UP.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*, 1-33.

Smith, L. B. (1999). Children's noun learning: How general learning processes make specialized learning mechanisms. In B. MacWhinney (Ed.), *The emergence of language*. Mahwah, NJ: LEA.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard UP.

Welder, A. N., & Graham, S. A. (2001). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development, 72*, 1653-1673.

A Model of Spatio-Temporal Coding of Memory for Multidimensional Stimuli

Todd R. Johnson (Todd.R.Johnson@uth.tmc.edu)
Hongbin Wang (Hongbin.Wang@uth.tmc.edu)
Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu)
Yue Wang (Yue.Wang@uth.tmc.edu)
University of Texas Health Science Center at Houston
School of Health Information Sciences, 7000 Fannin Suite 600
Houston, TX 77030 USA

Abstract

This paper presents a model of memory for multidimensional stimuli. The model captures the independence of features in memory, their recovery using spatial location and temporal cues, and the role of verbal recoding in building integrative feature memories. The model fits data showing that object features may be retrieved independently when given a location cue, but that correct retrieval of missing features given a feature cue depends on the correct retrieval of location. The model also suggests that positional codes implicated in many memory models may be the result of the initial positional encoding of stimuli by perception.

Introduction

Although perception appears to integrate multidimensional stimuli, mounting evidence suggests that object features, including color, form, motion, orientation, texture and location are independently processed by our visual system and can even remain independent in memory (e.g., Healthcote, Walker, & Hitch, 1994). This paper reviews the evidence for the independence and re-integration of features in memory and proposes a model of the spatio-temporal coding of memory for multidimensional stimuli. The model is implemented as a modification of the ACT-R cognitive architecture (Anderson & Lebiere, 1998) and shown to fit the results of a representative experiment.

Feature Independence in Memory

Evidence for the independent encoding of features in memory typically involves conjunction errors in recall or recognition tests (Reinitz, Lammers, & Cochran, 1992). In a recognition test, a conjunction error occurs when a subject reports previously seeing a new stimulus that consists of a conjunction of features from old stimuli. In a recall test, a conjunction error occurs when subjects recall a stimulus that erroneously conjoins features of previously seen stimuli. Conjunction errors have been demonstrated for a variety of stimuli, including faces (Reinitz et al., 1992; Treisman, Sykes, & Galade, 1977), two syllable nonsense words (Reinitz et al., 1992), colored forms (Stefurak & Boynton, 1986), and colored bars at different orientations (Isenberg, Nissen, & Marchak, 1990). Presentation times for study stimuli in these experiments range from 100 ms to several minutes, hence the results show that features are independently stored in both short- and long-term memory.

Nissen (1985) reported an experiment that suggested that visual features of objects (in this case color and shape) are stored separately, but are indexed or bound by their spatial location. Subjects were presented with four different shapes, each of a different color, and each in one of four positions, followed by either a location or color cue. When given a location or color cue, subjects were told to report the other two values indexed by that cue (color and shape, and shape and location, respectively). Subjects were tested in separate location-cue and color-cue conditions with 64 unique trials in each condition. Color, shape, location and cue were systematically randomized so as to ensure statistical independence among the stimuli and cues.

Nissen found that when the cue was a location, correct recall of color and shape were statistically independent; however, when the cue was a color, correct recall of shape depended on correct recall of location. These results suggest that object features are represented independently, with each feature associated with the object's spatial location. Thus, retrieving the shape of an object given its color as a cue requires one to first retrieve the location containing an object with that color, followed by retrieving the shape at that location.

Nissen's results showing independence in the location-cue condition were questioned by Monheit and Johnston (1994) who argued that because of the effects of guessing, very little deviation from independence was possible. By increasing the number of colors and forms (using letters instead of shapes) they reduced the effects of guessing and increased the expected deviation from independence. They also increased the number of trials to increase the chance of detecting a smaller deviation from independence. In a series of experiments that were similar to Nissen's location-cue condition, they found consistent evidence for the dependence of color and shape given a location cue. They explained their results by arguing that selective attention to an object tightly binds all features, but that in the Nissen experiment subjects have only enough time to selectively attend to a subset of the objects. Features of attended objects tend to be reported correctly, whereas features of unattended objects must be guessed. The combination of correct conjunction trials and those involving guessing produced the amount of dependence observed in their experiments.

Despite Monheit and Johnston's results, Nissen's experiment still supports a special role for location in binding object features. Monheit and Johnston's critique of

Nissen's experiment should apply equally well to the color-cue condition, meaning that it would have been just as difficult to detect dependence. However, Nissen found dependence in both the aggregate data and 8 of the 9 individual subjects. The fact that Nissen's experiment was sufficient to detect dependency in the color-cue condition suggests that location still plays an important role in binding object features. However, it is possible that selective attention may increase the association among object features, making location less important in the recovery of feature conjunctions. Indeed, Wolfe and Cave (Wolfe & Cave, 1999) suggested that pre-attentive features are loosely bound, whereas features of attended objects are more tightly bound.

Additional evidence suggests that features may also be bound by temporal cues (Treisman, 1977). In one experiment a series of several letters, a number, and several more letters were rapidly presented either at the same location or were alternated above and below the fixation point (Keele, Cohen, Ivry, Liotti, & Yee, 1988). Subjects were told to report the color of the background surrounding the digit. When the items were presented at the same location, more errors came from reporting the color of letters at the –1 and +1 temporal positions, items that appeared just before and just after the target. However, when the items were alternated among two locations, more errors came from the –2 and +2 temporal positions, items that occurred at the same spatial location as the target, but that were temporally more distant than the –1 and +1 items. Based on the results of several similar experiments, the authors argued that spatial contiguity is the dominant requirement for binding features and that temporal contiguity is of use only when features appear in the same location. However, the dominance of location may be an artifact of the task. Other researchers have argued that subjects may use multiple strategies to recover feature conjunctions, depending on the available cues at study and test (Heathcote, Walker, & Hitch, 1994).

There is also evidence that conjunction errors are affected by the distance and similarity among stimuli. Several experiments have found that subjects are more likely to erroneously conjoin features of adjacent or similar stimuli (for a review see Ashby, Prinzmetal, Ivry, & Maddox, 1996).

Other lines of research have shown that verbalization can result in an integrated stimulus memory. In a recognition task using colored animal shapes and long presentation times, Stefurak and Boynton (1986) demonstrated that subjects had memory for feature conjunctions unless they were prevented from naming study stimuli by engaging in a secondary verbal task, in which case they appeared to have absolutely no memory of feature conjunctions. In addition to suppressing verbalization, their experimental task provided neither temporal nor location cues, because the study stimuli were presented simultaneously, and the test stimulus was not presented in its study location. As a result, the suppressed verbalization condition did not provide any of the cues (verbal, temporal, or spatial) that are thought to mediate feature integration.

Because verbalization appears to result in an integrated feature memory, it appears that verbal codes act in a different manner than spatial and temporal codes. Instead of acting as a tag for separate perceptual memories, it seems likely that the perceptual features are simply recoded as verbal cues. For example, the features "red" and "triangle" may be recoded as a verbal chunk "red triangle" that may be retrieved as a whole. Likewise, a display containing multiple stimuli may be recoded as a verbal list, such as "red triangle," "blue square" where each item is given a temporal position code.

To summarize, features of multidimensional stimuli appear to be represented independently in memory, but bound by temporal and spatial cues. Integrated feature representations are possible, but only if verbalization is possible.

ACT-R 5.0

Our model of feature integration in memory is embedded in the ACT-R 5 cognitive architecture (Bothell, 2002), where it adopts ACT-R's theory of memory and cognition (as described below), but slightly modifies ACT-R's perceptual system. This section describes ACT-R 5. Modifications needed to support the model are described in the next section.

Unlike previous versions of ACT-R, ACT-R 5 (hereafter called ACT-R) consists of several interacting, asynchronous modules for perception, cognition, memory, and action. The cognitive module consists of a procedural (production rule) long-term memory and a goal buffer that holds the current goal and goal-relevant information. The declarative memory module consists of declarative memory chunks and a retrieval buffer that holds the last item retrieved. Each declarative chunk has a unique identifier, a type, and zero or more attributes and values, such as:

Obj1 isa shape-map feature triangle location loc1

where "Obj1" is the identifier of the memory chunk, "shape-map" is the chunk type, "feature" and "location" are attributes, and "triangle" and "loc1" are their respective values.

The perceptual-motor module has subsystems for vision, hearing, speech production, and motor commands. The visual module has a buffer that holds the currently attended visual location and the visual stimulus at that location. It accepts commands from cognition (via production rules) to conduct visual search and shift visual attention. The motor module accepts commands from cognition to do simple computer-based physical tasks, such as moving the mouse to a certain location, pressing a mouse button, and typing commands.

Much of the coordination between perception and action is done by production rules. The condition side of a rule is limited to testing the buffers (including whether a particular

module is busy), whereas the action side can only initiate a limited set of actions that modify buffers or send commands to one of the other modules. When a rule fires, its action side initiates commands to the other modules, such as shifting visual attention or retrieving a red object from memory, after which the rule system is free to fire additional rules. The other modules in ACT-R handle these actions asynchronously, usually resulting (after some delay) in changes to the buffers. Rules can then detect these changes and take appropriate actions. Although more than one rule can match at a given time, ACT-R only fires one rule in each cycle. A psychologically realistic conflict-resolution mechanism, based on cost and probability of success, determines which of several matching rules will fire.

To understand how this works, suppose that ACT-R is given a cued recall task, where it must report a remembered shape with a cued color. Furthermore, assume that ACT-R is attending to a fixation point that changes to the cue word "red." When the visual system detects the change, it updates the visual buffer to indicate that the word "red" is now attended. A production rule that is conditioned on seeing a word in the visual buffer fires and initiates a memory recall request for a red shape, plus notes on the goal that such a request was initiated. As the declarative memory module begins to process this request, the rule system continues to check for and fire any matching rules. This allows ACT-R to engage in additional cognitive processing, including initiating commands to the perceptual-motor system, while the memory system processes the retrieval request. When the retrieval request is complete, the retrieval buffer is filled with either the newly retrieved chunk or an indication of a retrieval failure. Two separate rules, both sensitive to the goal annotation indicating the retrieval request, handle these possibilities. One rule tests for a shape in the retrieval buffer and initiates a speech command to say the name of the shape, the other rule tests for a retrieval failure and initiates a second retrieval to guess a shape.

To understand the model presented below, it is also necessary to understand how ACT-R processes retrieval requests. Retrieval requests specify a chunk type and one or more attribute-value pairs. The memory module returns the chunk of the specified type with the highest activation value, where activation of chunk i is determined by

$$A_i = B_i + \sum_j W_j S_{ji} + \sum_k P_k M_{ki} \qquad \text{(EQ 1)}$$

B_i is the base level activation of the chunk, reflecting how recently and frequently it has been retrieved. The first summation reflects associative priming of the chunk by chunks in the goal buffer, where W_j is the available activation and S_{ji} is the strength of association from chunk j to chunk i. W_j is typically set to 1/n, where n is the total number of chunks in the goal buffer. S_{ji} is initially set to S-ln(n), where S is a constant and n is the number of chunks that have chunk j as an attribute value. This setting produces the classic fan effect (Anderson, 1974).

The second summation in EQ 1 reflects similarity of the chunk i to the retrieval cue. M_{ki} is the similarity between the value of the *kth* attribute in the retrieval cue and the value in the corresponding attribute of chunk i. P_k (which defaults to 1) reflects the weighting given to the similarity of attribute k. By default, M_{ki} is 1 if the kth attribute value in the cue is identical to the corresponding value in chunk i, otherwise it is –10.

To model the random fluctuations of human memory, activations vary with time by adding noise as a logistic function of the parameter s, where s is related to the variance of the noise by

$$\sigma^2 = \frac{\pi^2}{3} s \qquad \text{(EQ 2)}$$

Finally, the activation threshold τ specifies the minimum activation value for retrieving a chunk. If all chunks matching the cue fall below this value, the retrieval request fails. As with chunks, the retrieval threshold varies from time to time according to the noise parameter s.

The approximate probability of retrieving a chunk i given k competitors (including the threshold and chunk i) is given by

$$P(i) = \frac{e^{A_i / s\sqrt{2}}}{\sum_k e^{A_k / s\sqrt{2}}} \qquad \text{(EQ 3)}$$

where A_n is the mean activation of chunk n.

A Model of Memory for Multidimensional Stimuli

The model assumes that attending to a multidimensional stimulus results in a set of feature chunks in memory, where each chunk encodes one feature along with one or more temporal and spatial tags. While attention is fixed on the stimulus these chunks also appear in the corresponding perceptual buffer. If a stimulus is recognized (either identified or classified or both), perception may also deliver a separate chunk encoding the identity (or class) of the stimulus along with spatial and temporal tags.

Suppose that ACT-R attends to a red square at location Loc22 on a computer screen at time t1. ACT-R's visual buffer is then filled with chunks encoding red at Loc22 t1, square(shape) at Loc22 t1, and square(class) at Loc22 t1. These same chunks are also added to ACT-R's declarative memory. This is shown graphically in Figure 1, where squares represent chunks and arrows indicate chunk attributes. Locations (e.g., loc22) are chunks that correspond to unique locations using the computer-screen as the frame of reference.

A spatial tag encodes where the feature occurs and may be given in any number of frames of reference (e.g., Wang, Johnson, Zhang, 2001). For instance, one spatial tag might give object heading and distance in egocentric (body-centered) coordinates, whereas another spatial tag might indicate the exocentric heading and bearing of the object from another object. Because ACT-R's perceptual-motor

system is designed to work with two-dimensional computer displays, the model provides a spatial tag relative to the frame of the display. Evidence for frame-relative location encoding has been found in both monkeys and humans. Rolls (1999) found that some neurons in the monkey hippocampus responded to where the monkey looked on a screen independent of the position of the monkey relative to the location of the screen. Hock, et al. (1989) showed that subjects unintentionally retained frame-relative locations of circles forming patterns in a frame, such that they could estimate the frequency with which circles appeared at a particular location within the frame.

Given enough time, rules may recode the features. For instance, a set of rules may verbalize the visual features "red" and "triangle" resulting in a redundant verbal code with appropriate temporal tags. Rules may also recode the features into an integrated representation, such as a single chunk that binds "red" and "triangle." Whether or not a stimulus is recoded, and the nature of the recoding, is dependent on the production rules, which in turn depend on the current goal and the strategy being used to achieve it.

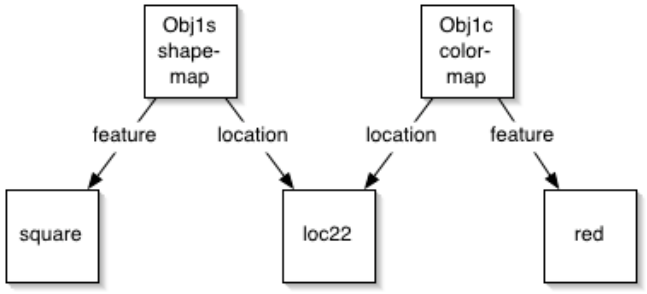


Figure 1. Representation of color, shape and location in the ACT-R model. Temporal cues are not shown.

The model assumes that the similarity (M in EQ 1) of temporal and spatial tags is inversely proportional to their temporal and spatial distance; however, the exact nature of this relationship is left to the model builder. As a result, the model will tend to confuse spatially adjacent features.

Applying the Model to the Nissen Task

As a partial test of the model, we applied it to the Nissen (1985) experiment described earlier. The critical phenomena in this experiment is that recall of color and shape is independent given a location cue, but when given a color cue, recall of shape is dependent on correct recall of location.

The ACT-R model contains production rules for attending to the four colored objects, attending to the cue, retrieving the answers, and pressing keys to record its responses. When presented with the 4 objects, the model visually attends to each object, resulting in automatic encoding of a color-map and shape-map chunk for each object. It then waits for the cue to appear, at which point it attends to the cue and begins the retrieval and response process.

The critical production rules for modeling the experimental results are those for retrieving location given a color, and those for retrieving color and shape given a location. When given a location cue the model first attempts to retrieve a chunk encoding the color at that location (such as obj1c in Figure 1), and then attempts to retrieve a chunk encoding the shape at the given location (such as obj1s). When given a color cue, the model attempts to retrieve a color-map chunk containing that color (e.g., obj1c). It then uses the location in this chunk to retrieve a chunk encoding the shape at that location.

If a rule fails to retrieve a chunk, the model will guess an appropriate value. For example, if the model fails in retrieving a color-map chunk with color red, it will simply guess a location, and then use that location when it attempts to retrieve the shape.

Fitting the Nissen data requires estimating the parameters in EQ 1, as well as the activation threshold, and the noise parameter. The activation threshold and the base level activation B_i for all stimulus chunks were set at 0—the default value. The amount of activation available for associative priming was also set at 0, because chunks in the goal buffer are redundant with attribute values in the retrieval cue. Similarity among matching attribute values, including locations, was set to 1 with mismatching values set to 0. The noise parameter s was the only parameter tuned to fit the data. We used EQ 3 and the results from the Nissen experiment to determine an initial value of s, then iteratively refined it over several model runs to produce the fit reported below (where s = 0.39).

A. Location-Cue Condition

		Shape: Correct	Shape: Incorrect	
Color: Correct		0.485 (0.450)	0.212 (0.219)	0.697 (0.669)
Color: Incorrect		0.213 (0.175)	0.090 (0.156)	0.303 (0.331)
		0.698 (0.625)	0.302 (0.375)	

B. Color-Cue Condition

		Shape: Correct	Shape: Incorrect	
Location: Correct		0.477 (0.494)	0.221 (0.234)	0.698 (0.728)
Location: Incorrect		0.032 (0.051)	0.271 (0.221)	0.303 (0.272)
		0.509 (0.545)	0.492 (0.455)	

Figure 2: Results of simulating 50 subjects for each condition. Values in parentheses are the experimental results from Nissen.

The results of running the model for 50 subjects in each of the two conditions are shown in Figure 2 along with the Nissen data. As expected, shape and color recall are statistically independent in the location-cue condition (χ^2 = 0.131, p = 0.72), whereas location and shape are dependent in the color-cue condition (χ^2 = 901.23, p < 0.01). The proportions of correct and incorrect recall across both conditions produce a good fit to the Nissen data: R^2= 0.95.

Conclusion

The model described in this paper can account for the basic phenomena of memory-based feature integration. The special role of location was demonstrated by applying the model to the Nissen task. The use of temporal cues, such as that reported by Keele, et al. and discussed earlier, is supported by the model's use of temporal tags for each feature. If location is given a stronger association to the features than is the temporal tag, this would produce Keele's results showing a role of temporal contiguity only for items that appear in the same spatial location. The tendency to erroneously conjoin spatially adjacent features and features of similar stimuli is captured in the model using ACT-R's theory of memory retrieval which tends to confuse similar stimuli (see the discussion of EQ 1). This means that features of spatially proximal objects will be confused more often than those of spatially distant objects. It also means that if the model is trying to recall the color of an oval, it would be more likely to confuse its color with that of a circle than with a square.

The effects of recoding, including verbalization, are captured in the model by assuming that the names of individual features or the name or semantic identity of an object may be memorized instead of the individual visual features. If there is enough time for object identification, the subject need only remember an object's identity and location during study. At test, the object's identification provides a cue for reporting essential object features. Such recoding will produce integrated memories, because the individual features are already well-learned and "bound" to the object identity. If a subject remembers seeing a banana, conceptual knowledge of bananas is sufficient to recall that it was yellow and crescent shaped—there is no need to encode the specific perceptual features in a new memory trace.

Such a strategy will not work, however, if the task presents bananas in unnatural colors. In this case, verbal rehearsal and the resulting verbal memory may be of use. For instance, given enough time a subject might remember objects in the Nissen task by verbally rehearsing "red triangle, blue square…" and so on. Recall of these items would then be subject to serial recall effects, such as the serial position curve and positional errors. It seems likely that such a strategy would result in fewer conjunction errors, which would explain why verbalization results in integrated feature memory.

The model provides a possible explanation for the need to use positional codes (instead of integrated codes or associative chaining) in cognitive models of memory tasks. Positional codes were used to account for chunk position effects in alphabetic retrieval response times (Klahr, Chase, & Lovelace, 1983) and positional errors in serial recall (Anderson, Bothell, Lebiere, & Matessa, 1998). It is possible that these codes may be the direct result of the positional (temporal or spatial) encoding of stimuli by perceptual processes.

One limitation of the model is that it treats all errors as memory retrieval errors. However, the model could be extended to include probabilities for correctly perceiving features, or a theory of feature perception. However, since our emphasis is on the representation of features in memory and their later reintegration, we saw no need to introduce additional theory.

Monheit and Johnston's demonstration of dependence of color and shape given a location cue provides a challenge to the model presented here. To account for the dependence our model must be modified to provide for some strength of association between features of attended objects. In the present model, knowing the color of an object does not activate the object's shape (e.g., the strength of association between color and shape, S_{ji} in EQ 1, is 0). In the revised model, the color of a previously attended object would activate its shape, allowing for some dependence among features of objects. This would make our model consistent with Wolfe and Cave's view that preattentive features are loosely bound, whereas features of objects that have been attended are more tightly bound.

Finally, the model is meant to provide a foundation for a comprehensive theory of spatial cognition embedded in ACT-R. By embedding the model in ACT-R other researchers can use it in their models, where it may provide additional constraints and enable more realistic memory representations and behavioral predictions.

Acknowledgments

This research was supported by the Office of Naval Research, Cognitive Science Program under Grant No. N00014-01-1-0074.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6*, 451-474.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*, 341-380.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Ashby, F. G., Prinzmetal, W., Ivry, R., & Maddox, W. T. (1996). A formal theory of feature binding in object perception. *Psychological Review, 103*(1), 165-192.

Bothell, D. (2002). *Act-R 5.0 Beta*. Retrieved February 6, 2002, from http://act.psy.cmu.edu/ACT-R_5.0/

Heathcote, D., Walker, P., & Hitch, G. J. (1994). Feature independence and the recovery of feature conjunctions. *The Journal of General Psychology, 121*(3), 253-266.

Hock, H. S., Smith, L. B., Escoffery, L., Bates, A., & Field, L. (1989). Evidence for the abstractive encoding of superficial position information in visual patterns. *Memory & Cognition, 17*(4), 490-502.

Isenberg, L., Nissen, M. J., & Marchak, L. C. (1990). Attentional processing and the independence of color and orientation. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 869-878.

Keele, S. W., Cohen, A., Ivry, R., Liotti, M., & Yee, P. (1988). Tests of a theory of attentional binding. *Journal of Experimental Psychology-Human Perception and Performance, 14*(3), 444-452.

Klahr, D., Chase, W. G., & Lovelace, E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Pschology: Learning, Memory, and Cognition, 9*(3), 462-477.

Monheit, M. A., & Johnston, J. C. (1994). Spatial attention to arrays of multidimensional objects. *J Exp Psychol Hum Percept Perform, 20*(4), 691-708.

Nissen, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and Performance XI* (pp. 205-219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reinitz, M. T., Lammers, W. J., & Cochran, B. P. (1992). Memory-conjunction errors: Miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition, 20*(1), 1-11.

Rolls, E. T. (1999). The representation of space in the primate hippocampus, and its role in memory. In N. Burgess, K. J. Jeffery & J. O'Keefe (Eds.), *The hippocampal and parietal foundations of spatial cognition* (pp. 320-344). Oxford: Oxford.

Stefurak, D. L., & Boynton, R. M. (1986). Independence of memory for categorically different colors and shapes. *Perception & Psychophysics, 39*(3), 164-174.

Treisman, A. (1977). Focussed attention in the perception and retrieval of multidimensional stimuli. *Perception & Psychophysics, 22*, 1-11.

Treisman, A., Sykes, M., & Galade, G. (1977). Selective attention and stimulus integration. In S. Dornic (Ed.), *Atention and performance VI* (pp. 333-361). Hillsdale, NJ: Erlbaum.

Wolfe, J. M., & Cave, K. R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron, 24*(1), 11-17, 111-125.

Wang, H., Johnson, T. R., Zhang, J. (2001). The mind's views of space. In *proceedings of the 4th International Conference of Cognitive Science.*

Analysis of the Dynamics of Reasoning Using Multiple Representations

Catholijn M. Jonker[1] **and Jan Treur**[1,2] **({jonker treur}@cs.vu.nl)**
[1] Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands URL: http://www.cs.vu.nl/~{jonker, treur}
[2] Universiteit Utrecht, Department of Philosophy
Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

Abstract

This paper presents a formalisation and analysis method for the dynamics of a reasoning process in which multiple representations play a role. Dynamics of reasoning processes are described by reasoning traces consisting of sequences of reasoning states over time. Reasoning states have a compositional structure; they are composed of different parts, for example, for different representations. Transitions between two reasoning states model reasoning steps. In relation to the compositional structure of the states, transitions are classified into a number of types. An example reasoning process involving multiple representations is used to illustrate how its dynamics can be formalised and analysed using the approach.

Introduction

Within Cognitive Science in recent years the dynamical perspective on cognitive phenomena has been emphasized and received much attention. In most literature focussing on the dynamics of cognition, the Dynamical Systems Theory (DST) is taken as a point of departure; e.g., (Port and Gelder, 1995). This theory assumes that, in contrast to the use of symbolic representations, modelling and analysis of dynamics of cognitive phenomena can be done more effectively by using representations based on real numbers and mathematical techniques, in particular difference and differential equations. The convincing examples illustrating the usefulness of this perspective often address lower level cognitive processes such as sensory or motor processing. Indeed one of the advantages of the Dynamical Systems Theory is that it is able to model the temporal aspects of events taking place on a continuous time scale, such as, for example, recognition time, response time, and time involved in motor patterns and locomotion.

Also some examples of higher level cognitive processes have been addressed using DST; for example the dynamic models for decision making developed by Busemeyer and Townsend (1993). Especially the continuous adaptive aspects of the decision making are covered nicely in this approach. Areas for which the quantitative approach based on DST is assumed to have less to offer are the dynamics of higher level processes with mainly a qualitative character, such as certain capabilities of language processing and reasoning. In the last two decades, within the areas of Computer Science and Artificial Intelligence alternative techniques have been developed to analyse the dynamics of phenomena using qualitative means. Examples are process algebra; transition systems; dynamic and temporal logic; event, situation and fluent calculus; e.g., (Eck, et al. 2001; Hölldobler and Tielscher, 1990; Kowalski and Sergot, 1986; Reiter, 2001). Just as difference or differential equations, these alternative techniques allow to consider and relate states of a process at different points in time. The form in which these relations are expressed can cover both quantitative and non-quantitative aspects. This paper illustrates the usefulness of such an approach for the analysis and formalisation of the dynamics of reasoning. Here a broad perspective is taken on reasoning, subsuming, for example, reasoning involving multiple representations.

A formal analysis method for the dynamics of reasoning is presented and illustrated by an example reasoning pattern involving geometric and arithmetic representations. This pattern is analysed and characterised in terms of a set of dynamic properties. The properties have been formalized, thus enabling automated support of analysis by an analysis environment that has been developed.

Below, first the dynamic perspective on reasoning is discussed in some more detail. Next, the example reasoning pattern is introduced, and the first steps of an analysis are made. Third, a number of dynamic properties identified for the example reasoning pattern are presented. Finally the analysis method is summarised and the contribution of the research presented in the paper is discussed.

Reasoning Dynamics

Analysis of the cognitive capability to perform reasoning has been addressed from different areas and angles. Within Cognitive Science, the two dominant streams are the syntactic approach (based on inference rules applied to syntactic expressions, as common in logic), e.g., (Rips, 1994), and the semantic approach (based on construction of mental models); e.g., (Johnson-Laird, 1983; Yang and Johnson-Laird, 1999).

Reasoning steps in natural contexts are usually not restricted to the application of logical inference rules. For example, a step in a reasoning process may involve translation of information from one representation form (e.g., geometrical) into another one (e.g., arithmetical). Or, an additional assumption can be made, thus using a dynamic set of premises within the reasoning process. Decisions made at specific points in time during the process, for example, on which representations to use or which assumptions to make, are an inherent part of the reasoning. Such reasoning processes or their outcomes cannot be

understood, justified or explained without taking into account these dynamic aspects.

To formalise the dynamics of a reasoning process, traces are used. *Reasoning traces* are time-indexed sequences of *reasoning states* over a time frame; for stepwise reasoning processes the set of natural numbers as a time frame is an appropriate choice. The set of all possible reasoning states defines the space where the reasoning takes place. Reasoning traces can be viewed as trajectories in this space, for which every (reasoning) step from one reasoning state to the next one is based on an *allowed transition*. If the possible reasoning states and the allowed reasoning steps or transitions are characterised, the set of proper reasoning traces can be defined as the set of all possible sequences of reasoning states consisting only of allowed transitions.

Reasoning States

A *reasoning state* formalises an intermediate state of a reasoning process. The content of such a reasoning state usually can be analysed according to different aspects or dimensions. For example part of the state may contain a geometric representation, another part an arithmetic representation. Accordingly, the reasoning state is structured as a composition of (i.e., a tuple of) a number of parts, indexed by some set I. This index set includes different aspects or views taken on the state, e.g., I = {geometric, arithmetic}. The set of reasoning states RS can be characterised as a Cartesian product $RS = \Pi_{i \in I} RS_i$ where RS_i is the set of all states for the aspect indicated by i. For example, $RS_{geometric}$ may denote the set of all possible geometric representations; note, however, that is also possible to use more dimensions, e.g., different types of geometric representations can be formalised. This Cartesian product formalises the multi-dimensional space where the reasoning takes place. For a reasoning state, which is a vector $S = (S_i)_{i \in I} \in RS$ in this space, the S_i are called its *components*.

Reasoning Steps: Transitions of Reasoning States

A transition from one reasoning state to another reasoning state, i.e., an element $< S, S' >$ of $RS \times RS$, formalises one *reasoning step*; sometimes also denoted by $S \rightarrow S'$. A *reasoning transition relation* is a relation on $RS \times RS$. Such a relation can be used to specify the *allowed transitions*. Transitions differ in the set of components that are involved. The most complex transitions change all components of the state in one step. However, within stepwise reasoning processes, usually transitions only involve a limited number of components of the state, e.g., only one or two. Transitions can be classified according to which set of components is involved. The most simple types of transition involve a single component transition. Next come transition types where two components are involved. In the current approach we concentrate on these two classes of transition types.

Single component transition types

For example, when a modification in the reasoning state is made solely within a geometric representation, only the geometric component of the state changes (geometric reasoning step). Or, if a calculation (arithmetic reasoning) step is performed, only the aritmetic component is changing. These single component transitions involve only that component and can be defined within one component only:

geometric $\rightarrow$ geometric
arithmetic $\rightarrow$ arithmetic

It is also possible that one component of a state is changed by information acquisition, importing information from an external source in the reasoning process.

Transitions involving two components of a reasoning state

Other types of transitions involve more than one component. For example, if information from a geometric representation is translated into an arithmetic form, thereby extending the arithmetic representation, then two components of the state are involved: the arithmetic component and the geometric component. Examples of transition types involving two components are:

geometric x arithmetic $\rightarrow$ geometric

(e.g., the geometric representation is extended or modified with results from the arithmetical representation)

arithmetic x geometric $\rightarrow$ arithmetic

(e.g., the arithmetic representation is extended or modified with results from the geometrical representation)

Reasoning Traces

Reasoning dynamics results from successive reasoning steps, i.e., successive transitions from one reasoning state to another. Thus a *reasoning trace* is constructed: a time-indexed sequence of reasoning states $(\gamma_t)_{t \in T}$, where T is the time frame used (the natural numbers). A reasoning trace can be viewed as a trajectory in the multi-dimensional space $RS = \Pi_{i \in I} RS_i$ of reasoning states. An example of such a reasoning trace will be discussed in Section 3; see also Figure 1. Reasoning traces are sequences of reasoning states subject to the constraint that each pair of successive reasoning states in this trace forms an allowed transition. A trace formalises one specific line of reasoning.

Example Reasoning Process

An example multi-representation reasoning process is used to illustrate the approach put forward: interaction between arithmetical reasoning and geometrical reasoning. The example focuses on how to determine the outcome of multiplications such as 23×36. Experiences on using such processes with children (8-9 years old) in class rooms have been reported, e.g., by Dekker et al. (1982), see also (Hutton, 1977). The example can also be extended to an example for children of 13 or 14 years to support algebra by geometric visualisations, e.g., the algebraic identity $(a + b)^2 = a^2 + 2ab + b^2$ interpreted as the area of a partitioned square of $(a + b) \times (a + b)$ in relation to areas of its parts: a square of $a \times a$, a square of $b \times b$, and two rectangles of $a \times b$. Also teaching quadratic equations can be supported by such visualisations as discussed, e.g., by Bruner (1968), pp. 59-63. The example pattern shows two types of one component transitions of reasoning states, and two transition types involving two components:

- an *arithmetical* reasoning step: arithmetic → arithmetic
- a *geometrical* reasoning step: geometric → geometric
- a *translation* of an arithmetical representation into a geometrical representation: geometric x arithmetic → geometric
- a *translation* of a geometrical representation into an arithmetical representation: arithmetic x geometric → arithmetic

The idea is that only simple arithmetical steps are required. The more complicated steps are performed via the geometrical representation. A number of skills are assumed. These skills can be defined in the form of transitions.

A. *Assumed arithmetic skills* arithmetic → arithmetic

aa1. splitting a number in 'tens' and single digits: 28 = 20 + 8
aa2. addition of a list of numbers of up to 4 digits, such as 1200 + 340 + 120 + 6
aa3. multiplication of two numbers starting with a nonzero digit, followed by zero or more zeros, such as 20 x 8, 60 x 30.

B. *Assumed geometric skills* geometric → geometric

gg1. partitioning a rectangle in non-overlapping areas based on partitionings of its sides
gg2. determining the area of a figure from the areas of a (non-overlapping) partition

C. *Assumed translation skills*

geometric x arithmetic → geometric:

ag1. drawing a rectangle with arithmetically given dimensions
ag2. partitioning a line segment according to a splitting of its length
ag3. determining the area of a rectangle from the multiplication of the lengths of its sides

arithmetic x geometric → arithmetic:

ga1. translating the area of a rectangle into the multiplication of the lengths of its sides
ga2. translating the area of a combination of nonoverlapping areas into the sum of the areas

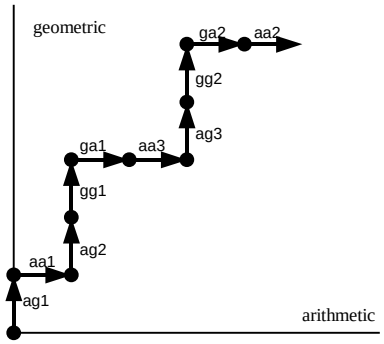


Figure 1: Reasoning trace as a trajectory in a two-dimensional reasoning state space.

The example reasoning trace, based on class room observation (cf. Dekker et al., 1982), forms a trajectory in the two-dimensional reasoning state space

$$RS = RS_{arithmetic} \times RS_{geometric}$$

This trajectory is depicted in Figure 1. Note that in this Figure only the changing component is visualised by an arrow, not what component affected this change. Therefore, e.g., both a geometric reasoning step and a translation of an arithmetic into a geometric representation are depicted by a vertical arrow. The detailed trace is presented below.

Starting problem What is the outcome of the multiplication 23 x 36 ?

Step 1 ag1 representation translation
Create a rectangle of 23 x 36.

Step 2 aa1 arithmetic reasoning
Split the numbers into the 'tens' and single digits: 23 = 20 + 3; 36 = 30 + 6

Step 3 ag2 representation translation
Translation of the arithmetical splitting of the numbers into partitions of the sides within the geometrical representation.

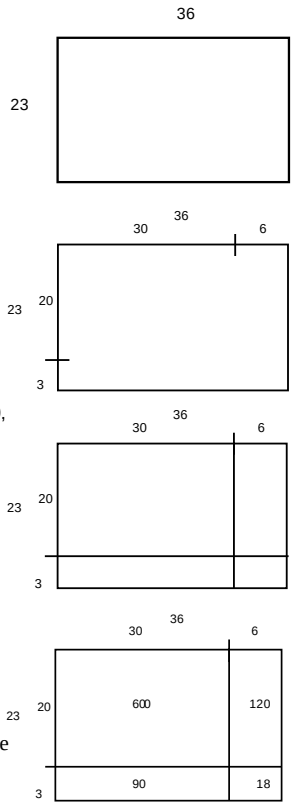
Step 4 gg1 geometric reasoning
Partition the area of the rectangle according to the partitioning of the sides.

Step 5 ga1 representation translation
For each part identify the correspond-ing arithmetical expression for its area: 20 x 30, 20 x 6, 3 x 30, 3 x 6

Step 6 aa3 arithmetic reasoning
Determine the outcomes of the four multiplications 20 x 30 = 600; 20 x 6 = 120; 3 x 30 = 90; 3 x 6 = 18

Step 7 ag3 representation translation
Identify the areas of the parts of the rectangle based on the outcomes of the multiplications.

Step 8 gg2 geometric reasoning
Assert that the area of the rectangle as a whole is the combination of the areas of the parts



Step 9 ga2 representation translation
Identify the corresponding arithmetical relation: 600 + 120 + 90 + 18

Step 10 aa2 arithmetic reasoning
Calculate the sum: 600 + 120 + 90 + 18 = 828

Dynamic Properties

To specify properties on the dynamics of a reasoning process, the temporal trace language TTL used by Herlea et al. (1999), and Jonker and Treur (1998) is adopted. This is a language in the family of languages to which also situation calculus (Reiter, 2001), event calculus (Kowalski and Sergot, 1986), and fluent calculus (Hölldobler and Tielscher, 1990) belong. In short, in TTL it is possible to express that in a given trace at a certain point in time the reasoning state has a certain (state) property. Moreover, it is possible to relate such state properties at different points in time. As an example, the following (global) property of a reasoning trace γ is considered, which expresses that all multiplication problems in two digits eventually will be solved.

GP1

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of a multiplication problem for numbers x and y < 100 is present,
then a time point t'≥t exists such that in the reasoning state in γ at t' an arithmetic representation of a solution z of this multiplication problem with z = x∗y is included.

The formalisation of this property in TTL is as follows.

$\forall t\ \forall x, y < 100$ state(γ, t, arithmetic) |== multiplication_problem(x, y)
$\Rightarrow\ \exists t' \geq t\ \exists z$ $z = x*y$ &
state(γ, t', arithmetic) |== is_solution_for_multiplication_of(z, x, y)

Note that for simplicity no maximal allowed response time has been specified. If desired, this can be simply added by putting a condition $t' \leq r$ in the consequent with r the maximal response time.

Milestone Properties

Within the overall reasoning process a number of milestones can be defined, and properties can be identified that express whether the process from one milestone to another one has been performed properly. Apart from the start and the finish, two intermediate milestones were defined: a reasoning state in which the problem has been represented in a geometric representation and it has been decomposed geometrically (after step 4 in the example trace), and a reasoning state in which a geometric representation with numbers in the areas occurs, i.e., in which the subproblems have been solved (after step 7 in the example trace). Accordingly, the following milestone properties have been formulated.

MP1

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of a multiplication problem for numbers x and y < 100 is present,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' a geometric representation of a rectangle ABCD is included with points P on AB and Q on AD, with |AB| = x and |AD| = y
and this rectangle is partitioned into four areas A_{11}, A_{12}, A_{21}, A_{22} by two lines PP'//AD and QQ'//AB with P' on CD and Q' on BC with $|AP| = x_1$, $|PB| = x_2$, $|AQ| = y_1$, and $|QD| = y_2$, where x_1, y_1 is the 10-part of x, resp. y, and x_2, y_2 is the digit part of x, resp. y.

Here, |AB| is the length of AB, and // is 'in parallel with'.

MP2

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included with points P on AB and Q on AD, with |AB| = x and |AD| = y,
and this rectangle is partitioned into four areas A_{11}, A_{12}, A_{21}, A_{22} by two lines PP'//AD and QQ'//AB with P' on CD and Q' on BC with $|AP| = x_1$, $|PB| = x_2$, $|AQ| = y_1$, and $|QD| = y_2$, where x_1, y_1 is the 10-part of x, resp. y, and x_2, y_2 is the digit part of x, resp. y,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' in each of these areas Aij a number zij is represented which equals $x_i * y_j$.

MP3

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included with |AB| = x and |AD| = y
and this rectangle is partitioned into four nonoverlapping rectangle areas A_{11}, A_{12}, A_{21}, A_{22},
and in each of these areas A_{ij} a number zij is represented which equals $x_i * y_j$, where $x = x_1 + x_2$, and $y = y_1 + y_2$,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' an arithmetic representation of a solution z with $z = x*y$ of the multiplication problem (x, y) is included.

Local Properties

In this section a number of properties are identified that characterise the reasoning in a more local manner: each property characterises one reasoning step. For the sake of simplicity, for the example reasoning process persistence of representations in reasoning states over time is assumed, so that persistence does not need to be formulated within each of the properties.

LP1 (arithmetic-geometric)

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of a multiplication problem for numbers x and y < 100 is present,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' a geometric representation of a rectangle ABCD with |AB| = x and |AD| = y is included.

This dynamic property expresses that in reasoning trace γ, if an arithmetically represented multiplication problem occurs, this eventually is translated into a geometric representation. The formalisation of this property in TTL is as follows.

$\forall t\ \forall x, y < 100$ state(γ, t, arithmetic) |== multiplication_problem(x, y)
$\Rightarrow\ \exists t' \geq t\ \exists A, B, C, D$
state(γ, t', geometric) |== rectangle(A, B, C, D) & |AB| = x & |AD| = y

Further local properties are the following (not in any particular order).

LP2 (arithmetic-arithmetic)

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of a multiplication problem for numbers x and y < 100 is present,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' an arithmetic representation of a splitting of the numbers x and y in 'tens' and digits occurs, i.e., $x = x_1 + x_2$, $y = y_1 + y_2$ with x_1, y_1 multiples of 10 and x_2, y_2 <10.

LP3 (arithmetic-arithmetic)

at any point in time t
if the reasoning state in trace γ at t contains an arithmetic representation of a multiplication problem for (x, y), with x, y multiple of 10 or less than 10,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' an arithmetic representation of a solution z with $z = x*y$ for this multiplication problem for (x, y) is included.

LP4 (arithmetic-arithmetic)

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of an addition problem for a finite list $z_1, ..., z_n$ of numbers of up to 4 digits is included,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' a solution $z = \Sigma_{1 \leq i \leq n} z_i$ of the addition problem is included.

LP5 (arithmetic-geometric)

at any point in time t
if in the reasoning state in trace γ at t an arithmetic representation of a splitting of the numbers x and y occurs, i.e.,
$x = x_1 + x_2$, $y = y_1 + y_2$,
then a time point $t' \geq t$ exists such that in the reasoning state in γ at t' a geometric representation of a rectangle ABCD with |AB| = x and |AD| = y is included with points P on AB and Q on AD such that $|AP| = x_1$, $|PB| = x_2$, $|AQ| = y_1$, and $|QD| = y_2$.

LP6 (geometric-geometric)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included with points P on AB and Q on AD,
then a time point t'≥t exists such that in the reasoning state in γ at t' the rectangle ABCD is partitioned into four areas A_{11}, A_{12}, A_{21}, A_{22} by two lines PP'//AD and QQ'//AB with P' on CD and Q' on BC.

LP7 (geometric-geometric)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included that is partioned into a number of nonoverlapping areas A_1, …, A_n,
then a time point t'≥t exists such that in the reasoning state in γ at t' it is asserted that the area of ABCD is the combination of the areas A_1, …, A_n.

LP8 (geometric-arithmetic)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD with |AB| = x and |AD| = y is included with points P on AB and Q on AD such that |AP| = x_1, |PB| = x_2, |AQ| = y_1, and |QD| = y_2,
and this rectangle is partioned into four areas A_{11}, A_{12}, A_{21}, A_{22} by two lines PP'//AD and QQ'//AB with P' on CD and Q' on BC,
then a time point t'≥t exists such that in the reasoning state in γ at t' arithmetic representations of multiplication problems for (x_1, y_1), (x_1, y_2), (x_2, y_1), and (x_2, y_2) are included.

LP9 (geometric&arithmetic-geometric)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included with points P on AB and Q on AD,
and this rectangle is partioned into four areas A_{11}, A_{12}, A_{21}, A_{22} by two lines PP'//AD and QQ'//AB with P' on CD and Q' on BC,
and arithmetic representations of solutions z_{11}, z_{12}, z_{21}, z_{22} for the multiplication problems for (|AP|, |AQ|), (|AP|, |QD|), (|PB|, |AQ|), and (|PB|, |QD|) are included.
then a time point t'≥t exists such that in the reasoning state in γ at t' within the geometric representation in each area A_{ij}, the number z_{ij} is represented.

LP10 (geometric-arithmetic)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included which is partioned into a number of areas A_1, …, A_n,
and within each of these areas A_i a number z_i is represented,
then a time point t'≥t exists such that in the reasoning state in γ at t' an arithmetic representation of an addition problem for z_1, …, z_n is included.

LP11 (geometric& arithmetic-arithmetic)

at any point in time t
if in the reasoning state in trace γ at t a geometric representation of a rectangle ABCD is included with |AB| = x and |AD| = y that is partitioned into a number of nonoverlapping areas A_1, …, A_n,
and within each of these areas A_i the number zi is represented,
and an arithmetic representation of a solution z of the addition problem for z_1, …, z_n is included,
then a time point t'≥t exists such that in the reasoning state in γ at t' an arithmetic representation of a solution z with z = x∗y of the multiplication problem (x, y) is included.

Relationships Between the Dynamic Properties

A number of logical relationships have been established between the properties above. First of all, the three milestone properties together imply the global property:

MP1 & MP2 & MP3 ⇒ GP1 (0)

Next, each of these milestone properties is implied by a number of local properties:

LP1 & LP2 & LP5 & LP6 ⇒ MP1 (1)
LP3 & LP8 & LP9 ⇒ MP2 (2)
LP4 & LP7 & LP10 & LP11 ⇒ MP3 (3)

These logical relationships, which can be depicted as an AND-tree, are helpful in the analysis of errors within a given reasoning trace. First it can be checked whether **GP1** holds. If this global property does not hold, the three properties **MP1**, **MP2**, **MP3** can be checked. Given the logical relationship (0), at least one of them will be found not to hold. This pinpoints the cause of the error in part of the process, say **MP3**. Next, (only) the local properties relating to **MP3** are checked, i.e, **LP4**, **LP7**, **LP10**, **LP11**. Again, due to (3) one of them will be found not to hold. This localises the error.

The Dynamic Analysis Method

The analysis method for the dynamics of reasoning processes as presented here is summarised as follows.

1. Identify the different dimensions or *components* of reasoning states.
2. Determine the different *types of transitions.*
3. Identify relevant *dynamic properties* for the reasoning
   a. for the process as a whole (global properties)
   b. for milestones within the process
   c. for reasoning steps (local properties)
4. Determine logical *relationships* between the different dynamic properties, in an AND-tree form; e.g.,
   a. local properties imply a milestone property, and
   b. milestone properties imply a global property.
5. For a given reasoning trace, *check* which of the dynamic properties hold and which do not hold. This can take the form of a diagnosis following the tree structure of the relationships between the dynamic properties. A software environment is available to support this checking process.

The dynamic properties identified can be of different types. Some may be assumed to hold for all proper reasoning traces, others may be used to distinguish different types of reasoning traces or reasoners.

Discussion

The analysis method for the dynamics of reasoning processes put forward and illustrated in this paper was validated on the basis of reports from experiments with 8-9 year old children in classrooms in the Netherlands (Dekker et al., 1982); a similar report has been made by Hutton

(1977). This paper shows how an analysis of these dynamics can be made using traces consisting of sequences of reasoning states over time to describe reasoning processes. It is shown for the example reasoning pattern, how characterising dynamic properties can be identified.

The language used to express dynamic allows for precise specification of these dynamic properties, covering both qualitative and quantitative aspects of states and their temporal relations. Moreover, software tools have been developed to (1) support specification of dynamic properties, and (2) automatically check specified dynamic properties against example traces to find out whether the properties hold for the traces. This provides a useful supporting software environment to evaluate empirical data on the dynamics of reasoning processes.

The same analytic method and software tools can also be applied to reasoning traces produced by software simulation models. This applicability supports the comparison of human reasoning with simulated reasoning.

Further experiments will be conducted, in which also a focus is on the control of the reasoning. For example, at what point in time a translation to a geometric representation is made, and why at that point in time? In the analysis the notion of reasoning strategy will be addressed. Due to the compositional structure of reasoning state a reasoning state can be extended with a component for control information.

Future research will also address the analysis of the dynamics of other types of practical reasoning, both from the syntactical and semantical stream, or their combination; e.g., (Johnson-Laird, 1983; Yang and Johnson-Laird, 1999; Yang and Bringsjord, 2001); see also (Stenning and Lambalgen, 2001). One component of the reasoning state may contain a syntactic formula structure, and another component a mental model or set of mental models. For example, a single component transition can be defined within a syntactic component including A and $A \rightarrow B$, for the derivation of B (and hence adding it to the component) based on the inference rule modus ponens. Yet another example, within a semantic component is a transition of a set of mental models, thus providing a formalisation of the dynamics of reasoning based on mental models.

References

Bruner, J.S. (1968). *Toward a Theory of Instruction*. Norton & Company, Inc. New York.

Busemeyer, J., and Townsend, J.T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, vol. 100, pp. 432-459.

Dekker, A., Heege, H. ter, and Treffers, A. (1982). *Cijferend vermenigvuldigen en delen volgens Wiskobas*. Universiteit Utrecht, Freudenthal Institute.

Eck, P.A.T. van, Engelfriet, J., Fensel, D., Harmelen, F. van, Venema, Y. and Willems, M. (2001). A Survey of Languages for Specifying Dynamics: A Knowledge Engineering Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 13(3):462-496, May/June 2001.

Herlea, D.E., Jonker, C.M., Treur, J., and Wijngaards, N.J.E. (1999). Specification of Behavioural Requirements within Compositional Multi-Agent System Design. In: F.J. Garijo, M. Boman (eds.), *Multi-Agent System Engineering, Proc. of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99*. Lecture Notes in AI, vol. 1647, Springer Verlag, 1999, pp. 8-27.

Hölldobler, S., and Thielscher, M. (1990). A new deductive approach to planning. *New Generation Computing*, 8:225-244, 1990.

Hutton, J. (1977). Memoirs of a Maths Teacher 5: Logical Reasoning. In: *Mathematics Teaching*, vol. 81, pp. 8-12.

Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Jonker, C.M., and Treur, J. (1998). Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. In: W.P. de Roever, H. Langmaack, A. Pnueli (eds.), *Proceedings of the International Workshop on Compositionality, COMPOS'97*. Lecture Notes in Computer Science, vol. 1536, Springer Verlag, 1998, pp. 350-380. Extended version in: *International Journal of Cooperative Information Systems.* To appear, 2002.

Kowalski, R., and Sergot, M. (1986). A logic-based calculus of events. *New Generation Computing*, 4:67-95, 1986.

Port, R.F., Gelder, T. van (eds.) (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, Mass.

Reiter, R. (2001). *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, 2001.

Rips, L.J. (1994). *The Psychology of Proof: Deductive reasoning in human thinking*. MIT Pres, Cambridge, Mass.

Stenning, K., and Lambalgen, M. van (2001). Semantics as a foundation for Psychology: A Case Study of Wason's Selection Task. *Journal of Logic, Language and Information*, vol. 10, pp. 273-317.

Yang, Y., and Johnson-Laird, P.N. (1999). A study of complex reasoning: The case GRE 'logical' problems. In M. A. Gernsbacher & S. J. Derry (Eds.) *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 767-771.

Yang, Y., and Bringsjord, S. (2001). Mental MetaLogic: a New Paradigm in Psychology of Reasoning. In: L. Chen, Y. Zhuo (eds.), *Proc. of the Third International Conference on Cognitive Science, ICCS 2001*. Beijing, pp. 199-204.

Cue Abstraction and Exemplars in Multiple-Cue Judgment

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Henrik Olsson (henrik.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Anna-Carin Olsson (anna-carin.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Abstract

Although categorization and multiple-cue judgment are similar tasks, categorization models emphasize *exemplar memory*, while multiple cue judgment routinely is interpreted in terms of mental integration of cue weights that are abstracted in training. We investigate if these conclusions derive from genuine differences in the processes in the two tasks or are accidental to different research methods. The results reveal large individual differences and a shift from exemplar memory to mental cue-abstraction when the criterion is changed from classification to continuous. This suggests that people switch between qualitatively distinct processes in the two tasks.

Introduction

A categorization task typically requires a probe described by a number of binary *features* to be classified into one, of usually two, *categories*. A multiple-cue judgment involves a probe defined by binary or continuous *cues* and typically requires judgment of a continuous *criterion*. Both tasks require inference from known variables to an unknown variable. Despite the structural similarity of the tasks (Figure 1), the most successful cognitive models in the two domains are profoundly different in terms of the computations, cognitive processing, and neural substrate that they imply. Research on categorization often emphasize *exemplar memory* (e.g., Nosofsky & Johansen, 2000): retrieval of memory traces of concrete objects from different categories. In research on multiple-cue judgment, the (explicit or implicit) interpretation is generally that people retrieve abstracted knowledge of cue weights, which is then mentally integrated to perform a judgment (e.g., Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

In this article, we report an investigation into the reasons for these divergent conclusions. From the outset, we can identify two possible answers. The first is that research on multiple-cue judgment has not benefited from the designs and the cognitive modeling needed to disclose the importance of exemplar memory. From this point of view, the conclusions are accidental to different research paradigms and once that we scrutinize the processes carefully we find that they are essentially the same. A second answer is that the different conclusions derive from the differences that nonetheless distinguish the two tasks; for example, the use of a binary criterion in categorization tasks and a continuous criterion in multiple cue judgment tasks. The latter answer suggests a cognitive system with multiple levels of qualitatively distinct representations that compete to control behavior depending on the requirements (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Jones, Juslin, Olsson, & Winman, 2000; Juslin, Olsson, & Olsson, 2002).

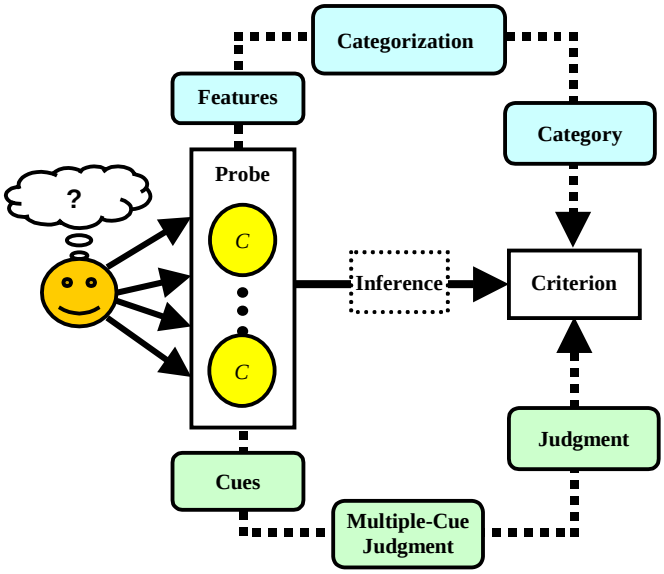


Figure 1: The structural similarity between a categorization task and a multiple-cue judgment task.

The Judgment Task

The task requires participants to use four binary cues to infer a binary or continuous criterion (Jones et al., 2000; Juslin et al., 2002). The judgments involve the toxicity of subspecies of an exotic (but fictitious) Death Bug. The different subspecies vary in concentration of poison from 50 ppm to 60 ppm (a continuous criterion), where concentrations below 55 ppm are harmless but concentrations above 55 ppm are lethal (a binary criterion, harmless vs. dangerous). The toxicity can be in-

ferred from four binary cues of the subspecies (e.g., short or long legs, spots or no spots on the fore-back).

The cues take on values 1 or 0 and the toxicity *c* of a subspecies is a linear, additive function of the cues:

$$c = 50 + 4 \cdot C_1 + 3 \cdot C_2 + 2 \cdot C_3 + 1 \cdot C_4 . \qquad (1)$$

C_1 is the most important cue with *coefficient* 4 (i.e., a relative weight .4), C_2 is the second to most important with coefficient 3, and so forth. The binary criterion *b* is formed from the continuous criterion by assigning *c*<55 *b*=0 (harmless), *c*>55 *b*=1 (dangerous), and *c*=55 randomly as *b*=1 or *b*=0. A subspecies with feature vector (0, 0, 0, 0) thus has poison concentration 50 ppm and is harmless; a subspecies with feature vector (1, 1, 1, 1) has 60 ppm and is dangerous. The continuous and the binary criteria for all the 16 subspecies (i.e., possible cue configurations) are summarized in Table 1.

In *training*, the participants encounter 11 subspecies and make either *binary judgments* about the toxicity of each subspecies (i.e., "harmless" or "dangerous") or *continuous judgments* about their toxicity (e.g., "The amount of poison is 57 ppm"). As indicated in the two right-most columns of Table 1, five subspecies are omitted in training. (Sets A and B, respectively, denote two different training sets where three omitted subspecies are counter-balanced.) In a test phase, the participants make the same judgments as in the training phase, but for all 16 subspecies and without feedback.

Table 1: Structure of the judgment task. The outbalanced constrained training sets are denoted A and B.

Exemplar	Cues				Criteria		Set	
#	C_1	C_2	C_3	C_4	Cont.	Bin.	A	B
1	1	1	1	1	60	1	E	E
2	1	1	1	0	59	1	T	T
3	1	1	0	1	58	1	T	T
4	1	1	0	0	57	1	O	N
5	1	0	1	1	57	1	N	O
6	1	0	1	0	56	1	N	O
7	1	0	0	1	55	*p*=.5	N	O
8	1	0	0	0	54	0	T	T
9	0	1	1	1	56	1	O	N
10	0	1	1	0	55	*p*=.5	O	N
11	0	1	0	1	54	0	T	T
12	0	1	0	0	53	0	T	T
13	0	0	1	1	53	0	T	T
14	0	0	1	0	52	0	T	T
15	0	0	0	1	51	0	T	T
16	0	0	0	0	50	0	E	E

Note: E = Extrapolation exemplar, T = training exemplar, O = Old comparison exemplar presented in training, matched on the criterion to one of the new exemplars, N = New comparison exemplar presented the first time at test, *p*=.5 assigns binary criterion 1 to the exemplar with probability .5.

A criticism of previous studies that support exemplar models is that often the artificial categories used essentially contain no structure at all. There is thus, in a sense, no other way to solve the task than to memorize the exemplars (Smith & Minda, 2000). Our task is neutral in this respect because it allows perfect performance in training both by exemplar memory and by induction of the task structure (i.e., by inducing Eq. 1).

Cognitive Models

The *cue-abstraction model* assumes that participants abstract explicit cue-criterion relations in training which are mentally integrated at the time of judgment. When presented with a probe the participants retrieve rules connecting cues to the criterion from memory (e.g., "Green back goes with being poisonous"). The rules specify the sign of the contingency and the importance of the cue with a cue weight. For example, after training the rule for cue C_1 may specify that C_1=1 goes with a large increase in the toxicity of a subspecies.

With a continuous criterion, cue abstraction suggests that the participants compute an estimate of the continuous criterion *c*. For each cue, the appropriate rule is retrieved and the estimate of *c* is adjusted according to the cue weight ω_i (*i*=1…4). The final estimate $\hat{c}_R$ of *c* is a linear additive function of the cue values C_i,

$$\hat{c}_R = k + \sum_{i=1}^{4} \omega_i \cdot C_i , \qquad (2)$$

where k = 50 + .5(10-$\sum\omega_i$). If ω_1 =4, ω_2=3, ω_3=2, and ω_4=1, Eq's 1 and 2 are identical and the model produces perfect judgments. The intercept *k* constrains the function relating judgments to criteria to be regressive around the midpoint (55) of the interval [50, 60] specified by the task instructions[1]. This formulation essentially provides a cognitive interpretation of the linear additive model known to provide a good account of multiple-cue judgment data (Brehmer, 1994). Predictions by the cue-abstraction model in a continuous task are illustrated in Figure 2A.

The binary judgment involves classification of subspecies into two categories based on their continuous criterion. One way to obtain such judgments from Eq. 2 is by assigning all subspecies with $\hat{c}_R$ <.5 as harmless and all subspecies with $\hat{c}_R$ >.5 as dangerous. Whenever the estimates are correct ($\hat{c}_R$ =*c*) this implies a relation between classification proportions $p_R(b=1)$ and the

[1] The constrained formulation captures the regression effect within the interval [50, 60] that is introduced by a random error in the cue weights or the process of cue abstraction. For example, for the extreme subspecies, (0, 0, 0, 0: *c*=50) and (1, 1, 1, 1: *c*=60), random error may produce judgments that deviate from 50 and 60, respectively. However, for exemplar (0, 0, 0, 0: *c*=50) we expect the errors to more often produce a judgment above than below 50. For exemplar (1, 1, 1, 1: *c*=60) we expect the errors to more often produce a judgment that is below than above 60. Second: it holds to a good approximation in the data reported below. Third, it provides a four-parameter implementation that is more easily compared to the four-parameter exemplar model described below in terms of the number of free parameters.

criterion *c* that is a step function. Taking into account that the process is likely to involve error in cue abstraction and decision making, we allow for a sigmoid function in the form of a logistic function (see Figure 1A):

$$\hat{p}_R(b=1) = \frac{e^{k+\sum W_i \cdot C_i}}{1+e^{k+\sum W_i \cdot C_i}}. \qquad (3)$$

where W_i are the cue weights in a logistic regression and k = -.5 $\sum W_i$. The intercept k implies a crossover from binary judgment 0 to b 1 at toxicity 55, as implied by the instructions. When the cue-abstraction model is fitted to binary judgments below, we rely on Eq. 3.

Exemplar models suggest that the participants make judgments by retrieving similar exemplars (subspecies) from long-term memory. The *context model* of perceptual classification (Medin & Schaffer, 1978) suggests that the probability $\hat{p}_E(b=1)$ of categorization as dangerous equals the ratio between the summed similarity of the judgment probe to the dangerous exemplars and the summed similarity to all exemplars:

$$\hat{p}_E(b=1) = \frac{\sum_{j=1}^{J} S(p,x_j)\cdot b_j}{\sum_{j=1}^{J} S(p,x_j)}, \qquad (4)$$

where p is the probe to be judged, x_j is stored exemplar j (j=1…J), $S(p,x_j)$ is the similarity between the probe p and exemplar x_j, and b_j is the binary criterion stored with exemplar j (b_j=1 for dangerous, b_j=0 for harmless). J depends on the size of training set of exemplars.

The similarity between probe p and exemplar x_j is computed by the multiplicative similarity rule of the context model (Medin & Schaffer, 1978):

$$S(p,x_j) = \prod_{i=1}^{4} d_i, \qquad (5)$$

where d_i is an index that takes value 1 if the cue values on cue dimension i coincide (i.e., both are 0 or both are 1), and s_i if they deviate (i.e., one is 0, the other is 1). s_i are four parameters in the interval [0, 1] that capture the impact of deviating cues (features) on the overall perceived similarity $S(p,x_j)$. s_i close to 1 implies that a deviating feature on this cue dimension has no impact and is considered irrelevant. s_i close to 0 means that the overall similarity $S(p,x_j)$ is close to 0 if this feature is deviating, assigning crucial importance to the feature. The parameters s_i capture the similarity relations between stimuli and the attention paid to each cue dimension, where a lower s_i signifies higher attention.

The context model was developed for classification. To generate predictions also for judgments of a continuous criterion we relax the model by allowing the outcome index b_j to be a continuous value. The estimate $\hat{c}_E$ of c is then a weighted average of the criteria c_j stored for the exemplars, with similarity $S(p,x_j)$ as the weights (see e.g., DeLosh, Busemeyer, & McDaniel, 1997; Juslin & Persson, 2000; Smith & Zarate, 1992).

Predictions

The predictions are summarized in Figures 2 (binary criterion) and 3 (continuous criterion). In both tasks, the models produce similar predictions when all exemplars are presented both at training and test (the upper panels). Both models thus provide accurate representations of the environment, albeit by different means. Figures 3A and 3B illustrate that the good fit of a linear additive model need not be informative in regard to whether cues are really mentally integrated according to a linear model: predictions by an exemplar model are identical. When the extreme exemplars (c= 50 & 60) and three intermediate exemplars (c=.55, 56, & 57) are withheld in training, the models produce distinct predictions.

As illustrated in the lower panels of Figures 2 and 3, the cue abstraction model allows accurate extrapolation beyond the distribution of criteria in the training set [51, 59]. Whenever the correct signs of the cue weights are identified, the most extreme judgments are made for exemplars 1 (c=60) and 16 (c=50). The exemplar model that computes a weighted average of the criteria observed in training can never produce a judgment outside the observed range (Delosh et al., 1997). The most extreme judgments are made for criteria c=51 and 59.

With the cue abstraction model there should be no systematic difference between judgments for the "New" and "Old" exemplars with c=55, 56, and 57: the process is essentially the same in both cases. However, with the exemplar model there is more accurate judgments for Old exemplars: these judgments benefit from retrieval of identical exemplars with the correct criterion.

One way to predict the relative importance of mental cue abstraction and exemplar memory in the binary and the continuous tasks is by computational considerations (see Juslin et al., 2002). For judgments of a continuous variable—assuming a linear additive model, as people tend to (e.g., Brehmer, 1994)—observation of five exemplars with their criteria is, in principle, sufficient to identify the structure of the task. This system of five linear equations has the unique solution provided by Eq. 1. Given a psychological bias towards linear additive models, the task thus has a well-defined rule-based solution that can be induced from a small number of observations. Binary judgment affords no unique solution, even if the correct function form is assumed and all 16 exemplars are considered. Given the difficulty of inducing a rule-based solution, the participants may have little alternative but to rely on exemplar memory (see Smith & Minda, 2000, for similar arguments).

Note the alternative hypothesis suggested by a *single-systems account* (Nosofsky & Johansen, 2000): that the participants rely on exemplar memory in both tasks. On computational grounds there seems to be no reason why exemplar memory should not be equally applied in both tasks. Both tasks allow a linear combination of criteria stored with exemplars. The hypothesis proposed here is based on a *dual-process account*. Because rule-based knowledge affords better communication and system-

atic elaboration, we expect explicit rule-based processes to be applied when the task structure and the feedback allow participants to induce the task structure, whereas exemplar memory provides a general and flexible back-up system when the task structure or feedback is poor.

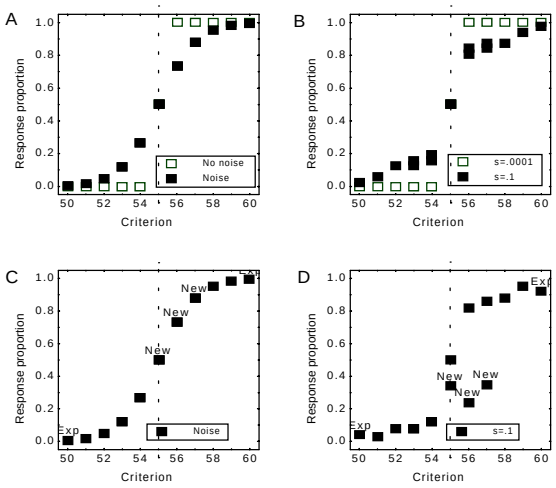


Figure 2: Predictions for the binary task. Panel A: Cue abstraction models with no noise and noise for the complete training set. Panel B: Exemplar model with all similarity parameter *s* equal to .0001 and .1 for the complete set. Panel C: Cue abstraction model with noise for the constrained set. Panel D: Exemplar model with similarity parameter *s*=.1 for the constrained set.

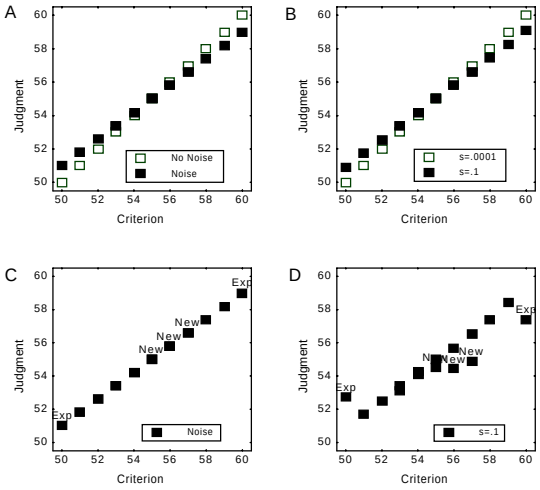


Figure 3: Predictions for the continuous task. Panel A: Cue abstraction models with no noise and noise for the complete training set. Panel B: Exemplar model with all similarity parameter *s* equal to .0001 and .1 for the complete set. Panel C: Cue abstraction model with noise for the constrained set. Panel D: Exemplar model with similarity parameter *s*=.1 for the constrained set.

Method

Participants

Sixty-four persons participated in the experiment (35 women and 29 men, with an average age of 23.5 years). All participants were undergraduate students at Umeå University and rewarded with 70 SEK (app. 7 US $) for their participation in the experiment.

Materials and Procedure

The written instructions informed the participants that there were different subspecies of a Death bug. The subspecies differed in toxicity between 50 and 60 ppm, toxicity below 55 is harmless and toxicity above 55 is dangerous. In the binary task condition, the instruction asked the participants to categorize the subspecies into dangerous and harmless. The training phase provided trial-by-trial outcome feedback about the binary criterion ("This bug is dangerous"). In the continuous task condition, the task was to directly estimate the *toxicity* of the subspecies as a number between 50 and 60. In training, the participants received feedback about the continuous criterion ("This bug has toxicity 57 ppm"). The question on the computer screen was "Is this subspecies harmless or dangerous? (binary task)" or "What is the toxicity of this subspecies? (continuous task)".

The subspecies varied in terms of four binary cues; leg length (short or long), nose length (short or long), spots or no spots on the fore back, and two patterns on the buttock. The cues had the weights 4, 3, 2, and 1 (Eq. 1). The weights determine the portion of toxicity that each cues adds to the total amount. In the analogue stimulus condition, the participants were presented with pictures of the subspecies, in the propositional stimulus condition they were presented with four propositions that provided information about the cue values.

The training phase consisted of 220 trials, where the 11 training exemplars in Table 1 were presented 20 times each. The remaining five exemplars were omitted in the training phase. Two different training sets were used (Sets A and B in Table 1). In Set A, Exemplars 5, 6, and 7 were omitted; in Set B, Exemplars 4, 9, and 10. The exemplars in the two training sets were pair-wise equal in toxicity and the omission of these exemplars was thus counterbalanced across the training sets.

In the test phase, all participants judged all 16 exemplars, twice with an analogue stimulus format and twice with a propositional stimulus format. The stimulus formats were presented in two 2x16 blocks, the order of which was counterbalanced across the participants. No feedback was provided in the test phase. Half of the participants were trained with analogue stimuli and the other half with propositional stimuli, whereas all participants were tested with both presentation formats.

Results

Results and model fits were collapsed over the analogue and propositional conditions, as the aim of this paper is to investigate the relative importance of mental cue abstraction and exemplar memory in binary and continuous tasks. Figure 4 presents model fits (r^2 & *Root Means Square Deviation*) and mean judgments.

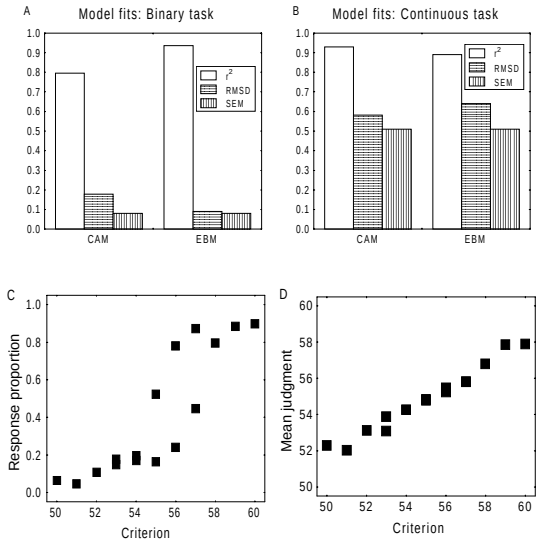


Figure 4: Panel A: Model fits for the binary task. Panel B: Model fits for the continuous task. Panel C: Response proportions in the binary task. Panel D: Mean judgments in the continuous task.

Inspection of Figure 4C suggests exemplar effects in the binary judgment task (notice the difference between new and old exemplars with criterion 55, 56, and 57). The mean difference in proportion of dangerous decisions between old and new intermediate exemplars was -.33 (95% CI: -.45 — -.21). Both the cue abstraction model and the exemplar model were fitted to data. The four parameters in each model were estimated with a Quasi-Newton procedure that minimized the sum of squared deviations between data and model predictions for the last 110 trials in the *training block*. These parameters were used to predict data in the *test phase* (i.e., all free parameters were determined by training data and thus produce cross-validation for training exemplars and genuine predictions for new exemplars).

The exemplar model is clearly superior in the binary judgment task. The model fit indexes in Figure 4A for the binary task, *RMSD* and r^2, suggests predominant use of exemplar processes with model fits almost identical to the mean standard error in data (*SEM*) and r^2 above .90. In the continuous task, the model fits are more ambiguous, although there is a slight advantage for the cue abstraction model (Figure 4B). However, there are nonetheless signs of exemplar effects (e.g., the judgments for the extreme exemplars are at the level of, or below, the judgments for the second-to-most extreme). A comparison between the two tasks revealed that the percentage of participants showing exemplar effects dropped from 81% in the binary task to 63% in the continuous task (p = .06).

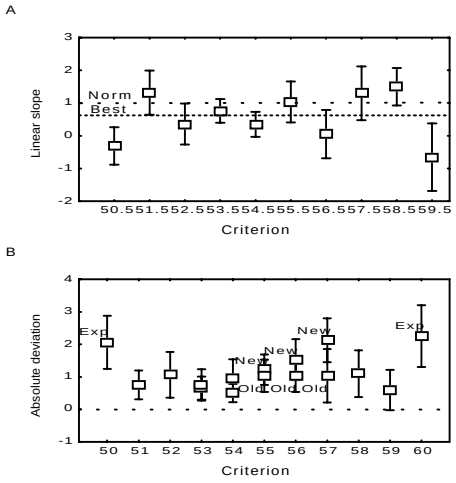


Figure 5: Exemplar effect in the data for continuous judgments with the same training and test stimuli. Panel A: Mean difference (slope) with 95% CI between each successive data point. Panel B: Mean absolute difference from the correct value for each data point.

The signs of exemplar effects are evident in Figure 5 presenting data for continuous judgments with the same training and test stimuli. Panel 5A plots the difference between each successive data point (slope) in a graph like Figure 4D. This slope is 1 for perfect judgments. Panel 5A also provides the slope of the best fitting linear regression of the mean judgments on the criterion. Panel 5B presents the mean absolute error of judgment for each criterion. It is clear that the slopes turn negative for the extreme criteria (inability to extrapolate) with more error in the judgments for new exemplars. There is poorer ability to extrapolate the continuous judgments when training and test stimuli were in the same format ($F(1, 30) = 4.62$, $p = .04$), thus suggesting more exemplar retrieval.

The ambiguous results for the continuous judgments suggest that the group-level data may actually be a mix of the two processes. Investigation of individual participants indeed revealed individual differences. Some participants relied on cue-abstraction, others on exemplar retrieval (Figure 6). Somewhat arbitrarily, but as bench-mark, we deemed best-fitting models accounting for more than 70% of the variance in individual data as producing acceptable fit. On this criterion, 11 participants (34%) were best accounted for by the exemplar model, 13 (41%) were best accounted for by the cue abstraction model, whereas 8 (25%) were not accounted for (r^2<.7 for both models). In sum: although there are exemplar effects also with a continuous criterion, there

is an increased prevalence of cue abstraction with some participants clearly relying on cue abstraction.

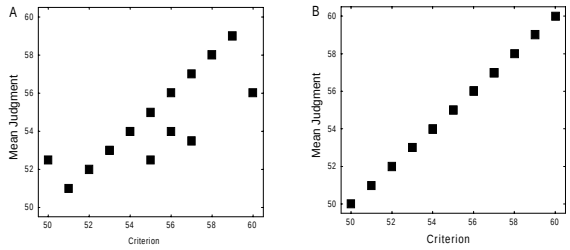


Figure 6: Individual-participant data: Panel A exemplifies a participant guided by exemplar retrieval and panel B a participant guided by cue abstraction.

Discussion

The question addressed in this article is why the theoretical conclusions from categorization and multiple cue judgment research are different, considering that the task structure is so similar (Figure 1). Perhaps, the most salient difference between the paradigms is that categorization often involves a binary whereas multiple-cue judgment often involves a continuous criterion.

The results suggest that the differential emphasis in the conclusions is not accidental to different research traditions, with more cognitive modeling in categorization research and more statistical modeling in multiple cue judgment research. Changing the criterion from binary to continuous thus creates a shift from exemplar memory to a mix of exemplar- and rule-based processing that involves cue abstraction in training and cue-integration at the time of judgment. In the continuous judgment condition just as many individual participants extrapolated appropriately and relied on cue abstraction as on exemplar memory. Figure 6 highlights the individual differences in preferred representational mode (see Shanks & Darby, 1998, for similar results). These results raise the question of the appropriateness of the routine procedure of applying quantitative models to group-level data. The exemplar retrieval with continuous judgments moreover seems to increase when training and test conditions coincide.

There is no reason why exemplar memory should not be used in both tasks (as it indeed was by some participants). Exemplar retrieval is an equally efficient way to solve both tasks. However, it seems that as soon as the feedback is informative enough, people eagerly induce explicit rule-based representations, corresponding to the "rule-bias" suggested by Ashby et al. (1998). This suggests that people change between qualitatively distinct representation levels depending on the task properties (Ashby et al., 1998; Jones et al, 2000; Juslin et al., 2002). Jones et al. showed that people spontaneously tend to integrate cues in a task like the one used here, either explicitly by cue abstraction or implicitly by exemplar retrieval. A principled understanding of the interplay between – and properties of – these distinct levels of representation in human judgment and categorization should be a prime goal of cognitive science.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442-481.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87,* 137-154.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968-986.

Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review, 86,* 465-485.

Jones, S., Juslin, P., Olsson, H., & Winman, A. (2000). Algorithm, heuristic or exemplar: Processes and representation in multiple-cue judgment. In L. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 244-249). Hillsdale, NJ: Erlbaum.

Juslin, P., Olsson, H., & Olsson, A-C. (2002). *Abstract and concrete knowledge in categorization and multiple-cue judgment.* Manuscript submitted for publication. Department. of Psychology, Umeå University, Umeå, Sweden.

Juslin, P., & Persson, M. (2000). *PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge*. Manuscript submitted for publication. Department. of Psychology, Umeå University, Umeå, Sweden.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning, *Psychological Review, 85,* 207-238.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7,* 375-402.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in categorization. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 405-415.

Smith, J. D, & Minda, J. P (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning , Memory and Cognition, 26, 3-27.*

Smith, E. R., & Zarate, M. A. (1992). Exemplar model of social judgment. *Psychological Review, 99,* 3-21.

Predicting Noun and Verb Latencies: Influential Variables and Task Effects

Natalie Kacinik (kacinn01@student.ucr.edu)
Christine Chiarello (christine.chiarello@ucr.edu)
University of California, Riverside
Life Sciences Psychology Building
Riverside, CA 92521-0426, U.S.A

Abstract

Natural language comprehension involves processing a multitude of words that vary along many dimensions, some of which may reflect statistical regularities in language. These variables may differ in their relative importance across various types of words and tasks. This study used a multiple regression approach to investigate potentially important predictors of noun-verb processing across naming, grammatical categorization, and sentence completion tasks. Although there were some indications of different predictors for nouns vs verb processing, the strongest predictors of response latencies were primarily determined by the types of processing most important for a given task. One variable of particular interest was the newly created Noun-Verb Distributional Difference (NVDD) metric developed by Chiarello et al. (1999). NVDD values reflect statistical regularities in language regarding the typicality of the contexts in which nouns and verbs tend to occur. The results suggest that although noun-verb typicality, as assessed via the NVDD, is a valid measure of regularities in noun-verb contexts within a linguistic corpus, individuals may not be very sensitive to this dimension in standard psycholinguistic processing tasks.

Introduction

Single word recognition is a central component of language processing. The typical approach has been to use a naming or lexical decision (LD) task and a factorial design to investigate the processing effect of one or more variables such as familiarity or imageability while holding other potentially confounding variables constant. In addition, most single word recognition research has tended to use words of different parts of speech without considering grammatical class (e.g., nouns vs. verbs), or has focused on concrete, imageable nouns. Natural language comprehension, however, involves processing a multitude of words varying along many dimensions. These dimensions may be relatively more or less important for various word types, and their relative importance is likely to vary across different forms of language processing (e.g., word pronunciation vs grammatical identification vs sentence integration).

With a few exceptions (e.g., Balota & Chumbley, 1984; Balota, Cortese, & Pilotti, 1999), there have been few attempts to investigate the relative importance of various orthographic and semantic dimensions for responding to words across tasks using multiple regression procedures. This approach provides the opportunity to study many variables simultaneously, to determine which lexical dimensions account for the greatest amount of variance in reaction time (RT) and accuracy for a particular task, and to assess whether the variance accounted for is unique, or is shared by other variables.

Such a regression approach was used in the present study to investigate the relative importance of different lexical dimensions across three language tasks. To our knowledge no prior regression study has examined whether various predictor variables are equally applicable to words of different grammatical class. This is an important issue because neuropsychological research has shown that nouns and verbs appear to be processed differently in the brain (e.g., Daniele et al., 1994; Koenig & Lehmann, 1996; Sereno, 1999). It is unclear whether these differences are due to neurally separate noun and verb processing systems, or whether these processing differences are mainly due to different semantic dimensions that covary with word class. Investigating several potentially relevant dimensions using a regression approach may be informative regarding these processing differences between nouns and verbs.

One possible reason word recognition research has generally been limited to concrete, imageable nouns, is the lack of word norming corpora available for other word types. A recent study by Chiarello, Shears, and Lund (1999), however, provides imageability ratings, frequency values from the Usenet text corpus of the Lund and Burgess (1996) Hyperspace Analog to Language (HAL) model, and a new measure of noun-verb distributional typicality (the Noun-Verb Distributional Difference, NVDD, metric), for a set of 1197 words: 555 “pure” nouns, 427 “pure” verbs, and 215 words “balanced” for noun-verb usage, as classified by the Francis and Kucera (FK, 1982) norms.

Noun-Verb Distributional Typicality

The new measure of noun-verb usage developed by Chiarello et al. (1999) uses context vectors from the Lund and Burgess (1996) HAL model, where words occurring in similar phrasal and sentential contexts are nearby in high dimensional context space. Context distances were computed between each word and each of the 555 “pure” nouns (according to Francis & Kucera, 1982) and averaged to get a mean noun context distance score. Mean verb distance scores were similarly obtained by computing and

averaging the context distances between each word and the 427 "pure" verbs (Francis & Kucera, 1982). The mean verb context distance was then subtracted from the mean noun context distance, for each word, resulting in a measure referred to as NVDD (Noun-Verb Distributional Difference), indicating the extent to which the word occurs in contexts that are more typical of nouns or verbs. Chiarello et al. (1999) validated the NVDD measure by demonstrating their strong correlation with Francis & Kucera (1982) estimates of noun-verb usage, as well as by examining actual part of speech occurrences for a subset of words in sentence contexts from the Usenet corpus.

The NVDD measure is purely computational, however, reflecting statistical regularities of noun-verb usage in the language corpus. Given the many compelling demonstrations of how information about statistical regularities in a learner's environment can be extracted to result in the learning of various language phenomena (e.g., Elman, 2001; Saffran, 2001a,b), it was of interest to determine whether the NVDD is psychologically relevant for language processing. We therefore examined whether individuals would demonstrate a sensitivity to the contextual typicality of nouns and verbs in their performance on psycholinguistic tasks. The present paper further investigates the psychological validity of NVDD across three different linguistic tasks, whose findings can be compared with our previous lexical decision results.

Prior Lexical Decision Results

Kacinik, Shears, and Chiarello (2000) reported regression results investigating the influence of imageability (ease with which a word arouses a mental image), word length, experiential familiarity, NVDD, and 2 measures of frequency (FK, 1982, and Usenet corpus) on noun and verb lexical decision (LD) response times (RTs).

The results indicated that noun-verb typicality (NVDD) accounted for a significant portion of verb RT on its own, but it did not account for any unique variance with the other variables in the model. For nouns, the opposite occurred, such that NVDD did account for a small, but significant portion of unique variance. These LD results only partially support the psychological relevance of the recently developed NVDD metric, because it was not a very important predictor. LD, however, requires discriminating words from nonwords, and as nouns and verbs are both words, noun-verb typicality may not really be a relevant dimension for making this discrimination. Experiments 2 and 3 of the current study investigated grammatical categorization and sentence completion tasks, both of which should involve language processes for which noun-verb typicality could be more relevant.

The LD results also showed different variables to be more or less important for predicting noun vs. verb RT. Specifically, imageability appeared to be a more important predictor for verb responses, whereas frequency appeared to be more important for nouns. Most importantly, however, the results indicated that although frequency, imageability, and NVDD could each account for a portion of the RT variance individually, they failed to account for much unique variance.

The biggest predictor of RT was familiarity, individually accounting for 50% or 60% of the variance, for nouns and verbs, respectively. Moreover, when contributions of the other variables were partialled out, around half of the RT variance accounted for by familiarity appeared to be unique (33% for nouns and 30% for verbs). Familiarity's importance in predicting LD RTs is not surprising (e.g., Balota et al., 1999; Gernsbacher, 1984), suggesting that familiarity is probably the most important dimension for discriminating between words and nonwords. More surprising, however, was that about half of the variance accounted for by familiarity was unique.

To continue examining predictors of noun-verb processing, and the psychological relevance of the NVDD metric across various linguistic processes, Experiment 1 examined word naming, Experiment 2 involved noun-verb decision, and Experiment 3 investigated sentence completion.

Experiment 1

This experiment investigated which variables would be the most important predictors for noun and verb naming latencies. Word naming entails activating phonological representations to produce a vocal pronunciation response, and is not thought to require much semantic processing (Balota et al., 1999). This is in contrast to lexical decision, as well as the grammatical categorization and sentence completion tasks examined in Experiments 2 and 3, which mainly involve activating semantics to make a decision and subsequent key press response. Familiarity, frequency, and length were expected to be most influential because of their likely importance in the initial recognition processes involved in activating phonological representations (Balota et al., 1999). In addition, because nouns and verbs primarily differ in meaning and grammatical class, we did not expect differences for the relative importance of predictor variables in this pronunciation task.

Method

Mean latencies from Spieler and Balota's (1997) young adult naming corpus were obtained for 251 nouns and 131 verbs found in our database of 1197 words. Spieler and Balota report that these mean RTs were obtained from 31 Washington University undergraduates (mean age = 22.6), who named a total of 2870 monosyllabic words.

The word length, NVDD, FK and Usenet frequencies, imageability, and familiarity values for each of these words were taken from the Chiarello et al. (1999) database. These 6 predictor variables were combined with the Spieler and Balota naming latencies as the dependent variable.

Results and Discussion

The multiple regression results for noun and verb naming latencies are shown in Table 1. When all 6 variables were in

Table 1: Noun and verb regression analyses for the naming task, Experiment 1.

NOUNS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p≤
NVDD	.00	.00	-.00	-0.08	ns
Image	.01	.00	-.01	-0.11	ns
Length	**.11**	**.12**	**.35**	**6.08**	**.0001**
FK Freq	**.06**	.01	-.24	-1.69	ns
Usenet Freq	**.05**	.00	.09	0.64	ns
Fam	**.05**	**.03**	**-.19**	**-2.78**	**.01**
VERBS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p≤
NVDD	.01	.02	.17	1.67	ns
Image	.00	.00	-.01	-0.15	ns
Length	**.07**	**.05**	**.23**	**2.52**	**.05**
FK Freq	.02	.00	.04	0.33	ns
Usenet Freq	.03	.00	-.10	-0.73	ns
Fam	**.05**	**.03**	**-.25**	**-2.02**	**.05**

Note: R^2 = .21, $F(6, 245) = 10.57$, $p<.0001$, for nouns, and R^2 = .12, $F(6, 121) = 2.87$, $p<.05$, for verbs

the regression, they significantly accounted for 21% of the RT variance for nouns, and only 12% of variance for verbs. It thus appears that the lexical dimensions typically thought to influence word naming are relatively more important for nouns than verbs. Further examination of the results, however, suggests that although both frequency measures appear to be additionally influential for nouns, they do not account for any unique variance. The variance they account for appears to be subsumed by length and familiarity. Indeed, the main conclusion from these results is that similar dimensions (length and familiarity) were most important predictors for both noun and verb RTs. These findings are in contrast to the previous LD results where, after familiarity, there was some indication that different predictors were important for noun vs verb processing. This suggests that noun-verb processing differences are due to semantics and/or result from postlexical processing.

NVDD was not found to be a significant predictor for noun latencies, and was only marginally significant for verbs, accounting for 2% unique variance. Thus, both the LD and naming results provide minimal support that individuals are sensitive to regularities in the contexts in which nouns and verbs occur. Another possibility is that neither of these tasks involves explicitly activating word class information. Perhaps a task that does require explicit activation of noun or verb meaning and/or grammatical class information, will show greater effects of noun-verb typicality.

Experiment 2

This experiment investigated the relevance of the typicality of contexts in which nouns and verbs tend to occur (NVDD) for deciding whether a word is a noun or verb. In contrast to LD and naming, this task requires the explicit activation of grammatical class information in order to make the noun-verb decision.

Method

Participants

Forty native English speaking University of California, Riverside undergraduates (20 males) participated in the experiment in exchange for course credit or pay ($6.00).

Stimuli

The same set of 152 nouns and 137 verbs, varying in NVDD, from Kacinik et al. (2000) were employed here.

Procedure

Each trial began with the presentation of a 400 ms fixation point, followed by 100 ms blank screen, which was followed by presentation of the target word. Participants were required to decide whether each item was a noun or a verb by making a button press response as quickly as possible. Targets remained on the screen until they responded, and the inter-trial interval was 1500 ms. Participants were told that nouns were words naming a quality, person, place, or thing, while verbs are words that express an action or the occurrence of an event, and given some examples of each. Fifteen practice trials preceded the experiment.

Results and Discussion

The noun and verb multiple regression analyses are presented in Table 2. When all the variables were included in the regression model, it was better at predicting noun-verb decision than naming latency, such that 44% of the RT

Table 2: Noun and verb regression analyses for the noun-verb decision task, Experiment 2.

NOUNS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p≤
NVDD	**.04**	**.03**	**.19**	**2.87**	**.005**
Image	**.37**	**.28**	**-.57**	**-8.50**	**.0001**
Length	.02	.01	.12	1.82	ns
FK Freq	.01	.00	-.01	-0.06	ns
Usenet Freq	.01	.00	-.10	-0.89	ns
Fam	**.05**	.01	-.11	-1.43	ns
VERBS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p≤
NVDD	**.10**	**.07**	**-.28**	**-3.79**	**.001**
Image	**.24**	**.17**	**-.44**	**-6.08**	**.0001**
Length	.00	.00	.01	0.09	ns
FK Freq	.01	.01	.11	1.16	ns
Usenet Freq	.02	.01	-.12	-1.30	ns
Fam	**.18**	**.03**	**-.20**	**-2.40**	**.05**

Note: R^2 = .44, $F(6, 144) = 18.57$, $p<.0001$, for nouns, and R^2 = .40, $F(6, 123) = 14.39$, $p<.0001$, for verbs

variance for nouns, and 40% of the RT variance for verbs, was accounted for.

For nouns, imageability followed by NVDD was the most significant predictor of noun-verb decision times. Although familiarity individually accounted for 5% of the noun RT variance, it was not found to contribute a significant amount of unique variance. Imageability and NVDD were also the most important predictors of unique variance for verbs. Familiarity was also a strong predictor of verb latency on its own, but in contrast to the noun results, it also accounted for a small significant unique amount of variance.

Imageability therefore appears to be the most influential dimension for deciding whether a word is a noun or a verb. Furthermore, much of imageability's contribution and the variance it accounts for seems to be unique. There is also some indication that imageability is more important for nouns than verbs in the noun-verb decision task. This is contrary to the LD results, where imageability appeared somewhat more important for verbs. A possible explanation for these results is that since nouns are generally more imageable than verbs, a highly imageable concrete word encountered in the context of the noun-verb decision task must be a noun and could be responded to very rapidly (Chiarello et al., 1999). A negative correlation should thus be expected between imageability and decision latencies for verbs, because low imageability words are more likely to be verbs (Chiarello et al., 1999), and should be responded to quickly. The obtained correlation, however, was positive, suggesting that imageability does not facilitate the noun-verb decision per se. Instead, we suggest it facilitates earlier processes such as the speed of word meaning activation, enabling subsequent noun-verb decisions to be made more rapidly.

As expected, noun-verb typicality (NVDD) was significantly correlated with both noun and verb decision latencies (r = .21, $p<.01$, and r = -.32, $p<.0001$, respectively). Given that noun-verb decision does explicitly involve processing part of speech information, it is surprising that NVDD was not more important and only accounted for 3% and 7% unique variance for nouns and verbs, respectively. Recall, however, that the NVDD is a measure of the typicality of contexts in which nouns and verbs occur. All tasks investigated thus far involved single word processing, and may not reflect the influence of the fundamentally contextual nature of the NVDD metric. It is possible that NVDD may be most relevant for processing words in sentence contexts.

Experiment 3

This experiment investigated whether noun-verb contextual typicality would influence response speed for deciding whether a word could be sensibly integrated into an incomplete sentence. If sentence context is assumed to constrain possible completions, faster latencies would be expected for words that are highly typical nouns or verbs than for words less typical of their grammatical class.

Method

Participants

Forty native English speaking University of California, Riverside undergraduates (20 males) participated in the experiment in exchange for course credit or pay ($6.00).

Stimuli

Incomplete sentence frames, which could be sensibly completed by either a high or low NVDD noun or verb, were created. For example, *punish* and *smack* are high and low NVDD verb completions for *"the father wanted to*

_____ *his son"*, respectively, and *tavern* and *pub* are high and low typicality noun completions for *"They walked into the _____"*. We created such incomplete sentences for 80 verbs (40 high and 40 low) and 88 nouns (44 high and low) from the set of 152 nouns and 137 verbs used by Kacinik et al. (2000). Sentences were normed and balanced for sensibility and cloze probability. Nonsensical completions were created by re-pairing sentence frames and target words *(e.g., She had to punish the text).*

Procedure

Each trial began with the presentation of a 500 ms fixation point, immediately followed by the appearance of the incomplete sentence. After 1200 ms, the target word appeared above the sentence, allowing participants enough time to read the sentence prior to the target's appearance. Both the sentence and target word remained on the screen until they responded. Participants had to decide whether the target word was a sensible completion to the sentence and responded by pressing a key as quickly and accurately as possible. The inter-trial interval was 1000 ms. Twenty-four practice trials were completed prior to the experiment.

Results and Discussion

Table 3 presents results from regression analyses for the related condition (i.e., when the noun or verb was a good completion to the sentence). Contrary to our predictions, noun-verb contextual typicality (NVDD) was not found to be relevant for sentence integration, a task where it was expected to strongly influence processing. Indeed, it is surprising that with the exception of Usenet frequency for nouns ($r = -.22$, $p<.01$) and imageability for verbs ($r = -.25$, $p<.01$), none of the variables were significantly correlated with decision latencies. Noun RT did not correlate with NVDD ($r = -.18$, ns), imageability ($r = -.20$, ns), length ($r = .07$, ns), FK freq ($r = -.12$, ns), and familiarity ($r = -.11$, ns). Verb RT was also not correlated with NVDD ($r = -.16$, ns), length ($r = .09$, ns), FK freq ($r = -.01$, ns), Usenet freq ($r = -.10$, ns), and familiarity ($r = -.11$, ns). These results suggest that lexical- semantic dimensions identified as being important for single word recognition are much less relevant for integrating words into sentences. Indeed, when all the variables are in the regression model, they only account for a marginally significant (13%) portion of variance for nouns, and a non-significant 12% of the variance for verbs.

One possible explanation is that these are lexical variables, representing characteristics of single words, and this task is primarily measuring sentence integration. The nature of the sentence, therefore, is also an important source of variance for this task. Indeed, it may even be the most important, suggesting that perhaps "sentence-level variables" such as sentence length, imageability, or meaningfulness, would be better predictors of decision latency. Another possibility is that these lexical variables are still important for initial word recognition and meaning activation, but that their influence dissipates once a word has been recognized, such that they are not involved in higher-level sentence integration processes. This would predict these lexical dimensions, and possibly NVDD, to significantly contribute to initial word recognition processes that happen in on-line sentence comprehension.

General Discussion

The relative influence of variables on processing nouns and verbs in naming, noun-verb decision, and sentence completion tasks was investigated. There were some indications in our previous lexical decision results that, after familiarity, different dimensions might vary in terms of their relative importance for noun versus verb processing.

Table 3: Noun-verb regression analyses for the sentence completion task, Experiment 3.

NOUNS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p<
NVDD	.03	.03	-.18	-1.57	ns
Image	.04	**.05**	**-.25**	**-2.24**	**.05**
Length	.00	.00	.01	-0.08	ns
FK Freq	.01	.00	.03	0.18	ns
Usenet Freq	**.05**	**.04**	**-.28**	**-1.85**	**.10**
Fam	.01	.01	.11	0.81	ns
VERBS	Sole Predictor Variance	Unique Variance (semi-partial r^2)	Beta	t value	t-test sig. of β p<
NVDD	.03	.03	-.17	-1.45	ns
Image	**.06**	**.05**	**-.25**	**-2.12**	**.05**
Length	.01	.00	.01	0.10	ns
FK Freq	.00	.01	.19	1.04	ns
Usenet Freq	.01	.01	-.20	-1.02	ns
Fam	.04	.00	-.05	-0.38	ns

Note: $R^2 = .13$, $F(6, 81) = 1.96$, $p<.10$, for nouns, and $R^2 = .12$, $F(6, 73) = 1.66$, ns, for verbs

Although the present study also found some evidence that different variables were relatively more important for the processing of nouns vs verbs across tasks, these differences were fairly subtle. They generally involved differences in the strengths of relationships between variables and RT, or in the unique amount of variance accounted for by each variable. Thus, the main conclusion from this study should be that the strongest predictors of RT do not depend on word class per se, but are determined primarily by the type of processing necessary for a given task. Specifically, word length and familiarity were most important for naming both nouns and verbs, familiarity is the most important dimension for discriminating words and nonwords (regardless of whether the item is a noun or verb), imageability and noun-verb typicality were most relevant for deciding whether a word is a noun or verb, and imageability was the most relevant lexical variable for integrating words into sentences.

A variable of particular interest in the current study was the typicality of contexts in which nouns and verbs tend to occur, as measured by Chiarello et al.'s (1999) recent NVDD metric. Despite the fact NVDD captures statistical regularities of noun-verb usage based on the typicality of contexts in which they tend to occur (Chiarello et al., 1999), it does not seem very relevant for the processing of nouns and verbs across a variety of tasks. Some results did find noun-verb typicality to correlate with lexical and noun-verb decision latencies. In these cases, however, part of that correlation with RT was also accounted for by other variables, such that the unique portions of variance accounted for by NVDD were rather small. This was true even in tasks where noun-verb typicality was expected to influence processing.

There is no doubt that individuals are sensitive to some statistical regularities present in the language environment, many of which can affect processing (e.g., Saffran, 2001a), The present findings suggest, however, that this may not be true for all such regularities. The typicality of contexts in which nouns and verbs tend to occur, as measured via NVDD, appears to be a valid regularity within a linguistic corpus. Yet individuals may not be very sensitive to this dimension in standard psycholinguistic tasks.

References

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception & Performance, 10, 340-357.

Balota, D.A., Cortese, M.J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In Abstracts of the 40th Annual Meeting of the Psychonomics Society. Los Angeles, CA: Psychonomic Society.

Chiarello, C., Shears, C., & Lund, K. (1999). Imageability and distributional typicality measures of nouns and verbs in contemporary English. Behavior Research Methods, Instruments, & Computers, 31, 603-637.

Daniele, A., Guistolisi, L., Silveri, M.C., Colosimo, C., & Gainotti, G. (1994). Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. Neuropsychologia, 32, 1325-1342.

Elman, J. L. (2001). Connectionism and language acquisition. In M. Tomasello & E. Bates (Eds.), Language development: The essential readings (pp. 295-306). Malden, MA: Blackwell.

Francis, W., & Kucera, H. (1982). Frequency analysis of English usage: Lexicon and grammar. Boston: Houghton Mifflin.

Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. Journal of Experimental Psychology: General, 113, 256-281.

Kacinik, N., Shears, C., & Chiarello, C. (2000). Familiarity for nouns and verbs: Not the same as, and better than, frequency. Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, Philadelphia, PA.

Koenig, T., & Lehmann, D. (1996). Microstates in language-related brain potential maps show noun-verb differences. Brain & Language, 53, 169-182.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrences. Behavior Research Methods, Instruments & Computers, 28, 203-208.

Saffran, J. R. (2001a). The use of predictive dependencies in language learning. Journal of Memory & Language, 44, 493-515.

Saffran, J. R. (2001b). Words in a sea of sounds: The output of infant statistical learning. Cognition, 81, 149-169.

Sereno, J.A. (1999). Hemispheric differences in grammatical class. Brain & Language, 70, 13-28.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. Psychological Science, 8, 411-416.

Acknowledgments

We thank Connie Shears for her help in developing the database used in the current study, and Curt Burgess, Dan Ozer, and Robert Rosenthal, for their useful suggestions. This research was supported by National Science Foundation Grants No. SBR-9729009 and BCS-0079456 to the second author.

Graph Structure Supports Graph Description

Irvin R. Katz (ikatz@ets.org)
Center for New Constructs, Educational Testing Service
Princeton, NJ 08541 USA

Xiaoming Xi (xxm@ucla.edu)
Dept. of Applied Linguistics & TESL, Univ. of California
Los Angeles, CA 90095 USA

Hyun-Joo Kim (hk312@columbia.edu)
Applied Linguistics Program, Columbia University
New York, NY 10027-6696 USA

Peter C-H. Cheng (peter.cheng@nottingham.ac.uk)
School of Psychology, University of Nottingham
University Park, Nottingham, NG7 2RD U.K.

Abstract

This research adapts theories of graph comprehension to investigate the factors affecting how easily a graph can be described. We find that the structure of a graph—the number of *visual chunks* (visually distinct units of information) to be described—influences the communicative quality of elicited descriptions. The work extends our understanding of graph comprehension by investigating the relationship between comprehension and description processes. This research occurs in the context of understanding how to design graphical description tasks for the Test of Spoken English.

Introduction

Graphs are a ubiquitous communication tool. Instructors describe graphs to communicate concepts, perhaps requiring students to uncover a graph's main point. A doctor might describe a graph to a patient to make a point about treatment ("see how your cholesterol level has been decreasing since you began the new diet?"). Yet we know little about the cognitive processes engaged when people describe a graph. Research on graph description can contribute to our understanding of how people integrate visual and verbal information in the performance of everyday tasks. From a practical standpoint, such research can provide guidelines for designing graphs that facilitate description.

Instead, much of the research on graphs has focused on graph comprehension—how we encode and interpret elements of a graph to draw out key pieces of information (Carpenter & Shah, 1998; Lohse, 1993; Pinker, 1990), typically in response to relatively narrow tasks (e.g., "Who had a greater market share in 1983?"). The few studies that investigate spontaneous descriptions of graphs have focused on what is described (e.g., global trends vs. local, piecemeal descriptions [Carswell, 1993; Carswell et al., 1998]; trends vs. comparisons [Zacks & Tversky, 1999]) and the organization of the descriptions (Shah, Hegarty, & Mayer, 1999; see below) rather than on the *communicative quality* of the description. One reason for this oversight might be the lack of a rigorous measure of communicative quality.

In the work presented herein, we apply a theory of graph comprehension to predict the characteristics of graphs that facilitate descriptive communication. To measure the quality of descriptions produced by alternative graphs, we use a theoretically grounded and empirically validated measure of communicative quality: the scoring rubric from the Test of Spoken English (TSE®).

The next section provides some background on the TSE, its scoring rubric, and the real-world problem that motivated this research.

The Test of Spoken English

The real-world problem

The goal of the Test of Spoken English (TSE) is to measure a test-taker's communicative competence in Northern American English. It is taken by approximately 30,000 non-U.S. citizens each year, who are seeking to be teaching assistants or healthcare professionals in the U.S. The test consists of 12 questions that elicit a range of communication functions (e.g., describe, compare, state opinion) through a variety of visual and verbal prompts. The questions are presented visually in a booklet and aurally by a taped interviewer; test-takers' spoken responses are recorded. Responses are scored by trained raters employing a well-defined scoring rubric (see below).

One question (illustrated in Figure 1) prompts for a description of a statistical graph. Test-takers are given one minute to respond. The task mirrors the type of communication using graphs done by teaching assistants and healthcare professionals. None of the other 11 questions presents a data graph.

The graph below shows what people of two age groups value about their work. Describe the information given in the graph.

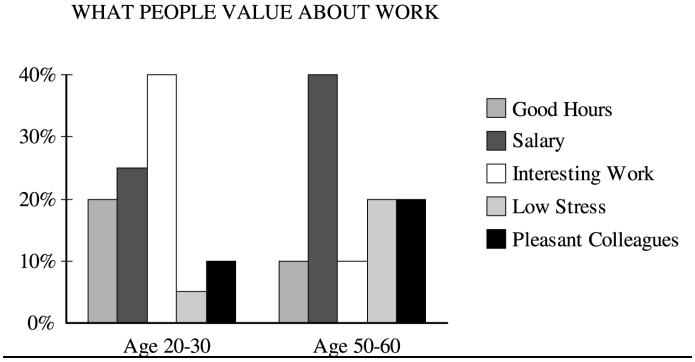


Figure 1: Illustrative graph question [Fewer visual chunks].

This type of graph-description question occasionally poses problems for TSE scoring. According to the raters, certain graphs elicit speech that displays a lower ability in

English than would be expected based on responses to all other test questions. However, many graphs evidenced no such difficulties. Analyses of data from the past two years of TSE administrations confirm that graph description questions are more likely than questions with non-graph prompts to elicit such discrepant performance (Katz, Xi, Kim, & Cheng, 2002).

The issue of what characteristics of a graph lead to descriptions that communicate better is critical to the TSE. If a graph is hard to describe, it might give an unfair advantage to test-takers with better graph-reading skills (i.e., a more sophisticated "graph schema"; Pinker, 1990), who can make sense of poorly constructed graphs. A test-taker's ability to read and interpret graphs should not influence their score on a graph question. Indeed, the accuracy of a person's response to a graph item is not considered in the score, only the degree to which the person evidences certain competencies associated with spoken English.

The challenge is to create graphs that contain enough information so as not to trivialize the description (which would eliminate any differences between test-takers) yet are straightforward to describe, allowing test-takers to show off their communicative skill without other factors getting in the way. Ultimately we seek to develop guidelines for the development of graph questions that validly measure communicative competence.

TSE Scoring Rubric

Responses to TSE prompts are scored according to the published "TSE SCORE BAND DESCRIPTOR CHART" (TOEFL, 2001). This scoring rubric defines four key communicative competencies: discourse, functional, sociolinguistic, and linguistic competence. The chart also specifies the types of response characteristics for these competencies at each of the five possible score levels (20, 30, 40, 50, and 60). Although these several competencies are considered during scoring, each response receives a single, holistic score representing the raters' judgment of which score band level was best evidenced in the response. The score band chart and associated training materials were developed based on research into the components of communicative competence (Douglas & Smith, 1997; Powers, Schedl, Wilson-Leung, & Butler, 1999).

Two communicative competencies are particularly relevant to the issue of graph comprehension: discourse competence and functional competence.

Discourse competence relates to the coherence and cohesiveness of a response. Is the response well organized and well developed, and does the speaker cue the listener to the organization (e.g., "First we see that…," "In contrast…")? For the graph in Figure 1, a partial response demonstrating low discourse competence is: (ellipses refer to short pauses in speech)

(1) *the good hours...ah for age...ah,...between age ...50 and 60 is ten percent....And...the pleasant ...colleagues...for...ah...for age...20 to 30...is ten percent...and...ah for...50 to 60 is twenty percent....*

Responses low in discourse competence tend to be list-like, consisting of phrases connected by "and" but showing neither a strong organizing structure nor development. A response showing stronger discourse competence is:

(2) *…for adults...uh,...between age two,...20 to 30,...they value interesting work as their most important thing....well...for the old man...that's not important....Other points I should compare is uh,...is the low stress ...for the old man they...they prefer low stress and...while for the younger men…*

This response guides the listener better by using phrases such as "for the old man…" and "Other points I should compare…".

Functional competence is the ability to use language to transfer information and ideas to accomplish a goal. It is demonstrated by the extent to which a person communicates an intended goal. For example, we all know people who "beat around the bush" while you are wondering when they will get to their point. For the graph in Figure 1, a partial response demonstrating low functional competence is:

(3) *Ok, people...around the age...20 to 30...I guess started like...ah...just youngsters...they are...um... they good hours up like twenty percent ...and... only...ah...at the age of 20 to 30 ...the people who are interested ...are only forty percent*

This response does not communicate what information was provided in the graph, partially because the speaker misrepresents the meaning of "good hours" and "interesting work." Response (1), in contrast, does a good job of describing the information and so was rated higher on functional competence than was response (3).

The other two competencies appear less likely to be affected by the particular characteristics of a graph. **Sociolinguistic competence** is the ability to demonstrate an awareness of audience and situation. **Linguistic competence** refers to more basic speech issues such as vocabulary selection, pronunciation, and syntax.

The Theory

Most theories of graph comprehension include the processes of (a) encoding a visual feature of the graph or data (sometimes referred to as a "visual chunk") and (b) interpreting that feature with respect to basic graph knowledge (e.g., a line going up means something is increasing) and specific graph content (e.g., "bicycle sales are increasing"). Carpenter and Shah (1998) provide evidence that comprehension occurs through repeated cycles of encoding and interpretation, building up more inclusive understanding of the graph. Thus, the more information (the greater the number of visual chunks) in a graph to integrate, the longer it takes to comprehend a graph.

We hypothesize that fewer visual chunks similarly lead to higher quality descriptions. Fewer pieces of information to describe leaves more time and cognitive resources for

communicative tasks such as providing cues for the listener as to the organization of the description, describing each piece of information succinctly, and so forth.

What are the visual chunks in multi-variable bar graphs? Shah, Hegarty, and Mayer (1999) argue that each group of bars associated with a particular value on the x-axis form a visual chunk. Consistent with this theory, participants' descriptions of bars graph tend to be organized around these chunks. However, this impoverished definition depends solely on the x-axis scale of the graph, accounting neither for the visual properties of the data nor what information is represented by each group of bars. The present work requires a richer definition of visual chunks.

Our theoretical claim is that a visual chunk should play the same role as a proposition in text comprehension models (e.g., Kintsch, 1998). That is, in addition to being visually distinct as guided by Gestalt principles, a visual chunk must encode a single unit of information. A group of bars need not be a single visual chunk as is claimed by Shah et al. Rather, that group would be encoded as a single unit only if it represented a single unit of information (e.g., "Older people value salary the most").

Consider the graphs shown in Figures 1 and 2. These graphs represent the same data set, but switch the variables represented along the x and z (bar shades) dimensions. Which should be easier to describe? Figure 1 incorporates fewer visual chunks than does Figure 2 (two vs. five), so according to our hypothesis should elicit descriptions with higher communicative quality. Figure 1 has two groups of bars, each with one category that is much higher than the rest: describing this feature succinctly summarizes the data represented in the group. Thus, a straightforward description would be to make the global comparison within one age group (e.g., "For Age 20-30, interesting work is the most important") and then the other age group. While such a response does not necessarily capture every nuance of the data, it does capture the essential difference between the two groups. By our enriched definition of visual chunks, it is important that each x-axis group of bars in Figure 1 contain an obviously maximal value. Otherwise, each group might be perceived as separate chunks (each bar), potentially diminishing the quality of descriptions that the graph elicits.

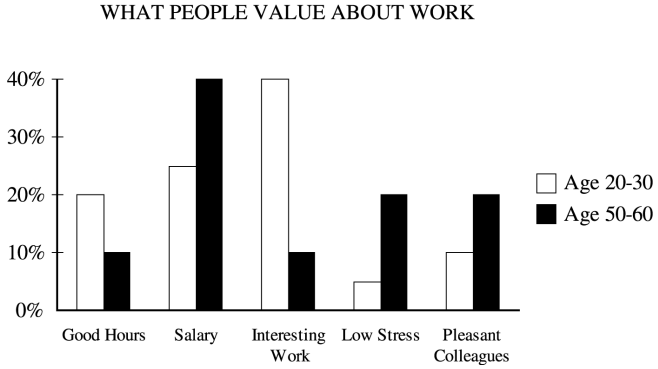


Figure 2: Alternative form of Figure 1
[More visual chunks]

Figure 2, in contrast, has five visual chunks: the relative height of the bars within each category. Thus, more time is needed to comprehend the graph, and the communicative quality of any descriptions of this graph should be lower than those of Figure 1.

This task analysis is not necessarily intuitively obvious. Although there are fewer visual chunks in Figure 1, the graph introduces five different shade-category mappings that might need to be either remembered or refreshed by looking at the legend (Lohse, 1993). From this alternative task analysis, Figure 1 might impose a heavier working-memory (WM) burden than Figure 2 because the latter has only two shades representing the two age groups. This alternative task analysis predicts that Figure 2 would elicit descriptions of superior communicative quality.

To test the visual chunk hypothesis, we conducted an experiment that manipulated two factors with the potential to affect the descriptive ease of a graph. First, as illustrated by Figures 1 and 2, we created two graph organizations for each of four data sets by switching the variables represented along the x-axis and by the differently shaded bars (the z-variable). One graph organization presents a smaller number of visual chunks (2-3 chunks depending on the data set) than the other organization (4-6 chunks). These two graph organizations will be referred to as the **few-chunks** (e.g., Figure 1) and **many-chunks** (e.g., Figure 2) graphs. The few-chunks graphs' organization minimizes the amount of information to be described, and is therefore predicted to elicit better descriptions.

An alternative to the visual chunks hypothesis is that a comparison between two groups is simply a more natural way to describe a graph. In other words, any superiority of the few-chunks graphs might be due to a particular descriptive strategy.

This alternative hypothesis suggests the possibility of drawing participants' attention to the fewer chunks even within a many-chunks graph (e.g., seeing the maximal values for the two age groups in the many-chunks graph). To investigate this possibility, we introduced alternative task prompts. **Open-ended** prompts were the same for all graphs and asked the participant to "Describe the information given in the graph." **Directive** prompts identified the critical contrast in the graph, suggesting more directly what should be described. For example, for Figure 1 the prompt was "Describe the changes in work values between the two age groups."

Method

Participants

Thirty-nine students (19 female, 18 male) participated in the experiment. Ten students[1] were recruited from each of four universities in the U.S., and students participated at their local institution. Eighty-five percent of participants were

[1] Due to technical difficulties, one participants' data were lost, so one school contributed only nine students.

doing graduate or post-graduate work; others were juniors or seniors. Participants ranged in age from 21 to 45, with an average age of 29. Students' reported fields of study were medicine (20%), math or science (18%), humanities (12%), business (8%), and social science (7%).

Each institution was asked to recruit eight non-native English speakers and two native English speakers. Most of the participants (n = 19) were native speakers of a Chinese dialect; other languages were reported by no more than two or three participants (a mix of Asian, European, and Middle Eastern languages). There were seven native English participants because one institution recruited only one native English speaker instead of the request two. Most of the students had been living in the U.S. for fewer than two years (n=22); the remaining students were evenly split between those that had lived in the U.S. 10 or more years (n=9) and between 2 and 10 years (n=8).

Materials

We constructed four data sets to be graphed as bar charts. Each data set had its own story line, which had been reviewed by professional test developers for comprehensibility to non-native speakers of English. The data represented the interaction of two independent variables, with one variable having fewer levels (2-3) than the other (3-5). The variables with fewer levels were either years or age groups (as in Figure 1). The other variables were either nominal categories (e.g., work values) or intervals (e.g., hours in a day).

We created two graphs from each data set, for a total of eight graphs. One graph in a pair placed the 2-3 level variable along the x-axis and represented the other variable on the z dimension (the different shades of bars)—this organization created the few-chunks graphs. As per our enriched definition of visual chunks, on the few-chunks graphs, each group of bars included one bar (unique to that group) clearly higher than the others. The many-chunks graph was created by switching the variables represented along the x and z dimensions.

Design

The independent variables of graph organization and prompt directness were implemented in a completely within-subjects design: each participant received all four graph types. The organization type alternated, with half the subjects receiving few-chunks graphs first and half receiving many-chunks graphs first. Because of the possibility of one prompt type influencing the next, that variable was implemented using an ABBA design, with half the subjects receiving an open-ended prompt first and half receiving a directive prompt first.

Preliminary analyses suggested no *a priori* differences among the participants from each school in terms of their communicative competence in English or in their familiarity with reading graphs.

Procedure

Each university conducted one data collection session of 10 students. Sessions were typically conducted in a language lab or similar equipped facility. Besides a test booklet, each student had a tape recorder and headphones. Students heard the prompts over their headphones and spoke their responses, which were recorded on audiotape.

The questions were administered in two sets, with a short break between the sets; each set consisted of nine non-graph questions followed by two of the experimental questions. After both sets were administered, students were given a brief graph familiarity questionnaire. The questionnaire consisted of several questions concerning graph interpretation, a section on self-reported graph familiarity, and a short demographic questionnaire.

Measures

We obtained three types of dependent measures from each response: response latency, holistic scores, and four component scores. **Response latency** is the number of seconds between the end of the spoken prompt and when the participant began speaking. The timing was done by a research assistant unaware of the purpose of the experiment, using an on-line stopwatch while listening to each tape.

Each response was also scored by highly experienced TSE raters, each rater having participated in many rating sessions each year for five or more years. Raters produced a **holistic score** in a way identical to how actual TSE responses are scored. To provide finer-grain scores than the 5-level scale described earlier, each rater was asked to indicate whether a score fell into the high, middle, or low end of the score band. Thus, raters provided scores such as "high 40" or "low 60." Raters often discuss responses in this way, so producing this additional information was not difficult. In converting these relative rankings into scores, "middle" scores were unadjusted to facilitate comparison between these scores and the typical score scale for the TSE. In the analyses, a "high" score adds 3.3 to the band level (e.g., "high 40" becomes 43.3) whereas a "low" score subtracts 3.3 from the band level ("low 60" becomes 56.7).

Finally, each rater was asked to provide a score for each of the **component competencies** in the TSE Score Band Chart, as described earlier. Thus, each response received a discourse, functional, sociolinguistic, and linguistic score. These scores were rated on the typical 5-level (20-60) scale.

Results

We look at the effects of graph organization and prompt type from three perspectives. First, what are the effects on response latency? According to Carpenter and Shah (1998), a greater number of visual chunks should lead to longer latencies because of the greater number of encode-interpret cycles need for comprehension. Second, what are the effects on holistic scores? As we are looking at within-subject performance, any effects suggest an influence other than a person's own communicative competence on the score (i.e., variance irrelevant to the construct intended to be

measured). Finally, as a follow-up to the effects on score, we look at the effects on the components of the score—the individual scores on discourse, functional, sociolinguistic, and linguistic competence.

We ran a 2x2 repeated-measures MANOVA, with graph organization (few- or many-chunks graphs) and prompt type (directive or open) as within-subjects factors and response latency as the dependent measure. There was a significant main effect of graph organization ($F(1,37^2)=4.0$, $p=.034$). Participants spent less time inspecting the few-chunks graphs before responding ($M=5.5$; $SD=3.7$) compared to the many-chunks graphs ($M=6.8$; $SD=4.6$). The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

Similar results were obtained for holistic scores. An identical 2x2 repeated-measures MANOVA revealed a significant effect of graph organization ($F(1,38)=8.1$, $p=.007$). Participants received higher scores when responding to the few-chunks graphs ($M=47.7$; $SD=9.1$) compared to the many-chunks graphs ($M=46.1$; $SD=9.5$). The main effect of prompt type was not significant nor was the interaction of graph organization and prompt.

The effects of graph organization on response latency and holistic scores were also observed in the sub-sample of seven native English speakers, albeit attenuated due to ceiling effects. Native speakers were quicker to respond to few-chunks graphs (3.6 sec) than to many-chunks graphs (4.2 sec) and produced better responses to those few-chunks (60.7 versus 59.5). These trends are consistent with the idea that the effects of graph structure are not just due to language skill, but rather that by using non-native speakers we accentuated differences that otherwise might have been difficult to detect.

Table 1. Mean (SD) scores by graph type.

Competence Component	Graph Type	
	Few-chunks	Many-chunks
Discourse	47.1 (8.6)	45.3* (9.9)
Functional	47.1 (8.7)	45.8 (9.9)
Sociolinguistic	46.2 (8.8)	45.5 (9.1)
Linguistic	48.0 (8.8)	47.2 (8.5)

Note. Each graph type score is the mean of the two scores for each participant. N = 37 per cell because one participant's component scores were unavailable. * $p < .05$

What types of effects does graph organization have on participants' responses? Are responses to few-chunks graphs more expressive or more linguistically precise? While we might expect graph organization to affect how well organized a response is (i.e., discourse competence), it might be the case that a poorly organized graph increases WM load, so impinges on all language competencies.

Table 1 shows the effect of graph organization on each of the competency scores. As expected, discourse scores were significantly higher (via two-tailed, paired-samples t-test) for the few-chunks graphs: responses to these graphs were rated as more coherent and cohesive. There was an almost significant difference on the functional scores, whereby participants' responses to few-chunks graphs reflected language more appropriate to the task than did their responses to many-chunks graphs. There were no differences between the graph types in participants' ability to express their knowledge of audience (sociolinguistic) or in their pronunciation or grammar (linguistic).

Thus far, the results are consistent with the model that better performance is achieved with graphs that have fewer visual chunks. But are participants describing the visual chunks predicted by the theory? For the few-chunks graph in Figure 1, participants' descriptions should include the global comparison between the highest category in a bar group and the other bars in that group (e.g., "Interesting Work is most important for the 20-30 year olds"). For the many-chunks graph, descriptions should instead include discrete comparisons within a category (e.g., "Interesting Work is more important to the 20-30 year olds than to the 50-60 year olds").

To address whether participants describe the expected visual chunks for these two graphs, we analyzed the first piece of information mentioned in their responses. Given the speeded nature of the task, the first graph feature mentioned should be the most salient to the participant.

Participants' descriptions were consistent with their describing the two graphs in terms of the predicted visual chunks (Table 2). Participants mentioned first the global features of the data significantly more often when the graph was organized to accentuate these features (few-chunks graph) and mentioned first the discrete comparisons (the relative-height visual chunks) of the many-chunks graph ($\chi^2(1)=11.8$, $p<.001$).

Table 2. Graph type by first description.

Graph Type	Global Comparison	Discrete Comparison
Few-chunks (Figure 1)	19	1
Many-chunks (Figure 2)	8	10

[2] Due to technical difficulty, one participants' latency was not obtained.

Discussion

The research presented in this paper replicates and extends basic research on graph comprehension. The results provide support for the hypothesis that graphs with fewer visual chunks are easier to describe. Participants took less time to scan the few-chunks graphs before speaking, which replicates Shah and Carpenters' (1998) results. Graphs with fewer chunks also elicited descriptions of greater communicative quality. Furthermore, the organization of a graph had a very specific influence on the descriptions provided by participants: graphs with fewer visual chunks led to more cohesive and coherent descriptions. If the many-chunks graphs were worse because of lower overall comprehensibility, we would expect more aspects of descriptive competence to be affected. Future research might further extend Shah and Carpenter's processing model to explain the mechanisms by which the higher quality descriptions are facilitated.

Interestingly, incorporating a directive prompt had no influence on participants' descriptions. Although it is dangerous to draw conclusions from null results, this lack of effect is consistent with the idea that visual chunks are a visual processing phenomenon and might not be influenced by directions on problem-solving strategy.

The visual chunks hypothesis—fewer visual chunks leading to descriptions of higher communicative quality—has practical implications, suggesting desirable characteristics of graph questions for the Test of Spoken English. For example, two or three visual chunks in a graph might be the limit of what is reasonably possible to describe within one minute. For multi-variable bar graphs, this recommendation would mean limiting the number of bar-groups placed along the x-axis and, as per the enriched definition of visual chunks, ensuring that each group encodes a single unit of information.

The visual chunks hypothesis is applicable to a wider range of graph types, as long as we can adequately define the visual chunks. For example, other research (Carpenter & Shah, 1998; Carswell, 1993; Shah, Hegarty, & Mayer, 1999) suggests definitions of visual chunks for multi-function line graphs: each non-parallel line is a visual chunk, although each "reversal" in a line (e.g., changing from an upwards to a downwards slope) is perceived as a separate chunk. By assuring that any line graphs have no more than two or so visual chunks according to these definitions, we would predict such graphs to be straightforward to describe.

In line with the overall theme of the conference, applied research should adapt theories and results from the basic research literature to solve real-world problems, and then contribute back to the theoretical literature from which it drew. By applying theories of graph comprehension to produce empirically supported recommendations for the design of TSE graph questions and, in the process, enriching the theoretical construct of visual chunks, the applied research presented in this paper achieves these goals.

Acknowledgments

This research was funded by the Test of Spoken English program of the TOEFL Policy Council. Peter Cheng was supported by the UK Economic and Social Research Council through the Centre for Research in Development, Instruction, and Training. We thank Shauna Cooper, Susan Lynn Martin, and Venus Mifsud for their assistance with this work, and Malcolm Bauer, Ann Gallagher, Patrick Kyllonen, and Valerie Shute for useful comments on earlier drafts of this paper. We are grateful to the TSE program staff—especially Emilie Pooler, John Miles, and Evelyne Aguirre Patterson—and to the TSE raters for their contributions to this project.

References

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*, 75-100.

Carswell, C. M. (1993). Stimulus complexity and information integration in the spontaneous interpretations of line graphs. *Applied Cognitive Psychology, 7*, 341-357.

Carswell, C. M., Bates, J. R., Pregliasco, N. R., Lonon, A., & Urban, J. (1998). Finding graphs useful: Linking preference to performance for one cognitive tool. *International Journal of Cognitive Technology*, *3*, 4-18.

Douglas, D., & Smith, J. (1997). Theoretical underpinings of the Test of Spoken English revision project (ETS Research Rep. No. RM-97-02). Princeton, NJ: Educational Testing Service.

Katz, I.R., Xi, X., Kim, H-J., & Cheng, P. C-H. (2002). Elicited speech from graph items on the Test of Spoken English (ETS Research Report RR-02-XX). Princeton, NJ: Educational Testing Service.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction, 8*, 353–388.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing*. Mahwah, NJ: Erlbaum.

Powers, D., Schedl, M., Wilson-Leung, S., & Butler, K. (1999). Validating the revisted Test of Spoken English against a criterion of communicative success. *Language Testing*, *16*, 399-425.

Shah, P., Hegarty, M., & Mayer, R. E. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology, 91*, 690-702.

Test of English as a Foreign Language (TOEFL) (2001). *TSE and SPEAK score user guide*. Princeton, NJ: Educational Testing Service [also available through http://www.toefl.org/pubs/pubsindx.html]

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition, 27*, 1073-1079.

Sex, Myths, and Adolescents' Conceptual Understanding of HIV

Alla Keselman (ak454@columbia.edu) and Vimla L. Patel (patel@dmi.columbia.edu)
Laboratory of Decision Making and Cognition, Department of Medical Informatics, Columbia University
622 West 168th Street, VC-5, New York, NY, 10032-372

Abstract

Research on knowledge organization and how this develops with education and training may provide insight into the alarmingly limited effectiveness of school HIV education curricula. The present study investigates the nature of adolescent knowledge of HIV and its relationship to reasoning. Middle and high school students were interviewed about their understanding of HIV and were also asked to critically examine problem scenarios that contained myths about HIV. The findings suggest that adolescents lack understanding of basic biological concepts around which they could build well-structured schemata of HIV. As a result, their HIV knowledge exists as a collection of disjointed facts, not conducive to effective application for reasoning. The implications for school-based HIV interventions are discussed.

Introduction

Despite growing awareness about HIV and AIDS, the outbreak of the disease continues unabated. Current assessments of the demographics of AIDS indicate that the disease disproportionately hurts the young, the poor, and urban minorities (CDC, 1999). Schools respond to the problem by producing educational interventions, aimed to teach adolescents about HIV risks and prevention. In particular, the New York City Board of Education mandates its schools to provide six hours of HIV Education annually at every grade level.

Unfortunately, in spite of such educational efforts, the statistics remain grim. Evaluations show that many existing interventions, while succeeding in increasing teenagers' knowledge about HIV and AIDS, do not lead to the decrease in high-risk behaviors (Brown et al., 1992; Langer & Tubman, 1997). These failures lead HIV educators to a conclusion, currently prevalent in HIV education literature, that knowledge about HIV has little bearing on real-life behavior.

We believe that in many previous studies, the relationship between knowledge of HIV and its real-life application was obscured by methodological weaknesses of HIV knowledge assessment measures. Typically, these studies assess knowledge as the ability to answer simple factual questions by selecting from true/false or multiple-choice answer options (Siegel et al., 1995). Such measures do not provide any insight into the nature and organization of adolescents' HIV knowledge which is critical to its applicability. The present study addresses two questions. First, what is the nature of adolescent knowledge about HIV? Second, to what extent do adolescents apply this knowledge when reasoning and evaluating information in the context of HIV? Answering these questions employing cognitive methods could provide important information for improving HIV Education curricula for American schools.

Research on expertise has long established that differences between expert and non-expert knowledge extend well beyond the difference in content richness. Studies show that some forms of knowledge organization are more suited for effective application than others. Expert knowledge is coherent and is organized in meaningful patterns around key concepts and ideas (Chi et al., 1981). In contrast, novices frequently organize their knowledge schemata around superficial surface attributes, rather than big ideas (Chi et al., 1982). Compared to novices', experts' knowledge schemata also contain more interrelations among individual concepts and ideas (Chi et al., 1981). As a result, experts have more efficient methods of deciding which chunks of information are essential for solving a particular problem, of retrieving that information efficiently and of applying it correctly. While novice and expert knowledge represent two endpoints of the trajectory, the development of expertise is a long process, which may be conceptualized as a gradual shift from flat and fragmentary to systematic and multi-layered knowledge structures (diSessa, 1993). This process is non-monotonic; often, an increase in knowledge results in a temporary drop in performance, while the new knowledge is being integrated with the existing knowledge (Patel & Groen, 1991).

Studies of lay understanding of health and disease provide us with domain-specific information about the kinds of knowledge that lay people use when reasoning about health issues. When reasoning about health, lay adults frequently rely on their intuition, as well as cultural, social and experiential knowledge (Sivaramakrishnan & Patel, 1993). In doing so, they often misattribute disease causality, viewing symptoms or co-factors of diseases as their causes. Lay scientific knowledge of relevant biological concepts is dissociated from experiential and cultural knowledge, fragmented and is often used opportunistically. This results in low internal consistency, self-contradictions, "loose ends", factual errors and misconceptions (Patel, Kaufman, & Arocha, 1999).

Findings from research on expertise and health cognition suggest that in order to assist effective real-life reasoning, adolescent models of HIV need to integrate superficial factual knowledge, conceptual biological knowledge, and experiential/practical knowledge into a coherent, uniform system. However, schools typically teach students about HIV within health education curriculum, which is separated from science/biology curriculum. HIV education is factual in nature. Moreover, while HIV education is usually introduced in early grades, the first comprehensive biology course is taught in high school, with little connection to adolescent real-life health concerns.

We hypothesize that in the prevalence of the current educational practice, integration of different kinds of knowledge and deep understanding of the mechanism of HIV will not occur. Therefore, adolescents are likely to have model of HIV that is incomplete and saturated with misconceptions, based on practical analogies and non-normative intuitive biology (Carey, 1985). In spite of its deficiency, this model is likely to include accurate factual knowledge of HIV risks and prevention factors. As a result, reliance on this model of HIV is likely to enable adolescents to successfully pass multiple-choice survey assessments, but is not likely to help them reason through complex real-life situations that require deeper understanding. We also hypothesize that while older adolescents may have more basic biological knowledge than younger adolescents, their knowledge is likely to be fragmented and not assimilated into coherent conceptual model, essential for effective reasoning and problem solving. As a result, adolescent ability to reason about novel situations in the context of HIV is expected to be limited. If confirmed, our hypotheses have important implications for the structure of knowledge-based HIV interventions.

Method

Subjects

The subjects include twenty adolescents from two New York inner-city schools, including ten seventh-grade middle school students (4 boys and 6 girls) and ten high school students from grades 9 through 12 (5 girls and 5 boys). The subjects are referred to by the school level (MS = middle school; HS = high school).

Procedure

Each subject participated in an individually administered 45-minute session which included two assessment measures, a semi-structured interview about HIV and a reasoning task that required evaluating information on a simulated website about ways to reduce the risk of HIV infection. The purpose of the interview was to assess students' knowledge about HIV risks and prevention, as well as their understanding of the underlying biological concepts. Questions of the interview were designed to cover the scope of HIV issues without requiring specialized biological knowledge. In the reasoning task, subjects were presented with a simulated web site about HIV, supposedly created by a group of high school students. The site contained four passages that presented and supported three erroneous claims and one accurate claim about ways to reduce the risk of HIV infection. Understanding the erroneous nature of the information in the passages required basic knowledge about HIV infection and disease progression. After reading each passage, students had to express and justify their opinion about the truthfulness of the information. This paper presents analysis of students' performance on one of the erroneous passages which is presented below.

Passage 2

If you had unprotected sex, you can minimize the risk of becoming HIV-positive by expelling the virus from your body through urine and sweat.

As you probably know, HIV is transmitted through bodily fluids: blood, sperm, etc. This means HIV lives in those fluids and travels with them. If one person's infected fluids get inside another person, the second person also becomes infected. Logically, if infection gets inside a body through fluids, it can also get out of the body through fluids. Fluids that leave human body are urine and sweat. So, if your condom broke, making a lot of fluid leave your body can minimize your risk of getting HIV. To lose fluids, drink lots of water (this will make you go to the bathroom a lot); put on warm clothes and do something physically active. The trick is to do these things early, before the virus has a chance to multiply and become strong.

Coding Scheme for Conceptual Understanding

The analysis of students' knowledge draws on cognitive research in science education concerned with characterizing progressions of conceptual understanding (e.g., Vosniadou, 1999) and on research in the development of biomedical understanding (Patel et al, 1995). Based on consultations with HIV educators, pilot testing and students' responses to the interview, three conceptual models of HIV were generated: *advanced*, *intermediate*, and *naïve*. The models reflect students' understanding of three concepts: the nature of HIV, the mechanism of HIV-infection, and disease progression. Students were assigned to a model, if at least two out of these three concepts were consistent with the model description. Two investigators scored a portion of the protocols, to ensure satisfactory level of inter-rater reliability.

Advanced Model involves understanding of HIV-relevant biological structures and processes on the cellular level, without requiring specialized biological knowledge. This understanding should be evident on the following dimensions: 1). Definition of HIV. Students recognize that HIV is a virus with specific cellular-level structural and functional components (e.g., lacks organelles). 2). Mechanism of HIV infection. HIV enters the body through exchange of bodily fluids and penetrates T-cells of the immune system. 3). Disease progression. HIV replicates within T-cells and eventually destroys them and disables the immune system.

Intermediate Model involves understanding of HIV on systemic, but not on the cellular level, as reflected on the following dimensions: 1). Definition of HIV. HIV is a biological entity (details of the viral structure and characteristics are not provided; replication is not mentioned) 2). Mechanism of HIV infection. HIV enters the body through exchange of bodily fluids. Entering T-cells is not mentioned. 3). Disease progression. HIV compromises the immune system, and the body succumbs to opportunistic infections. The role of T-cells may be mentioned, but without any notion of intracellular processes.

Naïve Model does not employ relevant biological concepts on either systemic or cellular level. Instead, it is build around intuitive lay concepts of health and disease. Scientific biological concepts are either not known, or not integrated with HIV knowledge. Students characterized by this model lack basic biological concepts around which they could organize their knowledge of HIV. This model, however, does not preclude individuals who hold it from knowing an extensive collection of facts about HIV risk factors and prevention measures. The naïve model is reflected in the following understanding of the three critical concepts: 1). Definition of HIV. HIV is an illness. No mention is made of the virus as a causal agent. 2). Mechanism of HIV infection. HIV enters the body. Bodily fluids are not implicated in the process. 3). Disease progression. It makes the person sick. No mention is made of the effect of HIV upon the immune system.

Coding Scheme for the Reasoning Task

The method of *semantic representations* was chosen for the analysis of excerpts of protocols of students' reasoning and information evaluation. Semantic network analysis is a formal method for representing relations among concepts through directed, labeled graph structures (Patel & Groen, 1986). In these structures, nodes represent concepts and links (directional arrows) represent relations among them (see Figures 1 and 2 for examples). The relations are binary relations, such as *causal*, *conditional*, *alternating or* and *exclusive or* relations. Semantic network is a method that allows one to analyze verbal protocols for the direction of reasoning (forward vs. backward), coherence and granularity of concepts.

Results and Discussion

The Results and Discussion section is organized into two parts. The first part, *Conceptual Understanding of HIV*, presents the analysis of students' models of HIV based on their responses to the semi-structured interview. The second part of the results, *Reasoning in the Context of HIV*, presents adolescents' responses to the reasoning task, described in the Methods section.

Conceptual Understanding of HIV

The classification of individual students' HIV models is presented in Table 1. On the basis of the stated criteria, 9 middle (MS) and 2 high school (HS) students' models were classified as naïve, 5 HS students were classified as intermediate and 3 HS were designated as advanced. Numbers in bold represent dimensions consistent with the model assignment for that student.

Table 1: Classification of students' HIV models.

	Naive	Intermed	Advanced
Virus			
Middle	**1, 2, 3, 4, 5, 6, 7, 8, 9**	**10**	
High	**1, 3**, 2, 5	**6, 8, 9**	**4, 7, 10**
Infection			
Middle	**1, 2, 3, 4, 5, 6, 7, 8, 9**, 10		
High	**1, 3**, 9	**2, 5, 6, 8**	**4, 7, 10**
Progress			
Middle	**2, 3, 4, 5, 6, 7, 9**	1, 8, **10**	
High	**1, 3**, 6	**2, 5, 8, 9**	**4, 7, 10**

Understanding at the Naive-Model Level. The eleven students assigned to this model (MS 1-9, HS1, HS3) showed almost no understanding of the biological concepts of virus, infection, and immune system, crucial for developing a more advanced model of HIV. When asked what a virus was, these students either characterized it as "sickness" (MS1, MS2, MS4, MS9), or provided specific examples (e.g., "stomach virus", "coughing virus") (MS3, MS7, HS1, HS2). Students' characterization of infection was similar to their characterization of virus. Nine students (MS2, MS3, HS1 and HS2) showed no evidence of having a conception of the immune system. Only two students (MS1 and MS8) were able to state that HIV destroyed

T-cells and weakened the immune system (e.g., MS1 described T-cells as "police officers" that help fight diseases).

Not having a basic understanding of infection, these students had no common theme that would unite different routes of HIV transmission (e.g., exchange of fluids). All the children associated HIV with sex. Older children also associated it with drugs (HS2) and exchange of blood (HS1 and HS2), thus illustrating that an increase in HIV knowledge may not correspond with an increase in understanding. Without seeing exchange of fluids as the common theme in all of the routes of transmission, adolescent' understanding of how protection measures work remains weak (e.g., Interviewer: "If you use a condom, can you still catch HIV or not?" MS3: "I don't know"). Such lack of understanding of how exactly HIV is transmitted during sexual intercourse may lead students to believe that they can control HIV by regulating the amount of sex that they have. Indeed, two adolescents (MS2 and MS4) mentioned that people who had HIV could make themselves feel better, if they stopped having sex.

Without perceiving a virus as a microorganism, students had to find an alternative causal agent for the disease. Some subjects avoided the challenge by providing no causal agent at all, while two subjects (MS4 and HS2) mentioned dirt. The following example illustrates a 7th-grader's (MS4) misconception about the process of infection, which involves sex and dirt: "See, most people, like, they don't actually wash after having sex. See, if the person is dirty, …, you know, like the dirt, it goes into skin, like, it stays there for a long time, then it starts to go further in, then it starts mixing with blood, and that's how AIDS could probably form." This misconception makes students vulnerable to the erroneous belief that washing after sex may prevent infection. Additional potential misconceptions result from students' lack of understanding of the process of disease. Not knowing how HIV affects the human body over time, students in this model also typically did not understand the connection between HIV and AIDS. They referred to them as two different diseases, with one being more dangerous than the other (MS1, MS2, MS3, MS7, MS9, HS1, HS2).

Understanding at the Intermediate-Model Level. The six students assigned to this level (MS10, HS2, HS5, HS6, HS8 and HS9) had some biological understanding of HIV at the systemic level. Although none of the students could describe the viral structure or life cycle, four of them realized that HIV was a particle with physical properties, such as shape and size (e.g., HS6: "It might be a cell-looking thing"). Their understanding that the process of HIV infection involved exchange of fluids allowed them to unite various routes of HIV transmission around a single theme. Five of the students understood that HIV affects the immune system - and thus destroys body's defenses - and defined AIDS as the advanced stage of HIV infection (e.g., HS3: "AIDS is the sickness itself."). Overall, these students had conception of HIV that was sufficiently biologically grounded to provide some framework for organizing facts about HIV. As a consequence, they did not share any of the misconceptions, exhibited by the naïve students.

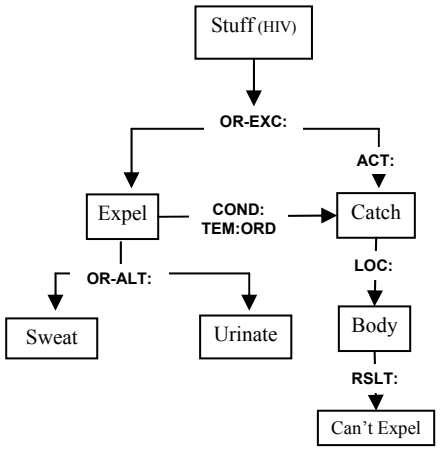
Understanding at the Advanced Model Level. The three students who demonstrated this level of understanding (HS4, HS7 and HS10) defined HIV as a virus, and described the virus as a microorganism that lacks organelles, but contains genetic material and can replicate inside of a host. They knew that the virus entered human body through exchange of fluids and penetrated white blood cells, replicated inside of these cells and eventually destroyed them. Both students stated that as the number of viral cells increased and the number of white blood cells decreased, the body became unable to fight opportunistic diseases.

Reasoning in the Context of HIV

Reasoning at the Naive-Model Level. Refuting the erroneous information in reasoning Passage 2 requires understanding that most HIV particles are "anchored" in white blood cells in the blood stream, and therefore, can not be expelled from the body through fluids. Students whose conception of HIV is at the naïve level do not have this understanding: in their model, HIV affects not specific cells that are located in the blood, and (for most, though not all students in this group) not even the blood in general, but "the body." Sweat and urine flow out of "the body", so the scenario should sound convincing to these students. The following justification of the agreement with passage provided by a seventh-grader (MS3) demonstrates the effect of naïve model of HIV on reasoning, "It makes sense… because the stuff is in the body, and you just need a release, release all of it out... before it really catches your body." Semantic network of this student's reasoning is provided in Figure 1. The network illustrates how this student's reasoning, while coherent, involves the level of conceptual granularity that is too crude to expose the fallacy of the myth.

Nine out of eleven naive-level students (MS 3-9, HS1 and HS2) agreed that HIV can be expelled through urine and sweat. Some of these students (MS 7 and MS 8) initially said that HIV could not be expelled, but later found the explanation in the passage convincing enough to change their mind. One of the students (HS1) gave additional support to his reasoning by providing an experiential analogy to the case of a person defeating cancer, "Yeah, this is true, this is true. Cause people can stop it like that. By exercising, like they said. Like that

lady, like I told you, she exercised her way out of cancer, so I think this is true, you can exercise your way out of HIV probably." Such opportunistic use of practical knowledge is typical of lay reasoning about health (Sivaramakrishnan & Patel, 1993). Notably, the subject who provided the analogy had explicitly stated in the course of the interview that HIV is incurable (HS1: "Once you get HIV, it's like, that's it, there is no coming back to it.")



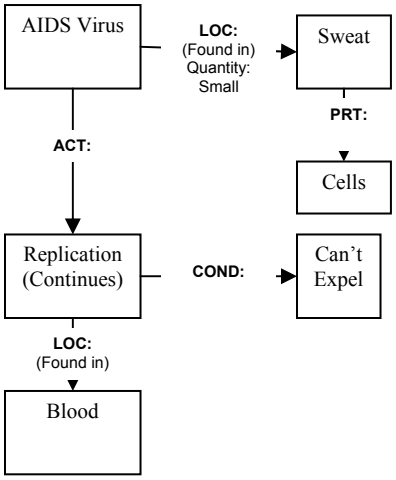
ACT: Action, ALT: Alternative, COND: Conditional Relation, EXC: Exclusive, LOC: Location, RSLT: Result , TEM: Temporality

Figure 1: Semantic network of naive-model reasoning.

Two students showed resistance to the myth presented in the passage, in spite of low biological understanding of HIV. One student (MS2) rejected the idea on the basis of common sense/practical reasoning, asserting that had information in the passage been true, HIV would have been defeated by now. However, she also conceded that the information in the passage was "a little-little true cause somebody wrote it." This illustrates that while common sense may be very useful in everyday health reasoning, its effectiveness is limited, if not supported by biological understanding. Only one naive-level student (MS1) refuted the myth on the basis of biological reasoning, stating that HIV could not be expelled from the body through urine and sweat.

Reasoning at the Intermediate-Model Level. Compared to the students assigned to the naïve model, the adolescents in this group had deeper understanding of the biological basis of HIV. However, the expertise literature characterizes a stage of development in which an increase in knowledge is accompanied by a temporary decrement in performance. Patel and Groen (1991) describe this as the intermediate effect, characteristic of a period during which recent knowledge is not yet fully assimilated. This phase of temporary disorientation is evident in the reasoning of intermediate level students. For example, unlike naïve model students, most intermediate model students understood the role of bodily fluids in HIV transmission. However, they often failed to utilize this information by inferring that different fluids contained different amounts of virus and did not interact with one another. As a consequence, four out of six students assigned to this model still found the "in with fluids, other with fluids" explanation convincing (e.g., HS6, "A fluid is what makes you get AIDS… drinking gets fluids out of your system… It probably would help.") Two students (HS2 and HS8) were able to refute the myth, although with some hesitations. Overall, intermediate model level understanding of HIV does not yet provide sufficient conceptual basis for consistent efficient reasoning in the context of HIV.

Reasoning at the Advanced-Model Level. All three students characterized at this level (HS3, HS7, HS10) rejected the myth with very high degree of certainty. This is not surprising, given their relatively rich conceptual model of HIV. In addition to knowing that HIV could enter human body with fluids and travel with them, these students also knew that the virus entered the blood stream, penetrated white blood cells and replicated within them. This allowed them to understand how the majority of the HIV particles were anchored in the bloodstream, thus uncovering the flaw of the passage, e.g., "AIDS virus, it's found in sweat… but only in minor quantities, so you can't just expel it, if you do it, there is going to be virus still in your blood that will just keep replicating."



ACT: Action, ALT: Alternative, COND: Conditional Relation, EXC: Exclusive, LOC: Location, PRT: Part

Figure 2: Semantic network of advanced model reasoning.

Figure 2 provides a semantic network of that student's (HS7) reasoning. The figure illustrates that while the complexity of the reasoning is comparable to that of the naïve model in Figure 1, the explanatory reasoning and use of conceptual knowledge demonstrates a greater degree of understanding.

Conclusions

In this study, students with adequate factual knowledge of HIV risks and prevention often lacked genuine conceptual understanding of HIV. With little conceptual understanding, students had difficulty evaluating dubious claims and reasoning about practical issues in the context of HIV. The dissociation between factual knowledge and conceptual understanding of HIV parallels the dissociation between HIV and science education in the schools. With biology taught separately from factual HIV education and introduced in later grades, adolescents have little understanding of the concepts of virus and immune system, which are critical to building accurate conceptual models of HIV. As a consequence, HIV knowledge remains in the form of a disjointed and sometimes erroneous collection of facts, which have minimal applicability to problem solving. When adolescents enroll in a high school biology course, they receive some grounding in the concepts that are relevant to understanding HIV. However, they do not receive an opportunity to integrate this biological knowledge with factual knowledge and the biological knowledge remains inert.

The present study cautions researchers against making a hasty conclusion about the lack of connection between knowledge of HIV and sexual behavior. We are not implying that understanding of HIV is the only factor that influences sexual risk taking. Non-cognitive factors (e.g., sexual arousal) can exert strong influence on decision-making (Lowenstein, 1996). However, in-depth understanding of basic HIV concepts is crucial to the success of any educational intervention. This study is part of a research program designed to impart robust conceptual understanding of HIV to adolescents.

Acknowledgments

This research is supported by American Educational Research Association postdoctoral fellowship award to the first author. We thank student who participated in this study, the educators who gave us their enthusiastic support and the schools that provided us with research sites. This paper has greatly benefited from critical comments and discussions provided by David Kaufman.

References

Brown, L.K., DiClemente, R.J., & Park, T. (1992). Predictors of condom use in sexually active adolescents. *Journal of Adolescent Health* 13, 651-657.

Carey, S., (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Centers for Disease Control and Prevention (1999). Young people at risk: HIV/AIDS among America's youth. www.cdc.gov/hiv/pubs/facts/youth.htm.

Chi, M.T.H., Feltovich, P.J., & Glaser, R., (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M.T.H., Glaser, R., & Rees, E., (1982). Expertise in problem solving. In R.J. Sternberg (Ed.). *Advances in the psychology of human intelligence (vol. 1*). Hillsdale, NJ: Erlbaum.

diSessa, A.A. (1988). Towards an epistemology of physics. *Cognition and Instruction, 10 (2 & 3)*, 105-225.

Langer, L.M., & Tubman, J.G. (1997). Risky sexual behavior among substance-abusing adolescents: psychosocial and contextual factors. *American Journal of Orthopsychiatry* 67, 315-322.

Lowenstein, G. (1996). Out of control: visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65(3),* 272-292.

Patel, V.L., & Groen, G.J. (1986). Knowledge-based solution strategies in medical reasoning. *Cognitive Science, 10*, 96-116.

Patel, V.L., & Groen, G.J. (1991). Developmental accounts of the transition from student to physician: some problems and suggestions. *Medical Education, 25*, 527-535.

Patel, V.L., Kaufman, D.R., & Arocha, J.F. (1999). Conceptual change in the biomedical and health sciences domain. In R. Glaser (Ed.), *Advances in Instructional Psychology, 5*, Mahwah, NJ: Erlbaum.

Siegel, D., DiClemente, R., Durbin, M., Krasnovsky, F., & Saliba, P. (1995). Change in junior high school students' AIDS-realated knowledge, misconceptions, attitudes, and HIV-preventive behaviors: Effects of a school-based intervention. *AIDS Education and Prevention, 7(6),* 534-543.

Simon, D.P., & Simon, H.A., (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: what develops?* Hillsdale, NJ: Erlbaum.

Sivaramakrishnan, M., & Patel, V.L. (1993). Reasoning about childhood nutritional deficiencies by mothers in rural India: A cognitive analysis. *Social Sciences and Medicine, 37 (7),* 937-952.

Vosniadou, S (1999). Conceptual change research: State of the art and future directions. In W. Schnotz, S. Vosniadou, & M. Carretero (eds.) New Perspectives on Conceptual Change. Amsterdam: Pergamon Press.

A Cognitive Task Analysis of Using Pictures To Support Pre-Algebraic Reasoning

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3890

Atsushi Terao (atsushi@cs.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3890 USA

Abstract

We present an analysis of hypothesized advantages of pictorial representations for improving learning and understanding of pre-algebraic quantitative reasoning. We discuss a "Picture Algebra" strategy that has been used successfully by 6th grade students as part of a new middle school mathematics curriculum. This strategy supports students in sense making both as they construct pictorial representations and as they use them to cue appropriate computations. Although we demonstrate that 6th grade students can use this strategy to successfully solve algebra-level problems, our detailed production rule analysis revealed limitations in our instructional approach and targeted areas for improvement.

Introduction

As part of a larger effort to develop a 6th grade mathematics course including both a textbook and Cognitive Tutor software (cf., Koedinger, Anderson, Hadley, & Mark, 1997), we have been exploring the use of pictorial representations to support student reasoning and learning (Rittle-Johnson & Koedinger, 2001). Here we investigate the claim that pictorial representations can help students gain early entry into algebraic reasoning and build a foundation that will facilitate more effective learning of formal algebra.

Why might using pictures or diagrams be advantageous? Cognitive scientists have presented arguments and experiments for the advantages of diagrams (e.g., Cheng, 1999; Larkin & Simon, 1987). According to Larkin and Simon (p. 98), "a diagram can be superior to a verbal description for solving problems" for three reasons. First, diagrams reduce problem-solving search by providing localized groupings of relevant information. Second, diagrams reduce the need for matching symbolic labels. Third, diagrams support perceptual inferences that are often easier than corresponding symbolic inferences.

Others have presented arguments for the use of diagrams for mathematics instruction in particular. The mathematics standards of the National Council of Teachers of Mathematics (NCTM, 2000) recommends use of pictures to support students in developing a conceptual understanding of mathematics. Pictorial representations are used extensively in Asian curricula (cf., Singapore Ministry, 1999). This usage *may* be a factor in the success of Asian countries on international mathematics assessments (TIMSS, 1996).

Despite these arguments for the advantages of pictures, there is also reason for caution. One argument for the use of alternative representations, like pictures, is that traditional instruction focuses too much on error-prone rote learning. However, students may also acquire rote procedures when learning to use alternative representations. Further, learning an alternative representation takes time that might be better spent learning the standard representation.

In this paper we introduce the "Picture Algebra" strategy, present student data on the use of it, and discuss a production rule model of the strategy and implications for transfer and instructional design.

Picture Algebra

Try to solve the Cans problem shown in Table 1 and reflect on the strategy that you use to do so. We have informally observed that many adults do not directly infer what arithmetic operations are needed to solve this problem. Instead, most begin by translating the problem statement to one or more algebraic equations, for instance, $x + (x + 9) + (x + 17) = 227$. They then perform transformations on the equation to arrive at a solution. Although a few use other means (cf., Hall et al., 1989), most find this problem difficult without the use of algebraic equations. In other words, this is arguably an "algebra problem" that we might expect to be out of reach of students without algebra instruction, for instance, 6th graders.

Figure 1A shows a 6th grade student's solution to this problem using a "Picture Algebra" strategy that was taught to students as part of our middle school mathematics curriculum. Like other problem-solving strategies, Picture Algebra can be described in two phases: a representation phase and a solution phase. In the representation phase, the student first translates the

Table 1. Three Problems Solved by 6th Graders

Cans: The sixth, seventh and eight grade classes brought in canned goods for the needy. They collected 227 cans between the grades. Sixth grade collected 9 more cans than eighth grade, and seventh grade collected 17 more cans than the 8th grade. How many cans did each grade collect?
Beanie: Robin has 6 fewer beanie babies than Angie. If they have 42 beanie babies altogether, how many beanie babies does Robin have?
CD: Carissa wants to buy the latest Out o' Sync CD. She also wants to buy a magazine and a poster about her favorite band. The magazine costs $8 less than CD, and the poster costs $12 less than CD. The total cost for all three items is $46. How much does each item cost?

phrase "sixth grade collected 9 more cans than eighth grade" into a box diagram by drawing a box to represent the cans collected by the 8th grade and a larger box to represent the cans collected by the 6th grade. The larger box is made up of two smaller boxes, one that is the same size as the 8th grade box and one that represents the 9 more cans the 6th grade collected. Similarly, the student draws two boxes to represent the 7th grade cans. In both cases, the extra box off to the right is labeled with the given "more-than" values, 9 and 17. In the final step of the representation phase, the student represents the total number of cans, 227, by drawing a bracket to the right of all the boxes.

This picture representation and the algebra equation, $x + (x + 9) + (x + 17) = 227$, are "informationally equivalent" in Larkin & Simon's terms. The three equal sized boxes on the left are analogous to the three x's in the equation, the extra boxes represent the + 9 and + 17, and the bracket and the 227 represent the equal sign and right hand side of the equation. Despite this similarity, the picture and equation representations are not "computationally equivalentÓ Whereas the equation is a 1D "horizontal" representation, the picture is 2D and takes advantage of both the horizontal and vertical dimensions to better support the inferences needed to solve the problem.

As we step through the Picture Algebra solution phase, notice how inferences are visually supported in a way they are not in the analogous equation solving steps. The first step in the solution phase is to parse or comprehend the representation, picture or equation. Past research has shown that both equation comprehension (Payne & Squibb, 1990) and equation production (Heffernan & Koedinger, 1998) are particularly difficult sets of skills for students to acquire. For instance, students must learn when they can and cannot rearrange elements of an equation. Although experts know they can in effect ignore the parentheses in "x + (x + 9) + (x + 17)" and transform the expression into "x + x + x + 9 + 17" and then "3x + 26", the steps involved are non-obvious to students and prone to error. In contrast, the analogous steps in the Picture Algebra strategy are visually supported. The vertical organization of the bars makes it more apparent that the three equal unknown boxes can be grouped separately from the two extra boxes for 9 and 17. As a consequence, students can "see" that if they subtract these extra boxes, they will be left with the three equal boxes. Notice the arithmetic on the left in Figure 1A where the student computes both 9 + 17 = 26 and 227 — 26 = 201. This step is analogous to the "subtract from both sides" transformation in equation solving, however, it is more perceptually intuitive as students can visualize chopping off the extra boxes. Next the student divides 201 by 3 (see the bottom middle of Figure 1A) to get the value, 67, of the three equal boxes. Finally, the student adds back the 9 and 17 to get the values of cans collected by the 6th and 7th grade.

Student Solution Data

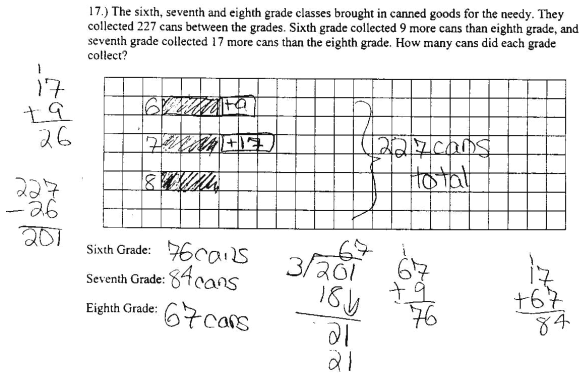
We analyzed 35 sixth graders' solutions to the three problems shown Table 1. The Cans and Beanie problems were given as part of a unit test. To get a sense for whether the representation or solution phase of Picture Algebra is more difficult for students, we provided a correct picture representation for some students but not for others. Half the students were given a test in which a diagram was given for the Cans problem but not for the Beanie problem. The remaining students were given a test in which a diagram was given for the Beanie problem, but not the Cans problem. Students solved the CD problem (with no diagram provided) as part of a warm-up activity on a later day. (One student was absent on each day.)

The Cans problem involves three unknown quantities (cans collected by 6th, 7th, and 8th grades) and more-than relations between these quantities. Students were initially instructed on Picture Algebra with simpler two quantity more-than problems. The Beanie problem also has two quantities but is more difficult because of the less-than (or "fewer") relation between them (Lewis, 1989). The CD problem combines both dimensions of difficulty as it involves three quantities and less-than relations. One question of interest is whether these dimensions of difficulty are independent.

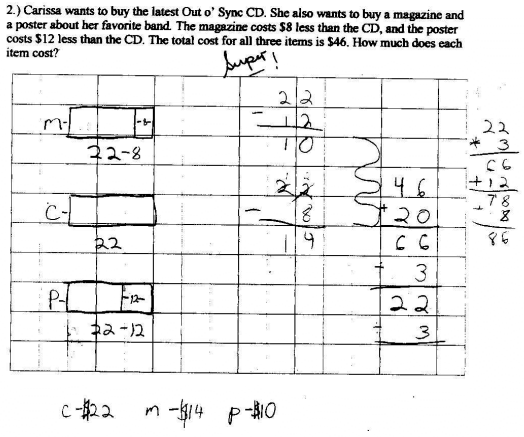
Overall Performance Results

The first result of note is that 6th graders can effectively use pictorial representations to solve "algebraic" problems. They were 68% correct on both the Cans and Beanie problems and 32% correct on the CD problem. Such problems are challenging for many older students. For instance, Bednarz & Janvier (1996, p. 120) found that pre-algebra students (same age as US 7th graders) were only 5% correct on the following problem:

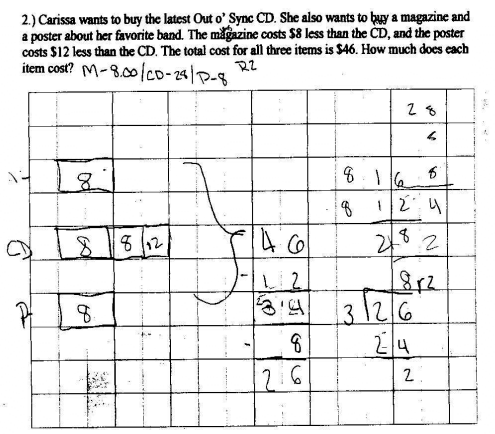
> 380 students are registered in sports activities for the season. Basketball has 76 more students than skating and swimming has 114 more than basketball. How many



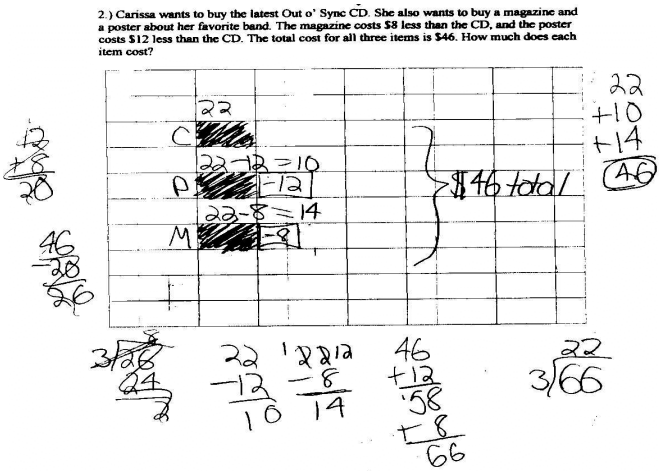
(A) Correct size-preserving picture for Cans problem



(B) Correct size-preserving picture for CD problem



(C) Wrong picture for CD problem



(D) correct abstract picture for CD problem

Figure 1. Correct and incorrect 6th grade students' Picture Algebra solutions.

students are there in each of the activities?

As another reference point, Koedinger & Alibali (1999) collected data on college students' performance on the following relatively simple problem:

> There are 38 students in class. If there are 6 more girls than boys, how many boys are in the class?

Students in a college algebra course were only 54% correct and even CMU students with a mean math SAT of 719 were 86% correct.

Students given a picture on the Cans and Beanie problems were 71% correct in both cases and were 65% correct on both without the picture. That being given the picture did not help much on these problems indicates that the representation phase was relatively easy for students. In contrast, the representation phase on the CD problem was not easy. Students drew a good picture in only 32% of CD solutions, but were 64% correct when they did. They were only 17% correct when a picture was missing or incorrect.

We were surprised at students relatively poor performance on the CD problem (32%) given that two key sources of difficulty in that problem, 3 quantities and less-than relations, were each present individually in either the Cans problem (3 quantities) or the Beanie problem (less-than relation). A simple surface-level analysis of problem difficulty factors would predict that students would be, at worst, 42% correct (.65*.65) on the CD problem. Clearly, such a surface level analysis is insufficient to understand why the CD problem was so hard for students. Our production rule analysis below provides a better explanation.

Diagramming Strategies

We classified the pictures students drew into 5 categories: size-preserving, incomplete, wrong, abstract, and no-diagram. In a "size-preserving" picture, all quantities, or parts of quantities, are represented by boxes whose sizes correspond with the

relative sizes of the quantities. That is, bigger quantities are represented by bigger boxes. Unknown quantities can be any size, but equal quantities must be represented by equal size boxes. All quantities are represented and all known quantities or parts of quantities are labeled with their given value. Figures 1A-B show examples of "size-preserving" pictures. An "incomplete" picture was one that was otherwise correct, but was missing quantities or labels. A "wrong" picture has boxes that are the wrong relative sizes. Figure 1C shows an example where the box for the magazine cost (top box) is the same size as the box for the poster cost (bottom box) when, in fact, the magazine cost is bigger than the poster cost.

An "abstract" picture is shown in Figure 1D. The boxes in Figure 1D violate the size preserving constraint. Whereas the CD cost is largest, it appears as the smallest box (the top one). However, some students correctly interpreted these extra boxes (labeled —12 and —8) as, in essence, having a negative size.

A potential advantage of the abstract picture is that it may incrementally move students toward the kinds of abstract inferences needed for symbolic algebra transformations. On the easier Cans and Beanie problems, the 7 solutions employing an abstract picture were all correct. However, one disadvantage of the abstract picture is that without the size-preserving cues, students can easily confuse the operations that need to be performed and, for instance, subtract when they should add. This confusion led 4 of the 10 students who used an abstract picture on the CD problem to an incorrect answer. Further, even among the 6 students whose abstract picture led to a correct solution, there is evidence of confusion. Notice the arithmetic on the left in Figure 1D where the student incorrectly subtracts 20 from 46 rather than adding 20 and 46. The student abandons this approach when he notices that 3 does not divide evenly into 26. This correct solution may not have come from understanding, but from a shallow school heuristic that problems usually come out even.

Production Rule Model of Picture Algebra

We performed a production-rule analysis of typical correct and incorrect student solutions to the three problems. Table 2 shows examples of key productions.

Production Trace of Cans Solution

The correct solution to the Cans problem with a size-preserving diagram (see Figure 1A) is traced with 13 productions. These productions are the same ones needed to solve a more-than problem dealing with two unknown quantities and thus we would expect good transfer from class instruction. To process 3 unknown quantities rather than 2, some productions are executed multiple times. When drawing a size-preserving picture the model first has to draw a "base" box (P2a in Table 2). This box represents the 8th grade cans. Other productions draw "equal-part" boxes for the equal parts of the 6th and 7th cans and "extra-part" boxes for the 9 and 17. The final production in the representation phase draws and labels the bracket with the total 227.

Table 2. Picture Algebra Production Rule Examples

P0a: Change-less-than-to-more-than If goal is to draw boxes to represent =Q1 is =X less than =Q2 then set goal to draw boxes to represent =Q2 is =X more than =Q1
P0b: Change-less-than–to-more-than-negative If goal is to draw boxes to represent =Q1 is =X less than =Q2, then set goal to draw boxes to represent =Q1 is minus =X more than =Q2
P2a: Draw-base-box-more-than If goal is to draw boxes to represent =Q2 is =X more than =Q1 and no box has been drawn then draw a first box for =Q1
P3c: Draw-comp-box-equal-part-repair If goal is to draw boxes to represent =Q2 is =Y more than =Q3 and first box for =Q2 with the size =S1 is drawn and no box for =Q3 has been drawn then draw a first box for =Q3 with the size =S1
P6a: Check-equal-size-boxes-with-extra If goal is to solve a problem with a box diagram and there are equal-size boxes and the value of an extra box for =Q1 is =X and the box with value =X has not been removed then set goal to remove the extra box

At this point, the solution phase begins with production P6a, which sets a goal to remove an extra box. This production fires twice in the Cans solution, once with =Q1 set to 6th grade cans and =X to 9 and a second time with =Q1 set to 7th grade cans and =X to 17. Once all quantities represented as extra boxes (i.e., 9 and 17) are subtracted from the total quantity (227), further productions find the base quantity by dividing the new total (201) by the number of equal-size boxes (3). The model finishes with productions for finding the other two quantities by adding the extra box values (9 and 17) to the value of the base quantity (67).

Production Trace of Beanie Solution

The key to the model's solution of the Beanie problem is production P0a that converts a less-than relation to a more-than relation. Once the model sees "Robin has 6 fewer beanie babies than Angie" as "Angie has 6 more beanie babies than Robin", it can solve the problem just as it would a more-than problem. Students' relatively good performance on the Cans and Beanie problems is consistent with a production rule transfer analysis (Singley & Anderson, 1989). Relative to the familiar two-quantity, more-than problems students were instructed on, only repetitions of already known

productions are needed for the Cans problem and only one new production is needed for the Beanie problem.

Production Traces of CD Solutions

It turns out the simple less-than to more-than transformation performed by P0a does not work so well for solving the CD problem. A production rule trace of the incorrect solution to the CD problem in Figure 1C illustrates this point. As in the Beanie trace, the model begins by firing P0a to convert "price of magazine is $8 less than the price of CD" to "price of CD is $8 more than the price of magazine". It then draws the single smaller box for the magazine cost (top of Figure 1C, but without 8 inside) and the equal-part and extra-part boxes for the CD cost (first two boxes in the CD row, without the 12 box or the first 8 label). The representation is correct at this point, but after converting "the price of poster is $12 less than the price of CD" relation to "the price of CD is $12 more than the price of poster" the model has trouble.

If the model only had the productions needed for the Cans and the Beanie problem, it would reach an impasse here. Usually the smaller base of the more-than statement (Poster in this case) has already been drawn and there are productions for drawing the equal- and extra-parts of the larger quantity (CD). However, in this case, it is the smaller box (Poster) that has not been drawn. Our model predicts that students reach an impasse at this point and must implement a "repair" (VanLehn, 1983). The production P3c in Table 2 represents a result of this repair. It draws a base box for the Poster cost that is the same size as the equal-part of the CD cost. Next, an over-general production adds an extra box to the CD row for the "12 more". This production is over-general because it is missing a constraint to check that an extra box has not already been drawn (the middle 8 box in the CD row). The model continues, like the student, to correctly compute values consistent with this incorrect representation of the problem.

With only one production difference from the Cans and Beanie problems this solution is a relatively easy transfer. This provides an explanation for the frequency of this odd incorrect solution strategy. 8 of 34 students drew pictures for the CD problem like the one shown in Figure 1C with two extra boxes in the CD row.

Students found two ways to be successful on the CD problem. Figure 1D shows an example of one of these ways. The model again starts by converting the given less-than relations to more-than relations, but not by changing the position of the quantities in the relation, but by negating the difference value (P0b in Table 2). For example, "price of magazine is $8 less than the price of CD" is converted into "price of magazine is -$8 more than the price of CD". The model then draws an abstract diagram with an extra box in the Magazine row that has a negative value. New productions are needed to deal with extra boxes with negative values. One of these productions removes an extra negative box by adding the absolute value of the box to the total (46 + 12 and + 8). A similar production is needed in the final steps of the solution to combine these negative values with the unknown value (22).

A second approach to the CD problem involves directly representing the less-than relations in the picture (see Figure 1B). In this approach, the larger CD cost remains the base quantity in both relations. That the magazine cost is $8 less is represented by labeling the space between the right end of the Magazine bar and the right end of the Poster bar. (It would be better if the student had not put a box around the 8 as this box implies the size of the Magazine cost includes this 8 when it does not.) Only one student on the CD problem drew a picture that was close to size preserving.

Both of the correct solution strategies for the CD problem involve more new productions (4 and 7) than the incorrect solution strategy (1). While the abstract picture strategy involves fewer new productions (4) than the size-preserving strategy (7), there are other relative advantages of the size-preserving strategy.

The new productions required for the size-preserving strategy are mostly straightforward analogies with existing productions, where less-than is substituted for more-than. More importantly, this strategy provides more reliable perceptual support for inferences. To get three equal bars, one must fill in the empty space to the right of the Magazine and Poster boxes (see Figure 1B). This "filling in" is a perceptual cue for addition. In contrast, in the abstract picture (see Figure 1D), the perceptual cues suggest "removing" or subtraction and, as we saw, many students fell for or were distracted by this cue. To succeed with this strategy, students have to do explicit symbolic processing to infer that taking away a negative is the same as adding.

An important consequence of this analysis was the recognition that students should be instructed *not* to draw boxes around extra missing parts (see the —8— and —12— in Figure 1C), but to use a double-headed arrow or bracket to mark the empty space.

Discussion

The Picture Algebra strategy benefits from a number of positive features of diagrammatic representations: grouping to facilitate search, reducing the need for symbolic labels, and substituting intuitive perceptual inferences for more difficult logical inferences (Larkin & Simon, 1987). The potential value of this strategy is evidenced by the observed success of 6th grade students on algebra-level problems that are quite difficult for many older students.

Teachers always encourage students to check their answers and students appear more likely do so when

using Picture Algebra (Fig 1B,C,D) than when using other strategies. We speculate that by evoking students' spatial intuitions, the pictorial representation puts students in a "sense making" mode that leads to greater self-monitoring (cf., Kalchman, Moss, & Case, 2001).

However, instruction based on pictures is not a panacea. Although pictures may facilitate students' reasoning and learning, it is not trivial for students to learn to use such representations flexibly and with understanding. Not all students using Picture Algebra engaged in sense making. Some made errors typical of equation solving (e.g., subtracting when they should add) and failed to catch them even with the visual support of the picture (cf., Lewis, 1989).

Although students performed relatively well on problem demands presented individually in the Cans and Beanie problems, they had much greater trouble with the CD problem that combined both demands. Our production rule analysis provides a detailed explanation for why this problem is significantly more difficult. The difficulty arises from the order in which the quantities are related to each other and potential traps from inappropriately selecting a "base" quantity.

A novel prediction of our model is that a 3-quantity more-than problem of the form "A is X more than B and A is Y more than C" should be almost as difficult as the CD problem (a 3-quantity less-than problem). This prediction follows from the fact that the model essentially converts the CD problem into this form. Similarly, the model predicts that a 3-quantity less-than problem of the form "A is X less than B and A is Y less than C" should actually be relatively easy, more like the Cans problem. The generation of such predictions has pushed us to rethink and improve our curriculum design to better address the previously hidden skills revealed by our production rule analysis.

This paper reports on theoretical and empirical steps toward answering the following research questions:

1. Can instruction on the use of Picture Algebra help younger students gain entry into algebraic reasoning sooner than direct instruction on formal algebra?
2. Can instruction on Picture Algebra help younger students build a foundation that will improve and accelerate later learning of formal algebra?

Because of the complexity and cost of performing classroom experiments that could directly address these questions, these questions are excellent candidates to test the applicability of cognitive theory. Our goal is to employ cognitive theory and lower cost empirical studies to provide strong arguments for and/or against these claims. With such arguments in hand, we are better prepared to assess the potential benefit of a costly classroom experiment and, perhaps more importantly, facilitate the design of an instructional method that is most likely to be successful.

Acknowledgments

This work was supported by NSF ROLE grant REC-0087396 and a curriculum development grant from Carnegie Learning, Inc. (www.carnegielearning.com).

References

Bednarz, N., & Janvier, B. (1996). Emergence and development of algebra as a problem-solving tool. In Bednarz, Carolyn, & Lee (Eds.), *Approaches to Algebra.* (pp. 115-145). Dordrecht, Netherlands: Kluwer.

Cheng, P. C.-H. (1999). Interactive law encoding diagrams for learning and instruction. *Learning and Instruction. 9*(4), 309-326.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the episodic structure of algebra story problem solving. Cognition and Instruction, 6, 223-283.

Heffernan, N.T. & Koedinger, K. R. (1998). A developmental model for algebra symbolization. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, (pp. 484-489). NJ: Erlbaum.

Kalchman, M., Moss, J. & Case, R. (2001). Psychological models for development of mathematical understanding. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 1-38). NJ: Erlbaum.

Koedinger, K. R. & Alibali, M. W. (1999). A developmental model of algebra problem solving. Annual meeting of the *American Educational Research Association.*

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

Lewis, A. B. (1989). Training students to represent arithmetic word problems. *Journal of Educational Psychology*, 81, 521-531.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics.* ISBN 0-87353-480-8. Reston, VA: NCTM.

Payne & Squibb (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14*, 445-491.

Rittle-Johnson, B. & Koedinger, K. R. (2001). Using cognitive models to guide instructional design: The case of fraction division. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society,* (pp. 857-862). NJ: Erlbaum.

Singapore Ministry of Education (1999). *Primary mathematics 6A Third Edition*. Singapore: Federal Publications. www.singaporemath.com.

Singley, M. K., & Anderson, J. R. (1989). *Transfer of Cognitive Skill*. Hillsdale, NJ: Erlbaum.

TIMSS (1996). Third International Mathematics and Science Study. National Research Coordinators: National Center for Educational Statistics.

Van Lehn, K. (1983). On the representation of procedure in repair theory. In H. P. Ginsburg (Eds.), *The development of mathematical thinking.* New York: Academic Press

Mutual Adaptive Meaning Acquisition by Paralanguage Information: Experimental Analysis of Communication Establishing Process

Takanori Komatsu (komatsu@cs.c.u-tokyo.ac.jp)
Kentaro Suzuki (suzuki@cs.c.u-tokyo.ac.jp)
Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Kazuo Hiraki (khiraki@idea.c.u-tokyo.ac.jp)
Department of System Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902 JAPAN
Natsuki Oka(oka@mrit.mei.co.jp)
Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd.
3-4 Hikaridai, Seika, Souraku, Kyoto 619-0237 JAPAN

Abstract

The effects of adjustments in teaching strategy and of paralanguage information in speech sound on the meaning acquisition process were determined by means of an experiment in which the game "Pong" was played by a team of two subjects. One subject (the teacher) coached the other one (the operator), instructing the operator in which direction to move the game paddle and when to hit the "ball." However, the teacher's speech was rendered linguistically incomprehensible. Three phenomena were observed. First, the use of a high-pitched voice by the teacher caused the operator to pay more attention to her/his actions. Second, meaning acquisition could be regarded as a reinforcement learning process based on a multi-reward system (i.e., one for successful game action and a different one from the teacher's high-pitched voice for the wrong action). Third, the subjects adapted to each other; that is, they learned to respond more appropriately to each other's behaviors (we call this **mutual adaptation**). These three phenomena are thought to play important roles in the acquisition of meaning from incomprehensible speech.

Introduction

How do people who speak different languages learn to communicate? How do they acquire the meanings of each other's speech? One way to interpret the meaning-acquisition process is to view it as statistical learning in which a certain speech sound is linked to the situation in which it was given. Several research groups have constructed general meaning-acquisition models based on this simple interpretation (for instance, Siskind, 1996). Testing of these models demonstrated that the word meanings can be acquired using statistical learning methodologies.

Some groups have investigated the teaching of word meanings to robot agents, which can move by themselves (Billard et al., 1998; Roy, 1999). For example, Kaplan (2000) showed that a four-legged robot could learn the names of a dozen objects shown in front of its "eyes" (camera) by the experimenter during its action experiences.

One study in particular demonstrated that an autonomous agent could learn not only the meanings of instruction words but also the meanings of evaluation instructions such as "good" and "bad". Suzuki et al. (2002) developed a learning agent model that could learn the meanings of words and evaluation instructions during its action experiences.

A large variety of learner models have thus been demonstrated. Siskind's learner model was a computer that processed string inputs; Kaplan's was a four-legged robot that could learn object names; and Suzuki's was a computer agent that could move by itself in a virtual environment. In contrast, only one type of "teacher" has been demonstrated, one who gave instruction based on unchangeable rules. In real life, however, good teachers adjust their teaching strategy to fit the learner's mode of learning. For example, a caregiver will speak to a preverbal infant using only simple words. Then, when the infant starts to speak, the caregiver will start to use more complex words and speak in different ways. So far, there have been no studies on how dynamically adjusting the teaching strategy affects the meaning acquisition process.

Investigating the process involved in acquiring meanings from speech sounds requires investigating the effects of not only phoneme information, which can be expressed using characters and text, but also of paralanguage information, such as prosody, speech speed and loudness, which cannot be expressed using characters and text. The effects of paralanguages information have been studied in various acoustic studies. For example, many studies have focused on the turn-taking mechanism (Pirrehumbert & Hirschberg, 1990) and emotional recognition (Hirose et al., 1997). However, so far, there have been none on the effect of paralanguage information on the meaning acquisition process.

We investigated the effects of adjustments in teaching strategy and of paralanguage information in speech sounds on the meaning acquisition process. We carried out a communication experiment, in which a team of two subjects played "Pong". One subject (the teacher) coached the other one (the operator), instructing her/him in which direction to move the game paddle and when to hit the "ball." However, the teacher's speech was rendered linguistically incomprehensible. We observed how the listener acquired the meanings of the given instructions. For the paralanguage information, we focused on prosodic information, because data extraction should be easier and the possibility of the engineering realization is higher. The results provide a new point of view about meaning acquisition process and can be used to develop

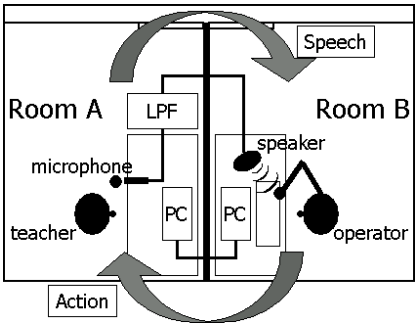


Figure 1: Game Environment

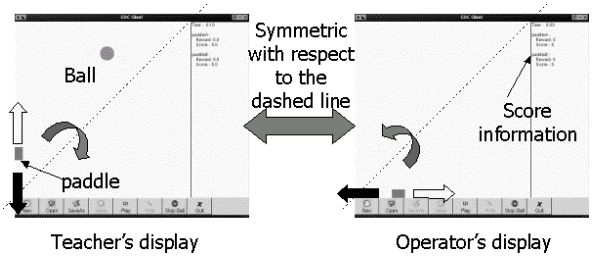


Figure 2: Different Display Settings

basic technologies for constructing interactive robots that can understand what is said to them, enabling them to communicate smoothly with people.

Experiment

Method

The experimental game environment is depicted in Figure 1. The two subjects were placed in separate rooms: the teacher in Room A, and the operator in Room B. The teacher's spoken instructions to the operator were transmitted through a low-pass-filter (LPF) and played over a speaker in Room B. The LPF masked the phoneme information, which included the symbolic elements of speech sound; it did not affect the prosodic elements. This means that the operator could hear only the prosodic information; s/he could not understand the teacher's instructions as meaningful linguistic information. They were simply noisy sound.

The team playing the game was awarded 10 points each time the operator hit the ball, and 10 points were deducted each time s/he missed it. As shown in Figure 2, the game display of the teacher differed from that of the operator: the teacher's display showed the ball, whereas the operator's did not. The operator could see only her/his paddle and the score information, indicating whether s/he had hit or missed the ball.

Even thought the instructions were masked by the LPF, the operator still might be able to guess which instruction was given based on the number of phonemes heard. For example, "right" and "left" in Japanese are *migi* and *hidari*, which have different numbers of phonemes. The display of the teacher and operator were thus made symmetrical with respect to the dashed line in Figure 2 to prevent biasing of the results.

Subjects

Eleven teams, each composed of two subjects (Japanese, 20-28 years old, 18 men and 4 women), participated in the experiment. The members of each team were required to know each other. One additional team participated in two control experiments: one without LPF and one without instructions.

Procedure

First, the experimenter explained to the subject team (one operator and one teacher) that the purpose of the experiment was to score as high a total score as possible by working together. Then, the experimenter let the team hear a vocal sample, with and without LPF masking, to demonstrate the filter's effects. The experimenter did not mention the differing displays.

The test was then started. Each team played two consecutive 10 minutes games, with 3 minutes of rest between them. The roles of teacher and operator were fixed, and the team members did not have an opportunity to talk face to face and share information.

Results

To evaluate team performance, values were assigned to two types of actions, moving the paddle and hitting the ball. For each move action, if the operator moved the paddle in the teacher's intended direction, **Correct Direction Value (CDV)** was assigned one point; if s/he moved it in a different direction, CDV was assigned zero point. For each hit action, if the operator hit the ball, **Hit Value (HV)** was assigned one point; if s/he missed it, HV was assigned zero points. Testing statistical hypothesis using binominal distribution was used to group the subjects, and actually the teams were divided into three groups as follows.

Group 1 Average CDV less than 0.8 point.

Group 2 Average CDV more than 0.8 point; Average HV less than 0.7 point.

Group 3 Average CDV more than 0.8 point; Average HV more than 0.7 point.

Table 1 shows the average values of the last ten hit and move actions for the three groups. Out of the 11 teams, two failed to understand any instructions (Group 1). Among the nine remaining teams, five succeeded in moving the paddle in the direction the teacher intended but could not hit the ball well (Group 2). We call this achievement "learning to recognize direction instructions." The four remaining teams could move to the teacher's intended position. We call this achievement "leaning to recognize distance instructions"(Group 3).

Table 1: Correct Direction Value (CDV) and Hit Value (HV)

Group	(CDV, HV)
Group 1 (2 pairs)	(0.5, 0.5), (0.3, 0.2)
Group 2 (5 pairs)	(0.9, 0.3), (1.0, 0.2), (1.0, 0.5) (0.8, 0.6), (1.0, 0.6)
Group 3 (4 pairs)	(1.0, 0.9), (1.0, 0.7) (1.0, 0.7), (0.9, 0.8)
*control	no instructions (0.5, 0.2), (0.4, 0.3) without LPF (1.0, 0.7), (1.0, 0.8)

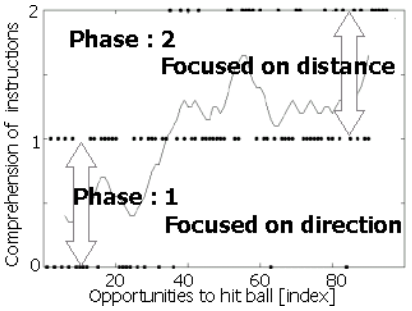


Figure 3: Trend in Comprehension of Instructions for Typical Group 3 Team

Table 1 shows that the values of two teams in Group 1 were the same as those for the control pair when they did not received instructions. Most of the teams in Group 2 and Group 3 had an average score between that of the two control settings, and some teams in Group 3 scored as high as the control group without the LPF.

We used **Comprehension of Instructions Value** to evaluate the operator's comprehension of the instruction. This value was assigned 0 points if the paddle was moved in the wrong direction, 1 point if it was moved in correct direction but the ball was not hit, and 2 points if it was moved in the correct direction and the ball was hit. The value for a typical team in Group 3 are plotted in Figure 3. The curve shows the moving average of the ten hit opportunities. When the curve was between 0 and 1 point, the operator is thought to have focused on understanding the direction instructions; when it was between 1 and 2 points, the operator is thought to have focused on understanding the distance instructions. The operator more quickly learned to comprehend the direction instructions than the distance ones. We investigate the processes of these two learning phases in terms of the prosodic elements of the instructions, the movement of the paddle, and the variety of instructions given. We focused on the pitch of the teacher's voice, which, among the prosodic elements, had the strongest relationship with the meaning of the instructions. As shown in Figure 6, when the pitch was increased, the operator tended to suddenly changed her/his action.

Phase 1: Focused on Direction

The operator on nine of the 11 teams learned to comprehend the direction instructions. A particular teaching/learning process was used by most of the nine teams. First, the teacher gave a wide variety of instructions to the operator, e.g., "move to the center of the display", "move a bit right", etc. The pitch curve of this initial teaching is depicted on the left side of Figure 4. It would have been difficult for the operator to discriminate between the "*ue, ue, ue...* (up, up, up...)" instructions for the right direction and "*shita, shita...* (down, down...)" instructions for the left.

As the teaching continued, the teacher gradually decreased the variety of instructions, typically converging on only two, such as *ue* and *shita*. In contrast, the teachers on the unsuccessful teams did not decrease the variety and continued to use a wide variety (Figure 5). Decreasing the variety of instructions apparently makes it easier to understand direction instructions.

In conjunction with this gradual decrease in the variety of instructions, the difference in the pitch curves of these two instructions became more distinct. Most teachers started to repeat the instructions, e.g., "*ueueue...*" and "*shitashita...*" The former was audible as one long, continuous voice, while the latter sounded like many choppy utterances (see the right side of Figure 4).

Here, the operator learned to recognize the type of instructions according to the differed sounds. At the same time, subjects were ready to make her/his paddle actions correspond to them. The operator then started exploring this correspondence by trial and error. If the operator succeeded in hitting the ball, even by chance, s/he learned to associate the given instruction with her/his last action. The teacher elicited a correct response by using higher vocal tones (depicted in the circle in Figure 6). When the pitch at a certain point in a series of teaching voices was about 20 [Hz] higher than the pitch at the onset, and when this higher pitch continued for at least 500 [msec], the operators intuitively recognized this as a warning from the teacher. Specifically, when the operator heard instructions delivered at a higher pitch, s/he recognized that her/his current action was wrong and modified it accordingly. Therefore, high-pitched vocalizations in Phase 1 served as a negative reinforcement of the operator's last action. The operators thus learned to compre-

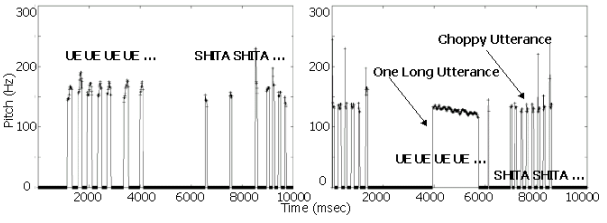


Figure 4: Pitch Curves of "*Ue*" and "*Shita*" in Phase 1. Left: 250 sec, Right: 540 sec

Figure 5: Variety of Instructions used by Successful/Unsuccessful Teams

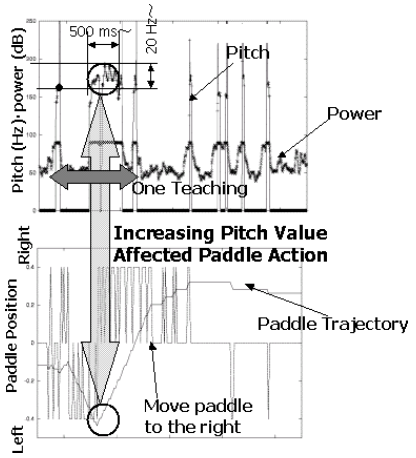


Figure 6: High-Pitch Element in Teaching Voice

hend the teacher's instructions by reinforcement learning based on a positive reward (hitting the ball) and a negative one (hearing a high-pitched voice).

Learning to recognize the instructions during Phase 1 was due not only to the teacher's efforts but also to the operator's action. The actions of operators on the successful teams were initially reluctant, concentrated on inferring the teachers' intentions from the given instructions. The operators started reacting actively after they inferred the intentions. Their actions thus indicated their comprehension of the given instructions. The operators on the unsuccessful teams were also reluctant at the beginning of the experiment, but over time they started moving actively, even though no instructions had been given: the operators seemed to disregard the instructions. In this case, the operator's actions did not indicate any comprehension of the instructions.

Although at a first glance it seems that meaning acquisition was achieved only by the operator adapting to the teacher's instructions, the teacher actually changed the instructions and method of delivery based on the operator's comprehensions, which was judged from the operator's reaction. The subjects actually learned to respond more appropriately to each other's behavior, so that meaning acquisition was a mutual adaptation process.

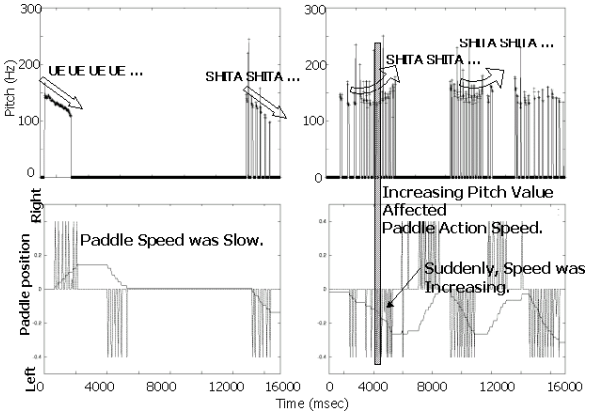


Figure 7: Pitch Curve of "*Ue*" and "*Shita*" in Phase 2. Left: 1070 sec, Right: 1270 sec

Phase 2: Focused on Distance

In this "Pong" game experiment, learning to recognize the direction instruction was not sufficient for achieving a high score. The distance information was also needed to consistently hit the ball with the paddle. Of the nine teams that learned to recognize the direction instructions, four were able to recognize the distance instructions. These four teams exhibited two distinct approaches to learning to recognize distance instructions. One team used the "stop" instruction, and three teams did not.

"Stop" Instruction Used One team developed a common understanding of the "stop" instruction before developing one for direction instructions. The sound of "stop" has a short, skipping sound, so it was not difficult for the operator to understand this literally as "stop". This team then developed a common understanding of direction instruction, like the repetitive "*ueueue...*" for move right and "*shitashita...*" for move left. In the case of using of "*ue*", the teacher had to give repetitive instructions; otherwise the independent usage of "*ue*" was recognized as a choppy sound, meaning the opposite (left, "*hidari*") direction. Therefore, the instruction to move right had to be repetitively elongated to avoid confusion with the usage of independent "*ue*". Otherwise, the operator tended to go past the teacher's intended position. When this "overrun" problem occurred, the teacher started saying "*ueueue, stop, shitashita*" to bring the operator back to the intended position. Finally, when the teacher said "stop" after a repetitive "*ueueue...*", the operator immediately reversed direction so as not to overrun the teacher's intended position. Thus, the operator took the "stop" instruction as the literal meaning or as meaning "go left". This is a typical example of mutual adaptation in this experiment.

"Stop" Instruction Not Used The three teams that did not use the "stop" instruction learned to recognize direction instructions in a manner similar to that of the other team. In a typical case, after teaching the operator to learn to recognize direction instructions, the teacher pitched her/his voice higher at the onset of utterance and lower at the end, i.e., the intonation decreased, as shown on the left side of Figure 7. When the distance from the paddle appeared to the teacher to be too short, the teacher gradually increased the pitch at the end of the utterance. Specifically, the teacher increased the pitch at the end of the instructions utterance when a long distance appeared to be needed (see the right side of Figure 7), whereas s/he reduced it when the distance appeared to be short. In this way, the teacher controlled the operator's action by increasing or decreasing the pitch at the end of the instruction utterance. High-pitched utterances spurred the operator's actions, which can be broadly interpreted as drawing the operator's attention to her/his actions.

Even among the nine teams who learned to recognize the direction instructions, five were unable to learn to recognize the distance instructions. The teachers on the successful teams came to realize that repetitive instructions were important to learning to recognize the direction instructions. They did this by observing the operators' actions. The teachers on the unsuccessful teams apparently did not came to this realization and were thus perplexed when the operators headed in the opposite direction when an "*ue*" instruction was given.

Comparison of the successful teams with the unsuccessful ones in the two meaning acquisition phases showed that the team members had to learn to respond appropriately to one another's behaviors to acquire a high score. The series of dynamic behaviors shown by the successful teams can be regarded as **mutual adaptation**.

Three points in particular were observed for the successful teams.

1. The use of a high-pitched voice by the teacher caused the operator to immediately focus on her/his action.

2. The operator learned to recognize the teacher's instructions by reinforcement learning composed of a positive reward (hitting the ball) and a negative one (hearing a high-pitched voice).

3. During the meaning acquisition process, mutual adaptation was observed. That is, team members learned to respond more appropriately to each other's behaviors.

Thus, paralanguage information functioned as a negative reward in the reinforcement learning process, and the teacher's adjustment in teaching strategy was observed as a mutual adaptation process. Paralanguage information and adjustment in teaching strategy thus play important roles in learning to recognize unknown teaching utterances.

Discussion

Effects of Adjusting Teaching Strategy

In this experiment, we observed a mutual adaptation process: the two members (the teacher and operator) learned to respond more appropriately to each other's behaviors. This study did not, however, consider how substantial differences between dynamic and static teaching strategies affect the meaning acquisition process. The actual effects of the teacher's dynamic instructions on meaning acquisition are thus unclear. We plan to investigate this effect by carrying out additional experiments in which we will control the behavior of the teachers. In any case, we observed that the teachers on the successful teams adjusted their teaching strategy according to the operator's comprehension of the given instructions. This observed behavior differs from one-directional, or fragmented communication aspect, such as that in the code model of Shannon & Weaver (1949). Wiener et al. (1972) argued that a symbol exchange model, which is a code model, is a primitive model of communication. Sperber and Wilson (1986) refuted this argument and argued instead that their "relevance theory" is a primitive model of communication. Although this theory might approach the essence of human communications and thus overcome the bottleneck of the code model: inferences in this theory can only be made using simple "deductive" inference rules (Kimura, 1997). Therefore, this theory is equivalent to the code model if we regard the inference "rules" as a complicated "code". This theory, moreover, has another problem: technical terms in the theory are not connected through physical existence. These unsolved problems reduce the theory's ability to explain actual communications.

In our experiment, it seemed at first that only the operator adapted to the teacher's behavior, because only the teacher could give spoken instructions. However, during the process of establishing communication, we observed that not only did the operator adapt to the teacher's behavior, the teacher adapted to the operator's behavior. While it is not uncommon for a teacher, who is a usually an information transmitter, to sometimes become an information receiver, in our experiment, the transmitting and receiving occurred simultaneously. This observed phenomenon cannot be explained by the code model. Thus, the results of our experiment suggest that the actual communication process cannot be expressed using a one-directional communication view like the code model. Instead, a mutual adaptive view that cannot be decoded into one-directional relationships is needed.

Effects of Paralanguage Information

In this experiment, we observed that the use of a high-pitched voice by the teacher caught the operator's attention. In general, the pitch of a speech sound decreases as the utterance draws to the end. If the pitch deviates from this pattern, the listener immediately catches the change. The observed function of the high-pitched voice in our experiment can be explained by this "prominence mech-

anism." Each team, regardless whether it was successful or not, actually made use of this function, which might be a universal trait among humans.

An infants' salient attention to a motherese voice, as revealed in various cognitive development studies, suggests that sensitivity to prosodic information is either an innate function or is learned just after birth. From our results, it might be said that infants react to the prosodic information in the caregiver's speech, not to the phoneme information. The meaning acquisition process observed in our experiment can be regarded as reinforcement learning with multiple-rewards, and the high-pitched voice among the various types of prosodic information functioned as one of two rewards. That is, prosodic information affects the meaning acquisition process at some basic levels. Thus, it can be assumed that the infants' language acquisition process and the meaning acquisition process in our experiment are similar in both depend on receiving prosodic information before acquiring the actual meanings of speech or teachings. Further research should reveal how this assumption is related to the language acquisition processes.

Application Based on Results

From the results of our experiment, we conclude that paralanguage information and adjustments in teaching strategy play important roles in learning the meaning of unknown utterance. If an autonomous robot had such abilities, it should be able to learn the meanings of human utterances and thus be able to communicate with people smoothly. In the traditional method of developing an interactive robot system, the designer needs to map specific utterance to the robot's functions. However, by applying the phenomena observed in our experiment, the designer need not to define such a mapping *a priori*, but simply needs to define "high-pitched voice = negative reward" and "successful action = positive reward". From these definitions, the robot can learn by itself the mapping between unknown utterances and its possible actions by using a reinforcement learning process based on the two types of rewards. Consequently, the robot will be able to adapt to its owner by trial and error, and over time the robot and owner can create an intimate relationship.

Conclusions

A communication experiment was carried out to observe the effects of a teacher's adjustment in teaching strategy and of paralanguage information in terms of learning the meanings of unknown utterances. An adjustment in teaching strategy was observed as a mutual adaptation process, in which the two subjects on a team learned to respond more appropriately to each other's behavior, and paralanguage information was observed to function as a negative reward in a reinforcement learning process for meaning acquisition. These two phenomena thus play important roles in learning the meaning of unknown utterances.

References

Billard, A., Dautenhahn, K., & Hayes, G. (1998). Experiments on human-robot communication with robota, and interactive learning and communicating doll robot. In B. Edmonds and K. Dautenhahn (Eds.), *Socially situated intelligence workshop (SAB98)* (pp. 4–16).

Hirose, K., Kawanami, H., & Ihara, N. (1997). Analysis of intonation in emotional speech. *Proceedings of ESCA tutorial and Research Workshop in Intonation: Theory, Models and Applications*.

Kaplan, F. (2000). Talking AIBO: First Experimentation of verbal interactions with an autonomous four-legged robot. In A. Nijohlt & D. Jokinen (Eds.), *Learning to behave: interacting agents CELE-TWENTE Workshop on Language Technology* (pp. 63–75).

Kimura, D. (1997). Information, regularity and communications – comparison between Shannon and Bateson. In Tani, Y (Eds.), *Communication no shizen-shi (in Japanese)*, Tokyo: Shin-yosha.

Pirrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, and M .E . Pollack (Eds.), *Intentions in Communication*, Cambridge, MA: MIT Press.

Shannon, C. & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Siskind, J. (1996). A computational study of cross–situational techniques for learning word to meaning mapping. *Cognition, 61*, 39–91.

Sperber, D. & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Suzuki, K., Ueda, K., & Hiraki, K. (2002). A Computational Model of an Instruction Learner: How to learn Good or Bad through Actions. *Cognitive Studies(In Japanese).*, in press.

Roy, D. (1999). *Learning from sights and sounds: A computational model*. Doctoral dissertation, MIT Media Laboratory.

Wiener, M., Davoe, S., Rubinow, S., & Geller, J. (1972). Nonverbal behavior and nonverbal communication. *Psychological Review, 79.*

Qualitative physics as a component in natural language semantics: A progress report

Sven E. Kuehne (skuehne@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
1890 Maple Avenue, Evanston, IL, 60201, USA

Kenneth D. Forbus (forbus@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
1890 Maple Avenue, Evanston, IL, 60201, USA

Abstract

We propose that qualitative physics can provide an important component of natural language semantics. Specifically, we describe how qualitative process theory can be recast in terms of frame semantics, as used in the Berkeley FrameNet project. This reformulation is important because it could allow the techniques of qualitative reasoning to be harnessed for natural language understanding and it expands the range of phenomena that can be described in NL semantics. We show that these ideas can account for a large percentage of a small corpus of explanatory text, and that they support the construction of QP models from such texts.

Introduction

Understanding the semantics of natural language is a central problem in cognitive science. Such an understanding must connect fundamentals of our conceptual structure to their realizations in linguistic forms, and thus must draw upon both insights about language and about conceptual structure. Significant progress is being made on the language side, with projects such as FrameNet (Fillmore et al 2001) developing broad systems that capture aspects of the meaning of words and linguistic constructions in terms of *frame semantics* (Fillmore & Atkins, 1994). Significant progress is also being made on understanding aspects of human conceptual structure, for example, the work carried out in the qualitative reasoning community. Qualitative reasoning focuses on the representations necessary to reason about the physical world, ranging from everyday phenomena to the work of scientists and engineers. While many QR efforts are aimed at applications, some efforts are aimed directly at modeling human reasoning about physical systems (cf. Bredeweg & Schut, 1991; Kuipers & Kassirer 1984; Kuipers et al 1988; Forbus & Gentner 1986, 1997). Ultimately these two lines of investigation, natural language semantics and the understanding of human conceptual structures, need to join forces. In the words of the FrameNet team[1]: "In the end it will be necessary to express frame notions in some formal knowledge-representation language which will allow valid inferences to be drawn from frame semantic representations of sentences, or which can serve in a precise way in the development of a cumulative representation of the content of an ongoing discourse."

This paper is a first step in linking these investigations. We propose that qualitative process theory (Forbus 1984) can provide such a knowledge representation language for aspects of frame semantics concerned with continuous parameters and continuous causation. QP theory concerns the structure of a class of physical theories, and has been successfully used in a variety of reasoning systems (Forbus, 1996). The hypothesis is that many mental models of physical phenomena can be expressed in this formalism. QP theory has been used to develop a wide range of models of phenomena, including economic and medical models in addition to physical models. This makes it an excellent candidate for a component in a larger system of natural language semantics.

We begin by building a theoretical bridge between QP theory and frame semantics, as exemplified in FrameNet. We briefly review the relevant aspects of QP theory and show how they can be recast in terms of frame semantics. This recasting provides a means for defining frames for physical processes and relationships involving continuous parameters. We illustrate how these ideas can be used to extend a natural language semantics by an analysis of flow.

With the theoretical bridge in place, we provide empirical support for it via a corpus analysis of an explanatory text. This analysis provides evidence concerning two questions. First, we explore how much of the explanation these frames can account for. Second, we analyze whether a qualitative model can be reconstructed from the text using these frames. Finally, we discuss some new issues raised by this approach and plans for future work.

Qualitative physics in frame semantics

We begin by recasting QP theory in terms of frame semantics, as used in FrameNet. We review the ontological assumptions underlying QP theory and their implications for the organization of the frame system. Next we discuss physical processes and their occurrences, followed by an analysis of how qualitative mathematics is expressed. We conclude this section by showing that this analysis is compatible with analyses of overlapping phenomena already in FrameNet.

[1] The FrameNet project's home is the International Computer Science Institute in Berkeley, CA. A detailed description of the project can be found at http://www.icsi.berkeley.edu/~framenet

Ontological assumptions

In QP theory, physical changes in continuous properties are caused by *physical processes*. Examples of physical processes include kinds of flows (e.g., heat, liquid, gas), phase changes (boiling, freezing), and some aspects of motion. Ontologically, physical processes serve as the mechanisms of physical causality: All naturally occurring changes (and many of the indirect effects of the actions of agents) are ultimately caused by the activity of one or more physical processes. Instances of physical processes exist when an appropriate configuration of *participants* occurs. Such process instances are *active* over any span of time for which their *conditions* hold. When a process instance is active, its *consequences* hold. For example, two thermal entities (i.e., having the continuous property *heat*) that are thermally in contact give rise to two instances of heat flow, one in each potential direction. Whether or not either of these is active depends in turn on the relative temperatures between the two bodies.

The consequences of a physical process are of three types. First, there are *direct influences* that represent the direct effects that a physical process has on the world. For example, heat flow causes the heat of the source of the flow to decrease while increasing the heat of the destination. Second, there are other dynamical properties defined, including new parameters and causal laws, which describe how changes propagate through continuous properties. For example, the rate at which heat flows is a continuous property, and it is determined by the difference between the temperatures. Third, other properties that hold while the process is occurring, such as appearance information, can be consequences. In everyday boiling, for instance, one typically sees bubbles.

Two key conceptual advances in qualitative modeling are the insights that (1) many important kinds of reasoning about dynamical systems can be done without numerical information or mathematical models and (2) qualitative relationships can be formulated which explicitly capture patterns of human causal reasoning. These causal connectives are summarized below; see (Forbus 1984) for details. The values of continuous parameters tend to be expressed in comparative terms, via ordinal relationships constraining a parameter with respect to other relevant properties. If an object participates in process instances of heat flow, for example, then its temperature is defined in terms of its relationships with the temperatures of the other objects participating in those heat flows.

This summary highlights three important properties of QP theory that makes it potentially a valuable component of natural language semantics. First, the notion of physical process it defines is psychologically plausible. Descriptions of physical processes are abundant in language concerning physical phenomena, and are routinely used in metaphors (cf. Lakeoff 1980, Gentner et al 2001). Second, the causal account QP theory provides is consistent with human causal explanations in most physical domains (Forbus & Gentner 1986, 1997). Third, the abstract level of information that qualitative representations support seems a natural fit for the level of specificity commonly found in natural language descriptions of physical principles and situations. One does not need to understand differential equations or carry out detailed simulations to understand physical metaphors ("her anger mounted until she boiled over").

Recall that in frame semantics, meaning is expressed in terms of systems of structured representations, *frames*, whose parts (called *frame elements*, abbreviated FE) are bound to parts of a text and have associated with them inferences that provide meaning (Fillmore & Atkins, 1994). The packaging of physical knowledge and principles in QP theory (inspired in part by Minsky's (1975) notion of frames) suggests a natural alignment with frame semantics. There is a basic *physical process* frame, whose structure provides the fundamental aspects of physical processes. Subframes describe particular categories of physical processes, with differences in their participants and consequences being the differentia that set them apart. Instances of these frames are combined with frames from other aspects of the semantics to create the frame system describing the meaning of a text. The qualitative causal mathematics of QP theory is expressed through another collection of frames. In addition to their role in physical process descriptions, these qualitative causal frames can be used for other domains with continuous parameters, such as economics or metaphorical extensions of physical concepts. The next three subsections outline these frame systems.

Processes and their occurrences

The `PhysicalProcess` frame involves four types of FEs:

- *Participant* specifies one of the participants in the physical process. Example: in "Heat flows from the hot brick to the cool room", "hot brick" and "cool room" are `Participants` in an instance of the HeatFlow frame.
- *Condition* specifies one of the conditions under which the process is active. Example: in "Heat flows from one place to another because the temperature of the two places is different." the `Condition` is the difference in temperature values (see ordinals below)
- *Status* specifies whether or not the process is active. Example: In "The radiator leak was stemmed by shoving a cloth into it." The word *stemmed* suggests that a flow which was enabled is now stopped. We say that the `Status` is *active* when the process is occurring, and *inactive* otherwise.
- *Consequence* specifies one of the direct consequences of the physical process. Example: In "Water flooded into the room when the valve broke." the liquid flow into the room has as one of its `Consequences` an increase in the amount of water in the room.

These frame elements can be directly mapped to the formal models that QP theory supports. For a process type or instance, the set of participants collectively define the collections of entities it occurs among. The union of the conditions is the set of conjuncts that comprise the necessary and sufficient conditions for it to be active. The

set of fillers for the consequences FEs constitute its direct consequences.

Our analysis of the syntactic realizations of these frame elements, and the others reported here, is work in progress, and we plan to analyze a much larger corpus to ensure that our results are robust. That, plus space limitations, will limit our discussion of syntactic realizations to a few stable highlights. Noun phrases that serve as the primary actor and object in a sentence tend to be participants, e.g., in "A hot brick loses heat to a cool room." "Hot brick" and "cool room" are participants. Certain frame elements already used in other FrameNet frames, e.g., `Source` and `Destination`, tend to be participants when a physical process is the actor. The patterns that indicate conditions include "`Condition` **causes** `Process`", "`Process` **occurs when** `Condition`.", and "`Process` **depends on** `Condition`." For consequences, there are two cases: influences and other consequences. Influences are discussed below. The other consequences, since they can range over almost any physical statement in principle (e.g., appearances, sounds, etc.), are difficult to characterize concisely. Example indicators are occurrences of the FrameNet FEs of `Manner` and `Result`.

Parameters and values

Continuous properties are represented by the `Quantity` frame, which has the following elements:

- *Entity* specifies what this property is a property of. Typically this is unique. Example: "brick" in "the temperature of the brick"
- *QuantityType* specifies the kind of parameter that this is. Example: "temperature" in "temperature of the brick."
- *Value* specifies the numerical value of the property. This FE is optional. Example: "3" in "3 liters of water".
- *Units* specifies the physical units of the property. This FE is optional. Example: "kilograms" in "3 kilograms of lead".

Ds specifies how the parameter is changing and stands for "sign of the derivative". This FE is optional. Example: In "The temperature is increasing." the sign is expressed by the word "increasing" which would be mapped to the value of 1. While syntactic realizations for quantity types, values and units are fairly obvious, `Ds` manifests itself in the text many different ways, e.g. *-1* could show up as "falling", "decreasing", etc.

Values and units are often not explicitly stated or even filled in via default, but `Ds` and comparative statements about values are common. These are expressed via the `Ordinal` frame, which has the following FEs:

- *Q1, Q2* specify the quantities being compared. Either is optional. Example: "the coffee's temperature".
- *OrdReln* specifies the relationship between the values of the quantities. It must be one of $<, >, =, \geq, \leq, =. \neq$, *same-order*, or *negligible*. Example: In "Evaporation can be ignored" the word "ignored" refers to a negligible `OrdReln` of the rate of an evaporation process compared to other processes being described.

Ordinal relations provide a useful qualitative notion of value because they often serve as conditions for physical processes and states (e.g., flows occur when a driving parameter is unequal, equilibriums occur when opposing effects are equal). Syntactic realizations of `Ordinals` are usually described via explicit comparisons (e.g., "`Q1` **is greater than** `Q2`") or as some type of comparative construction. One very common pattern is the use of ordered dimensional adjectives to set up a tacit comparison. For instance, from "hot brick" and "cool room", one knows an ordinal relationship involving their temperature due to the meanings of "hot" and "cool".

Qualitative mathematics and causality

The causal relationships between quantities are expressed via a qualitative mathematics that supports partial information about the nature of the connections between them. The basic frame is the `Influence`, whose FEs are

- *Constrained* specifies the dependent quantity, i.e., the effect.
- *Constrainer* specifies the independent quantity, i.e., a proximal cause for the constrained quantity.
- *Sign* specifies the direction, which can be + or -. It is expressed by words such as "up", "down", "greater", "more", "less" etc.

There are two subframes of the `Influence` frame, `DirectInfluence` and `Qprop`. These correspond to the QP theory primitives I+/I- and $\propto_{Q+}/\propto_{Q-}$ respectively (Forbus 1984). While the two subframes share frame elements, the underlying semantics is quite different. For direct influences, the constrainer is combined via addition to other constrainers to determine (qualitatively) the derivative of the constrained quantity, and the sign indicates whether it is a positive or negative contribution to that sum. For `Qprop`, the `Constrained` is functionally dependent on the `Constrainer`, and perhaps on other properties as well, with the sign indicating whether the dependence is increasing or decreasing monotonic. This is the weakest distinction that enables changes to be propagated through causal laws.

As their common heritage suggests, in some cases the syntactic realizations of these two kinds of influences can be quite close. However, many cases are straightforward. Some realizations for `Qprop` include "`Constrained` **depends on** `Constrainer`." and "**As** `Constrainer Ds`, `Constrained Ds`." For example, "As the air temperature goes up, the relative humidity goes down" is clearly a `Qprop`, with `Constrained` = "relative humidity", `Constrainer` = "air temperature", and `Sign` = -.

Syntactic realizations for `DirectInfluences` are more complex. In advanced texts one can find patterns such as "**The rate of** `Constrained` **depends on** `Constrainer`." but they do not seem common. In everyday texts explicit discussions of rates seem even rarer. Instead, `DirectInfluences` tend to occur in larger-scale patterns, often tied to a generalized notion of motion. For example, "Most water in the air comes from evaporation." is a

DirectInfluence, with Constrained = "water in the air", Constrainer = "[rate of] evaporation" and Sign = +. [2]

Compatibility with existing frame semantics

One implication for semantics is that, in addition to the frames associated with QP theory per se, there will be a collection of subframes corresponding to particular kinds of physical phenomena, such as flows, motion, and phase changes. And indeed FrameNet already has an existing analysis of motion that is compatible with QP theory. The FEs of Theme, Source, Goal, and Path are finer-grained distinctions of the FE Participant in the general PhysicalProcess frame. In QP theory models of this kind of motion (cf. Forbus 1984), there is a quantity Position that is referenced to the Path from Source to Goal. A DirectInfluence frame with Constrained = Position and Constrainer = Velocity is a Consequence of the Motion frame. An Ordinal frame with Q1 = Velocity, Q2 = zero, and OrdReln = ≠ is the Condition for the Motion frame. This compatibility is encouraging, since it means that the implications that can be drawn from qualitative reasoning could be made available in service of natural language understanding.

Example: An analysis of flow

Next we present an analysis of an important frame for physical phenomena, flow. We start with well-worked out ideas in the qualitative physics literature, using the framework above to recast them into frame semantics. This frame is used in our corpus analysis below.

The general Flow frame

The model of flow we are starting with is based on those in Forbus (1984). Several of the frame elements are specializations of Participant:

- *FlowSource* specifies the starting region of the flow.
- *FlowDest* specifies the region where what is flowing ends up.
- *FlowPath* specifies the path along which the flow occurs

These FEs determine the overall type of flow occurring:

- *FlowDriver* specifies the intensive quantity (e.g., something like pressure or temperature) whose difference at source and destination drives the flow.
- *FlowQ* specifies the extensive quantity (e.g., something like mass or heat) that is directly influenced by the flow. Optional.
- *FlowStuff* specifies the "stuff" which is considered to be flowing. Optional.

Typically texts mention either FlowQ or FlowStuff but not both. Many uses of the Flow frame are metaphorical from a scientific perspective (e.g., heat is not a substance), but may be literal from a common sense perspective, depending on the language user's mental models. FlowStuff must be continuous in nature (hence FlowQ must exist, even if not explicitly mentioned) for the idea of flow to make sense.

There are two Condition FEs for Flow. The first is an Ordinal, i.e., that the FlowQ(FlowSource(Flow)) is greater than FlowQ(FlowDest(Flow)). The second is that FlowPath not be blocked. The nature of being blocked depends on the subframe of Flow. For instance, a stopper can block liquid flow, but heat can still pass.

Flow has three Consequence FEs: a Quantity whose QuantityType is Rate, and two DirectInfluence frames, constraining the FlowQ of FlowSource and FlowDest via Rate(Flow) with the appropriate signs.

More sophisticated versions of this basic pattern are common in qualitative modeling; e.g., the rate is typically a function of the difference between the driving quantities, and also depends on path properties. Such elaborations do commonly appear in explanatory texts, and consequently the ability of qualitative modeling to support such incremental elaboration, which can be done via additional Qprop frames in this case, provides a necessary source of representational flexibility for natural language semantics.

Some empirical evidence

Up to now we have been concerned with expressing the concepts of QP theory in Frame Semantics, showing how they can fit into this larger system and some of their syntactic realizations in English. Here we examine the utility of doing this, in two ways. First, we turn the question around: How much do these ideas contribute to understanding the semantics of texts involving the physical world? Second, can we use these ideas to reconstruct from an explanatory text the physical ideas being communicated? We examine each in turn.

How far can a QP-based frame semantics go in accounting for the semantics of explanatory texts? One way to answer this question is to analyze a corpus of physical explanations, and see what fraction of the sentences require the frames of QP theory (and frames for mental models expressed in QP theory) for their interpretation. We have done this by using four chapters of a book on solar energy, *Sun Up to Sun Down* (Buckley, 1979). We chose this book because it is very clearly written, and we have been using it for a source of examples in other projects, since it uses both diagrams and analogies heavily.[3] We chose chapters 2 through 5 because they provide a basic exposition of heat, temperature, and types of heat flow.

Our analysis method was this. Two evaluators familiar with the theory independently scored each sentence. Then

[2] Note that 'from evaporation' refers to an internal quantity of a process (i.e. the evaporation rate), not to a participant (as in 'from the ocean'). The latter would be marked as a source, not as a constrainer.

[3] As additional corpus material, we have selected a college textbook on the weather as well as a children's book on weather, but while these have served as a source of data for our analysis of syntactic realizations, the results in this section are based only on the Buckley text.

they compared their results, discussing divergences until they came to agreement.

We looked at the linguistic realizations of physical processes in the text. Based on the QP frame semantics, we defined nine types of information about processes: the process name (P), information about subclasses of a process (i.e. a specialization) (SC), participants (PA), about conditions: antecedent activations (AA), antecedent ordinal relations (AO), antecedent relations (AR), and finally about consequences: indirect influences (CII), direct influences (CDI), and consequence relations other than influences (CR).[4] Multiple pieces of information can appear within a single sentence, so we scored number of phrases of particular types in addition to the number of sentences that they occurred in. Sentences can contain multiple types of information, so the same sentence can appear in multiple categories. We also distinguished between information from examples (identified through a preliminary analysis) and general information, since we have hypothesized (Forbus & Gentner, 1997) that common sense physics arises from within-domain analogies involving concrete descriptions. Tables 1 (general information) and 2 (exemplar-specific information) show our results.

Type	P	SC	PA	AA	AO	AR	CR	CDI	CII
#Sentences	10	1	8	15	5	1	9	8	15
#Phrases	11	4	14	16	5	1	18	16	18

Table 1: General statements using QP theory concepts

Type	P	SC	PA	AA	AO	AR	CR	CDI	CII
#Sentences	26	0	28	15	6	5	26	19	14
#Phrases	26	0	74	15	7	5	53	38	17

Table 2: Use of QP theory concepts in examples.

The data shows that the exemplar-specific data contains more than twice the number of processes, about five times the number of participants and a lot more information about the consequences of the mentioned processes. However, the amount of information about the conditions of a process (categories AA, AO, and AR) is nearly the same. As expected, any information about specialization of processes (SC) is only found in the general information.

What kind of coverage does QP theory provide? Of the 216 sentences, 94 of them mention at least one element from the QP frame system proposed here. That means that QP theory can account for roughly 43% of these chapters.

Let us turn to the second question, the reconstruction of a QP domain theory from the frame semantics that one might get from analyzing a text. Again we rely on Chapters 2 through 5 of Buckley (1979). These chapters yield six physical processes: General models of heat flow and volume flow (e.g., liquid flow), and four subclasses of heat flow (conduction, convection, radiation, and transport). Using this data and the information about the constituents of processes it contains, we attempted to manually reconstruct the models of the underlying physical processes. Figure 1 shows the reconstructed model of the generic heat flow process. One piece of information in Figure 1 marked with a star was not part of any general description of the heat flow process but originated from information about specific examples. It was generalized and included into the generic process model. By combining information from specific examples with general information, reasonable QP theory process descriptions of each were obtained.

These results are very encouraging. The frame semantics based on QP theory provides significant coverage of this corpus. The aspects that are not related to QP theory are not themselves physical laws or behaviors per se, but require frames of the types that would be found in other kinds of texts. Thus these results suggest that the frame semantics we propose using QP theory could play a useful role in a broad system of natural language semantics**.**

Discussion

This paper argues that qualitative physics, specifically QP theory, can be used as a component in a system of natural language semantics. We outlined how QP theory can be recast in terms of Fillmore's frame semantics, as used in the FrameNet project. The constructs of QP theory can be recast in terms of a collection of frames and subframes, which can be used to describe many causal mental models found in explanatory texts. As the syntactic realizations of these frames are further worked out, we believe that they will be a valuable extension to FrameNet semantics.

In addition to broadening the coverage of FrameNet to include a wide range of continuous phenomena and systems, our extension also grounds these new frames in terms of a well-worked out knowledge representation formalism capable of supporting qualitative reasoning. The compatibility of existing FrameNet motion descriptions with this model, and our analysis of a QP model of flow in frame semantics, lends support to our claim that this recasting of qualitative modeling can productively extend frame semantics for natural language.

The corpus analysis presented suggests that this extension can be useful, since 43% of the material in sample chapters from a typical science book can be captured in terms of them. Moreover, our analysis suggests that these frames could be composed to construct domain theories of a kind already used in qualitative reasoning.

Our results suggest that QP frame semantics can indeed play an important role in natural language semantics for physical texts. More investigation is needed on several questions, including:

- We want to refine our estimates of coverage by analyzing a larger corpus with a broader range of materials. These analyses are impractical by hand, so we are exploring the use of automated tools for subsequent analyses.

[4] Although the information involved in these categories has varying complexity, e.g., influences and ordinal relations are more complex than the process or participant names, we do not impose any ordering or weighting on these pieces of information.

```
Process HeatFlow
:participants
  loc1 ;; FlowSource
  loc2 ;; FlowDest
  path ;; FlowPath
:conditions
  (> (temp loc1) (temp loc2))
  *(heat-aligned path)
:consequences
  (from-location loc1)
  (to-location loc2)
  (qprop- flowrate thermal-resistance)
  (I- (heat loc1) flowrate)
  (I+ (heat loc2) flowrate)
```

Figure 1: One QP physical process description reconstructed from the Buckley text. The line marked with "*" was derived via generalization from specific examples.

- We view our work as complementary to that of Narayanan (1999), who is linking FrameNet semantics with sensory-motor schemata. Both will ultimately be needed, and their interplay will be interesting to explore.
- The same QP analysis used for literal language could be used to improve the productive understanding of many metaphors. For example, the FrameNet analysis of `heat` in the emotional domain has tied to it the lemma **boil** directly. If QP representations for heat, heat flow, and boiling were used instead, one could infer that making someone angry for longer could lead to boiling, and that if someone had "boiled over", starting a heat flow with them as the `Source`, could "cool them off".
- A fascinating set of questions arises from cross-linguistic comparisons. Are these ideas bundled up in the same way in all languages, or are they realized very differently (e.g., Bowerman's (1996) cross-linguistic analysis of spatial prepositions, Talmy's (1985) cross-linguistic analysis of verb semantics, and Imai and Gentner's (1993) analysis of the mass/count distinction)?

Another goal of our analysis is synthesis, i.e., to create a habitable controlled language that can be used in natural language processing systems that communicate with people about the physical world more fluently. Such software could be invaluable in creating new kinds of intelligent software, such as tutoring systems and monitoring systems.

Acknowledgments

We thank Jason Trost for help with the corpus analysis, and Larry Birnbaum and Dedre Gentner for insightful comments on the paper. This work was supported by the Artificial Intelligence program of the Office of Naval Research.

References

Bowerman, M. (1996). *Learning how to structure space for language: A cross-linguistic perspective.* In P. Bloom, M.A. Peterson, L. Hadel, & M.F. Garrett (eds.), *Language and space*. Cambridge, MA: MIT Press, 385-436.

Bredeweg, B. & Schut, C. (1991). *Cognitive plausibility of a conceptual framework for modeling problem solving expertise*. Proceedings of the 13th Annual Conference of the Cognitive Science Society. Hillsdale, New Jersey: Lawrence Erlbaum, 473-479.

Buckley, S. (1979). *Sun up to sun down*. New York: McGraw-Hill.

Falkenhainer, B. & Forbus, K. (1991). *Compositional Modeling: Finding the Right Model for the Job*, Artificial Intelligence, **51** (1-3).

Fillmore, C. J. & Atkins, B. T. S. (1994). *Starting where the dictionaries stop: The challenge for computational lexicography*, In Atkins, B. T. S. and A. Zampolli (eds.) Computational Approaches to the Lexicon. Clarendon Press.

Fillmore, C. J., Wooters, C. & Baker, C. F. (2001). *Building a Large Lexical Databank Which Provides Deep Semantics*. Proceedings of the Pacific Asian Conference on Language, Information and Computation. Hong Kong.

Forbus, K. (1984). *Qualitative Process Theory* Artificial Intelligence, (24): 85-168.

Forbus, K. (1996). *Qualitative Reasoning*. CRC Handbook of Computer Science and Engineering. CRC Press.

Forbus, K. & Gentner, D. (1986). *Causal reasoning about quantities*, Proceedings of the Eighth annual conference of the Cognitive Science Society, Amherst, Mass., August, 1986.

Forbus, K. & Gentner, D. (1997). *Qualitative mental models: Simulations or memories?* Proceedings of the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). *Metaphor is like analogy*. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (eds.) The analogical mind: Perspectives from cognitive science. Cambridge, MA: MIT Press, 199-253

Imai, M., & Gentner, D. (1993). *Linguistic relativity vs. universal ontology: Cross-linguistic studies of the object/substance distinction.* Proceedings of the Chicago Linguistic Society.

Kuipers, B. J. & Kassirer, J. P. (1984). *Causal reasoning in medicine: analysis of a protocol.* Cognitive Science 8: 363-385.

Kuipers, B. J., Moskowitz, A. J. & Kassirer, J. P. (1988). *Critical decisions under uncertainty: representation and structure.* Cognitive Science **12**: 177-210, 1988.

Lakeoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Minsky, M. (1975). *A framework for representing knowledge*. In: P. Winston (ed.) The Psychology of Computer Vision. New York: McGraw Hill: 211-7.

Narayanan, S. (1999). *Moving Right Along: A Computational Model of Metaphoric Reasoning about Events.* Proceedings of the National Conference on Artificial Intelligence (AAAI '99), Orlando, Florida, July 18-22, 1999, AAAI Press: 121-128.

Talmy, L. (1985). *Lexicalization patterns: Semantic structure in lexical form*, In T. Shopen (ed.), Grammatical categories and the lexicon, Vol. 3 of Language typology and syntactic description, CUP, Cambridge

Learning Causal Structure

David A. Lagnado (David_Lagnado@Brown.Edu)
Department of Cognitive and Linguistic Sciences, P.O.Box 1978
Providence, RI 02912 USA

Steven Sloman (Steven_Sloman@Brown.Edu)
Department of Cognitive and Linguistic Sciences, P.O.Box 1978
Providence, RI 02912 USA

Abstract

The central aims of this experiment were to compare observational and interventional learning of a simple causal chain, and to ascertain whether people represent their interventions in accordance with the normative model proposed by Pearl (2000). In the observation condition people treated putative causes as independent, and systematically selected the wrong model. In the intervention condition performance improved, in particular greater sensitivity was shown to the relevant conditional independencies. However, participants' likelihood judgments approximated the observed frequencies rather than reflecting the appropriate causal model.

Introduction

Our causal knowledge of the world is closely tied to our ability to control or manipulate certain aspects of it. On the one hand, we often learn about cause-effect relations by observing the effects of our own interventions (e.g., running controlled experiments). On the other, we can exploit such knowledge by manipulating the causes appropriate to our desired ends. Further, our causal knowledge allows us to predict or imagine the consequences of our actions, and is thus a prerequisite for deliberative decision-making.

Given the central role that intervention plays in causal reasoning, it has received scant attention in most accounts of human causal learning. In part this is due to the lack of a formal analysis of intervention, and the failure of standard probability theory to distinguish *action* from *observation* (Pearl, 2000). These lacunas appear to be addressed by a body of recent research in AI, which provides a normative analysis of causal inference and a formal means of representing the difference between observation and intervention (e.g., Glymour, 2001; Pearl, 2000).

The formulation of a normative model is at best only a first step towards an understanding of how people acquire and employ causal knowledge. The current experiment aims to gather some preliminary evidence about the difference between observational and interventional learning, and whether people represent their interventions in the manner suggested by this normative account.

Causal Models

The causal model framework offers a method for representing causal knowledge and formal rules for updating this knowledge in the light of either observation or intervention. Central to this formalism is the use of directed graphs to represent the mechanisms that underpin our causal knowledge of a domain, and the use of probability theory to reflect the uncertain and defeasible nature of this knowledge.

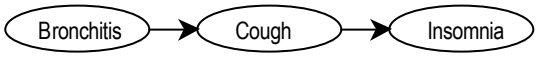


Figure 1: A causal graph

A causal model is made up of a set of nodes, a set of directed links between nodes, and a conditional probability distribution for each node. The nodes correspond to variables relevant to the domain (the pre-selection of which may be non-trivial); these may be binary, or take on a range of values. The directed links between variables correspond to the autonomous mechanisms that are supposed to mediate between these variables, and hence reflect the dependencies between them.

A simple causal graph is depicted in Fig. 1. In this example the model is restricted to three binary variables: *Bronchitis*, *Cough*, and *Insomnia*. There is presumed to be one mechanism that leads from *Bronchitis* to *Cough*, and another that leads from *Cough* to *Insomnia*.

Typically the dependencies between variables are probabilistic – reflecting either the incompleteness of the causal model or genuine noise. This uncertainty is represented by conditional probability distributions for each node (referred to as the *parameterization* of the

graph). Thus in our simple example the strength of dependency between *Bronchitis* and *Cough* is represented by two conditional probabilities – the probability of *Cough* given *Bronchitis*, and the probability of *Cough* given no *Bronchitis*. A high probability for the former would correspond to the belief that *Bronchitis* is very likely to cause *Cough*; a high probability for the latter would correspond to the belief that *Cough* is also very likely to be caused by other variables not represented in our simple model.

Given certain assumptions,[1] the structure of a causal graph will fully capture the probabilistic dependencies amongst all of the represented variables. A fundamental relation here is that of 'screening off' or conditional independence. For any three variables *A*, *B*, *C*: *A* and *B* are conditionally independent given *C* if P(*A*|*B*&*C*) = P(*A*|*C*); once you know the value of C, learning the value of *B* does not provide additional information about *A*. One causal graph representation that implies screening off is when *C* intercepts all directed paths between *A* and *B*. Thus in the causal graph in Fig.1, the fact that the *Cough* node is in between the nodes for *Bronchitis* and *Insomnia* implies that *Bronchitis* and *Insomnia* are conditionally independent given *Cough*. Once you know the value of *Cough*, learning the value of *Bronchitis* tells you nothing more about the value of *Insomnia*.

By representing conditional independencies in this way, causal graphs provide a powerful tool for organizing knowledge, and for inferring the effects of new observations. As the graphs increase in size, these independence relations can greatly simplify such computations. For example, one could supplement the simple model in Fig.1 with a complex network of nodes and links between *Bronchitis* and *Cough*, but so long as the variable *Cough* still intercepts all links from *Bronchitis* to *Insomnia*, knowledge of *Cough* is all one needs to make inferences about *Insomnia*.

Making inferences given new information

The structure of a causal graph, in combination with the parameterization of its nodes, determines what inferences we can make on the basis of new information. When this information takes the form of an observation, then Bayesian updating tells us how we ought to modify our probabilities. For example, given the causal model in Fig. 1, if we find out that Jim has a cough, we should increase (to some degree, depending on the parameters) both the probability that Jim has *Bronchitis*, and the probability that he has *Insomnia*. However, what if we changed the value of *Cough* by giving him a cough suppressant? Such an action warrants a change in our belief that he has *Insomnia*, but does not warrant any change in the probability we assign to him having *Bronchitis*.

More generally, the probabilistic inferences we are licensed to draw after observing the value of a variable may not be the same as those after intervening to set that variable to the same value. Bayesian updating, indeed any formal probability model, fails to recognize this. It does not differentiate between *observing* and *acting*. That is, the same conditional probability P(X|Y) is used to represent the probability of X given that Y *is observed*, and the probability of X given that *we do* Y. But these can be quite different, as our example illustrates – the probability of *Bronchitis* given the absence of a cough is distinct from the probability of *Bronchitis* given that we remove the cough.

The Representation of Intervention

One of the innovative features of the causal model framework is that it proposes a normative account for the representation of interventions, and for the inferences that they license. In so doing, it formalizes the difference between observation and intervention.

Pearl (2000) achieves this through the introduction of the '*do(•)*' operator. In short, this amounts to representing an intervention in terms of a minimal modification of the causal graph. Thus a simple intervention to set a variable to a particular value is represented by the removal of all arrows into that variable, without altering the other directed links in the graph. The effects of the intervention are then computable through Bayesian updating on this 'mutilated' graph.

To illustrate using the graph in Fig.1, consider an intervention (e.g., use of a cough suppressant) that sets *Cough* to the value low. This leads to the modified graph in Fig. 2: The directed link from *Bronchitis* to *Cough* is deleted whilst the link from *Cough* to *Insomnia* is left unchanged. In effect the intervention amounts to placing the variable *Cough* under the influence of a new mechanism that sets its value to low.

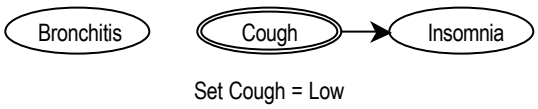


Figure 2: Causal graph after intervention

This account provides a normative model for the representation of both actual and imagined interventions, and tells us how these interventions will (or would) affect the values of the other variables in the system. In particular it dictates which probabilistic inferences we are entitled to make. Thus the modified graph in Fig. 2 permits us to infer a lower probability of *Insomnia*, but no change in the probability of

[1] For example, the explicit representation of any variable that affects two or more other variables in the model.

Bronchitis. The latter prohibition is reflected in Pearl's terminology by the difference between P(*Bronchitis*| ~*Cough*) and P(*Bronchitis*| *do*(~*Cough*)), and captures the basic asymmetry of the cause-effect relation: manipulating a cause can change an effect but not vice-versa.

Learning causal structure

The appropriate representation of intervention is not just critical to predicting the effects of our actions; it is also important for the discovery or learning of causal structure. Causal models can be learned from explicit instruction about how the world works, but we can also learn about causal structure through observation or through intervention. These are not exclusive, but it is useful to distinguish cases in which one is restricted to observational data alone from those in which one also has the opportunity to intervene.

Observational learning

The causal model literature in AI has developed various algorithms for inferring causal structure from observational data, many of which exploit the conditional dependencies encoded in the structure of a causal graph. So far none of these have been proposed as models of actual human discovery, although they do suggest some general principles that are relevant to such enquiries. For example, the establishment of conditional dependencies is a crucial starting point for the construction of a causal graph, so it is important that people are able to make judgments of conditional dependence versus independence. In contrast, the precise parameterization of those dependencies is not always required to discover correct causal structure

Moreover, the graphical approach clarifies which causal structures can be differentiated on the basis of observational data alone. It establishes equivalence classes of structures ("*Markov equivalence*") that share conditional dependencies and are thus indistinguishable on the basis of observation alone.[2] For example, in a model made up of just two nodes, *A* and *B*, ascertaining their probabilistic dependence does not tell us whether *A* causes *B*, or *B* causes *A*.

Even if causal structures are from different Markov equivalence classes, it might be difficult for people to distinguish them on the basis of observational data. Indeed, selecting between certain structures requires careful tracking of observed frequencies and subtle inferences based on what one would expect to see. For example, consider the two possible causal structures depicted in Fig. 3. In order to distinguish these on the basis of observation alone, one must determine whether or not blurred vision and headache are independent given high wine consumption (conditional independence would only hold if the data were produced by the model on the right). This may require many observations and careful tracking of the relevant relative frequencies.

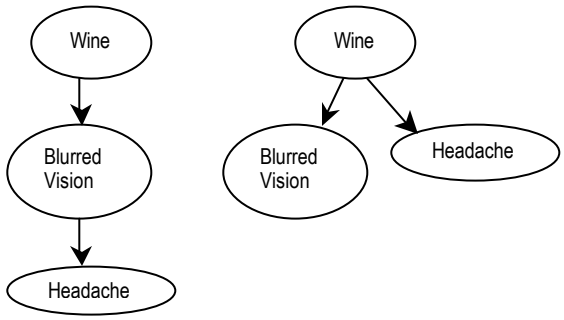


Figure 3: Two possible causal models

Interventional learning

Another way to learn about causal structure is to actively interact with the system under study and to observe the consequences. This seems to apply to the infant playing with a new toy as much as to the scientist running controlled experiments. Whilst this is often recognized as an important source of causal knowledge, it has received less attention in the human causal learning literature.[3]

Intuitively, the ability to intervene on a system should facilitate our learning about its causal structure. To take the simplest example, consider two variables that are known to be probabilistically dependent. Assuming no other relevant variables, the direction of this link can be determined by manipulating one of the variables and observing whether or not the other also changes. In a noisy system such learning may still require multiple trials and sensitivity to the observed frequencies. But interventional learning has several advantages over passive observation. Not only can it help to determine the direction of the causal links, it also allows selection of the kind of data to see, and thus to test out critical relations between variables. For example, let us return to the task of distinguishing between the two possible causal models in Fig. 3. One possible intervention is simply to drink a large amount of wine and then keep your eyes closed. If you don't get a headache, you can be reasonably sure that the chain model is the correct one. If the system is rather noisy you may have to repeat this experiment several times,

[2] One important qualification here is in the case of graphs in which causal links are necessary but not sufficient; that is, for a directed link from *A* to *B*: $1 > P(B|A) > 0$ and $P(B|\sim A) = 0$. Networks built from such links may be distinguishable even though they are Markov equivalent.

[3] The dominant approaches to human causal learning (e.g. Cheng, 1997; Dickinson, 2001; Shanks, 1995) concentrate on observational learning.

but it will still lead to greater confidence than making the distinction on the basis of observation alone.

Overview of Experiment

The central aims of this experiment were to compare the observational and interventional learning of a simple causal model, and to ascertain whether people represent their interventions in accordance with the normative model proposed by Pearl (2000). We used a typical observational learning paradigm (e.g. Shanks, 1995), but adapted it to include an interventional learning condition and a model selection task. The learning data were generated from a simple chain model (see Fig. 4).

Learning performance was assessed both through a model selection task and through the sensitivity of people's probability judgments to the appropriate conditional dependencies.

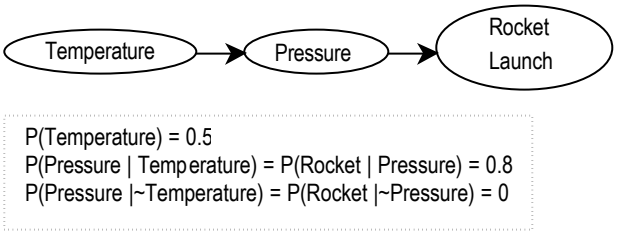


Figure 4: Causal graph used to generate stimuli for both observational and interventional tasks

Method

Participants. Thirty-three undergraduates from Brown University received course credit for their participation.

Materials and procedure. Initial instructions to the participants included an introduction to the notion of a causal model with examples of five candidate models. Each participant then completed both an observational and an interventional learning task. Two cover stories were used, one for each task (task order and scenario were counterbalanced across participants). Participants were asked to imagine that they were space engineers (chemists) running tests on a new rocket (perfume) in order to discover the underlying causal structure. They were told that previous tests had identified two variables as relevant to the success of the test. In the space engineer scenario the relevant variables were *Temperature* (either high or low) and *Pressure* (either high or low), and the outcome variable was whether or not the rocket launched. In the chemist scenario the variables were *Acid level* (either high or low) and *Ester level* (either high or low), and the outcome variable was whether or not the perfume was produced. In the *observation* task participants viewed the results of 50 test trials. On each trial they were shown the values of the two relevant variables, and then clicked on a button to view whether or not the outcome occurred. The learning set was constructed according to a chain model (see Fig. 4) and is shown in Table 1 (order of presentation was randomized for each participant).

Table 1: Frequency of presented instances in Observational Learning condition.

Temperature	Pressure	Rocket Launch	No	Prob
High	High	Yes	16	0.32
High	High	No	4	0.08
High	Low	Yes	0	0
High	Low	No	5	0.1
Low	High	Yes	0	0
Low	High	No	0	0
Low	Low	Yes	0	0
Low	Low	No	25	0.5

Participants then proceeded to a test phase, in which they made various conditional likelihood judgments (e.g., given that *Temperature* is high, and *Pressure* low, what is the likelihood that the rocket launches?) plus a model selection question. This question presented participants with five candidate causal models – two chains, two forks, and a collider (Fig.5 shows one model from each category) – and asked them to select the model that they believed was most likely to have produced the data.

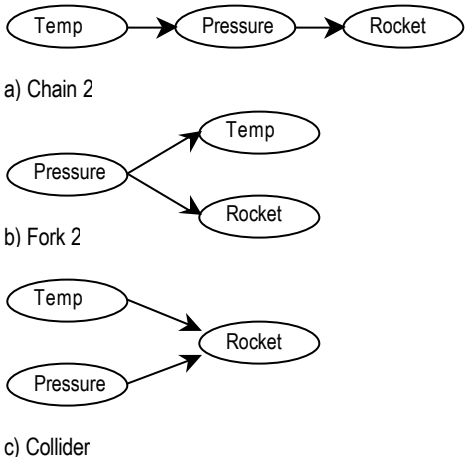


Figure 5: Three models from the selection task

In the learning phase of the intervention task, participants were able to set the value of one of the two relevant variables. They then viewed the resulting values of the outcome variable and the variable they had not intervened on. This learning set was generated from a pseudo-random table constructed in accordance with the same chain model. After running 50 tests they

proceeded to an identical test phase as in the observation task.

Results and discussion

Model Selection. The results for the model selection task are shown in Fig. 6, with the correct chain model designated as chain 2.[4] There were more correct model selections in the intervention than in the observation condition. However, whilst the correct model was the modal response in the intervention condition, overall responses were not significantly different from the uniform distribution ($\chi^2(4) = 2.91$, *ns.*). By contrast in the observational condition there was an overwhelming bias in favor of the collider ($\chi^2(4) = 40.79$, $p < 0.001$).

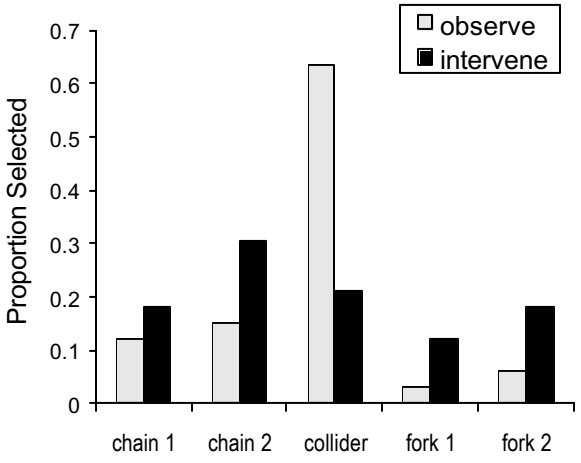


Figure 6: Model selection results in interventional and observational conditions (the correct model is chain 2).

Derived judgments of conditional independence. On the model used to generate the learning set (see Fig. 4), *Temperature* was independent of *Rocket launch* conditional on *Pressure*, that is: P($R|T\&P$) = P($R|P$). Participants' mean ratings for these two likelihoods are shown in Fig. 7. No significant difference obtained between the two likelihoods in the intervention condition, suggesting that participants were sensitive to this conditional independence. This is reinforced by the fact that 19 out of 33 participants judged the two likelihoods equal. This contrasts with the observation condition, in which the mean likelihoods differed substantially, and only 8 out of 33 participants judged them equal.

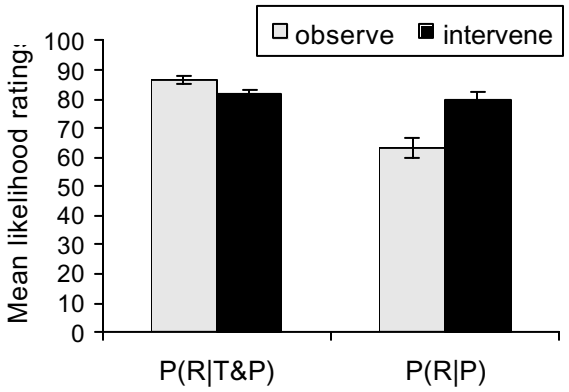


Figure 7: Mean conditional likelihood ratings for the outcome variable R (rocket launch).

Compatibility of judgments with the *do* operator. One way to assess the extent to which participants represent their interventions in line with the *do* operator is to look at their judgments of the likelihood that *Pressure* was high given that *Temperature* was low, P($P|\sim T$). Recall that the correct judgment for this likelihood is zero; *Pressure* is never high if *Temperature* is low. However, when participants intervene on the *Pressure* variable and set it to high they temporarily break the link between *Temperature* and *Pressure*. In such cases the value of *Temperature* is equally likely to be high or low (its base rate = 0.5). If participants fail to represent their interventions appropriately, by not 'mentally' removing the link from *Temperature* to *Pressure* when they intervene on *Pressure*, they may erroneously judge that P($P|\sim T$) > 0. This is because 50% of the time when they set *Pressure* high they will observe *Temperature* as low. In other words, they might fail to mark the distinction between action and observation.

To test out this possibility we compared people's judgments for P($P|\sim T$) with the relative frequencies they actually observed; i.e., with the proportion of times they observed both low *Temperature* and high *Pressure* (regardless of whether they intervened on *Temperature* or *Pressure*). As shown in Fig. 8, participants' mean judgments for P($P|\sim T$) were very close to the frequencies they observed, and significantly different from the normative value of zero.

[4] One complication is that the chain model used to generate the data is Markov equivalent to fork 2. However, although not inconsistent with the observational data, this model requires an idiosyncratic parameterization whereby one effect (temperature) occurs more often than its sole cause (pressure). Very few people chose this model in the observation condition.

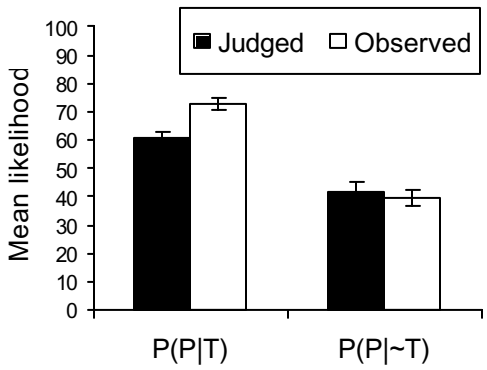


Figure 8: Mean likelihood ratings and observed relative frequencies in the intervention condition.

This result could indicate a failure by participants to implement the *do* operation when inferring the relation between *Pressure* and *Temperature*. However, there are alternative explanations for this finding. One possibility is that participants interpreted the likelihood question in terms of observational rather than interventional probabilities, and accurately reported the relative frequency with which low *Temperature* and high *Pressure* co-occurred, regardless of whether they believed that low *Temperature* would cause high *Pressure*. This fits with numerous studies showing that people encode the relative frequencies of events automatically, and often use these as a basis for their likelihood judgments (e.g., Hasher & Zacks, 1984).

Second, on Pearl's account the notion of an intervention is only well defined relative to a specific causal model. Thus if people uphold an incorrect model (as the majority of the participants did) they are unlikely to give appropriate estimates for the interventional probabilities. Moreover, even those participants that do select the correct model will have entertained various incorrect ones through the course of learning, and it may be very hard for them to retrospectively revise their prior observations.

Conclusions

This experiment demonstrated a contrast between observational and interventional learning, both with respect to people's model selection and their likelihood judgments. Under observational learning, participants exhibited a strong bias for the collider, despite the fact that the variables they judged to be independent were highly correlated in the data. This suggests that they were engaged in predictive learning of the outcome variable (e.g., *Rocket launch*) on the basis of two indicator cues (e.g., *Temperature* and *Pressure*), effectively treating them as independent causes of the outcome. This resonates with research on associative learning (e.g., Shanks, 1995), and multiple cue probability learning (e.g., Hammond, 1996), where models that assume the independence of causes fit the human data well. One factor likely to encourage this kind of learning was the manner in which the data were presented (e.g., indicator variables followed by outcome variable).

Interventional learning increased sensitivity to the appropriate conditional independencies and eliminated the bias for the collider, but the effect on model selection was not entirely beneficial. Although the correct chain was the modal choice, the majority of participants still chose the wrong model. Taken together with the observational results this implies that participants might have experienced too few trials to confidently discriminate between the models.

Whatever the precise reasons for sub-optimal performance in these tasks, the experiment shows that the automatic mechanisms that allow us to engage in the predictive learning and encoding of noisy information can sometimes override our discovery of the causal model that generates this information. Nevertheless, the difference we did find between observational and interventional learning encourages us that people are able to make use of the special kind of information afforded by intervention, and that future models of learning need to incorporate methods that represent the effect of action.

Acknowledgments

This work was funded by NASA grant NCC2-1217 to Steven Sloman. We thank Sean Stromsten and Dave Sobel for valuable comments.

References

Cheng, P. (1997). From covariation to causation: a causal power theory. *Psychological review*,104. 367-405.

Dickinson, A. (2001). Causal learning: an associative analysis. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.

Glymour, C. (2001). *The mind's arrows.* Cambridge, MA: MIT Press.

Hasher, L., & Zacks, R.T. (1984). The automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372-1388.

Hammond, K.R. (1996). *Human judgment and social policy*. New York: Oxford University Press.

Pearl, J. (2000*)*. *Causality*. Cambridge: Cambridge University Press.

Shanks, D.R. (1995). *The psychology of associative learning.* Cambridge: Cambridge University Press.

Data analysis of conceptual similarities of Finnish verbs

Krista Lagus (krista.lagus@hut.fi)
Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 9800, 02015 HUT, Finland

Anu Airola (anu.airola@helsinki.fi)
Department of General Linguistics, University of Helsinki
P.O.Box 9, 00014 University of Helsinki, Finland

Mathias Creutz (mathias.creutz@hut.fi)
Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 9800, 02015 HUT, Finland

Abstract

The study of the conceptual representations that underlie the use of language is a problem motivated from both a cognitive research point of view and that of construing language models for various language processing tasks. In this work, we organized 600 Finnish verbs using the SOM algorithm. Three experiments were conducted using different features to encode the verbs: morphosyntactic properties, individual nouns, and noun categories in the context of the verb. In general, the morphosyntactic properties seem to draw attention to semantic roles, whereas nouns as features seem to highlight clusters formed on grounds of topics in the text.

Introduction

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual representations that underlie the use of language is important for applications such as speech recognition. Due to the redundancy in communication, by studying large amounts of data it may be possible to induce the conceptual, system-internal representations which provide a grounding for meanings of words. Whether this is possible, and if so, how, is an interesting and controversial question.

A central problem in learning a language or in estimating a language model[1] from data is how to generalize from particular observations to new, similar instances. Generalization requires knowledge of similarities between words, concepts and other units of language and thought, i.e., similarity representations.

The hypothesis that the semantic similarity of two words correlates strongly with the similarity of their contexts has been widely discussed in linguistics and psychology (for recent treatments, see Levin, 1993 and Miller & Charles, 1991).

It has been proposed by Gärdenfors that a central part of our conceptual reprensentations are grounded in various low-dimensional *conceptual spaces*. A conceptual space is defined as a set of quality dimensions with a geometrical structure (Gärdenfors, 2000). Examples of conceptual spaces near our perceptual apparatus are colors and the pitch of sounds. For many higher order concepts a geometric interpretation can be found, as well. For example, comparative relations such as 'longer than' can be represented as a geometric relation between two elementary length spaces. Gärdenfors proposes a subset of concepts called *natural concepts*:

> A natural concept is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated.

An inherent and important property of the proposed conceptual spaces is that they provide a meaning representation that is ordered and offers means for representing similarities, often in terms of some continuous-valued underlying qualities. Gärdenfors gives examples of conceptual spaces that humans are likely to have. However, an open research question remains for both brain research and the study of language use: What are the possible conceptual dimensions that humans utilize?

In this work we analyze the use of Finnish[2] verbs with the following goals in mind: (a) to uncover possible conceptual spaces, i.e., underlying, organizing semantic qualities or properties, (b) to study semantic similarities of verbs in actual language use. In particular, we examine the kinds of semantic or conceptual ordering qualities that appear to affect the distribution of features in the immediate context of a verb, in particular (1) morphosyntactic properties of nearby words, and (2) the nearby nouns and (3) unsupervised categories of nearby nouns. In effect, we rely on the redundancy in communication and assume that certain regularities observed in the distributions of verb contexts will contain significant information about the semantics of the verb as well.

[1] For an introduction to statistical language modeling see (Manning & Schütze, 1999). Their applications include speech recognition, machine translation, and dialogue agents that converse with humans in order to perform tasks such as answering questions about train schedules and booking flights.

[2] Most of the research on language is carried out using English data only, which creates a too narrow or misleading picture of the modeling apparatus underlying language learning and use.

In order to obtain a simultaneous visualization and a clustering of the verbs (and in one experiment, nouns as well), we apply the self-organizing map (SOM) (Kohonen, 1995) algorithm. The visualized ordering of the verbs is studied qualitatively to obtain an understanding of the conceptual dimensions by which the verbs are likely to be organized. Furthermore, the obtained clustering is compared to a kind of 'ground truth', namely a semantic classification of Finnish verbs suggested by Pajunen (2001).

Self-Organizing Map (SOM) algorithm

The SOM (Kohonen, 1982; Kohonen, 1995; Kohonen et al., 1996) is an unsupervised neural network method that is able to arrange complex and high-dimensional data so that similar inputs are, in general, found near each other on the map. The ordered map display can then be utilized to illustrate various properties of the data set in a meaningful manner.

The algorithm automatically places a set of reference vectors—also called model vectors—into the input data space so that the data set is approximated by the model vectors. Each reference vector corresponds to a *map unit* on a two-dimensional regular grid. In effect, the grid and the vectors form a two-dimensional 'elastic net' in the high-dimensional input space: after application of the SOM algorithm, the map follows the data in a nonlinear fashion. The algorithm simultaneously obtains *a clustering* of the data onto the model vectors and a *nonlinear projection* of the input data from the high-dimensional input space onto the two-dimensional ordered map.

Prior work on unsupervised word categorization and projection

It has been shown that distributional information of word contexts can, at least for English, be used to induce syntactic categorization, and to some degree, semantic categorization as well using various methods (cf. e.g., Finch & Chater, 1992; Charniak, 1993; Honkela, 1997; Redington, Chater & Finch, 1998). The SOM has been applied to clustering English word forms based on the word forms in their immediate contexts in (Ritter & Kohonen, 1989; Honkela, 1997). Furthermore, the word categories obtained in such a manner have been used for encoding the meaning of documents e.g., in (Lagus et al., 1996).

Various alternatives to using SOM exist, including other clustering methods such as hierarchical clustering (used e.g. in Redington, M., Chater, N., & Finch, S., 1998; Pereira et al., 1993). Moreover, different metrics or clustering criteria can be applied, such as relative entropy (e.g. in Pereira et al., 1993) or minimum description length (in Li and Abe, 1998). Moreover, projection methods that do not form clusters but only project the data into a lower-dimensional space include a number of nonlinear projection methods under the name multidimensional scaling (MDS), and linear projection methods such as latent semantic analysis (LSA). Compared to these, a particular property of the SOM is that it simultaneously forms a grouping and a nonlinear projection of the data set.

In constructing the feature vectors for each word, one may look at the occurrence of individual words, or of syntactic word categories, or of morphological or derivational features (e.g. in Light, 1996). The position in which features are examined may be defined in terms of a wide window where more distant occurrences contribute less (e.g. Gallant et al. 1992; Lund and Burgess, 1996), or looking at only specific positions near the word, or according to some grammatical relationship with the word (e.g. Pereira et al., 1993; Schulte im Walde, 2000).

Experiments

We carried out three experiments on organizing and clustering verbs using the SOM algorithm. For the verb encoding, the following types of contextual features were explored: morphosyntactic properties, individual nouns and noun categories.

Corpus and data set

A corpus consisting of 13.6 million words of Finnish newspaper text [3] was used in the experiments. The examined set of verbs consisted of the 600 most frequent Finnish verbs, returned to their base forms[4]. Each context of a verb in the corpus was examined. The context was defined as the preceding and the two following words relative to the verb.

Encoding of the verbs

Morphosyntactic properties. In the first experiment, we preferred overtly marked morphosyntactic features. The selected features were 1) case endings (see example (i)), 2) endings of the nominal forms of verbs (see example (ii)), and two closed-class parts of speech, namely 3) adverbs and 4) adpositions (see example (iii)). In addition, the features NUM (for digits), and PUNCT (for punctuation marks) were included since they can be considered as visible features, too.

In Finnish, the primary means for coding various semantic-functional dependencies in a clause is the case-marking system. The case endings are added to stems, as shown in the following example:

(i)

Lapse-t	ajoivat	kaupunki-in.
child	drive	city
N-PL-NOM		N-SG-ILL

'The children drove to the city.'

The feature set used also includes two nominal (non-finite) verb forms, namely the 1st and the 3rd infinitive. The infinitives function in a sentence as nouns:

[3] Corpus by CSC, http://www.csc.fi/kielipankki/.

[4] The morphosyntactic analysis of word forms was performed using the Conexor FDG parser for Finnish, ©Conexor Oy and Anu Airola.

(ii)

Halua-n	lähte-ä	syömä-än.
want	go	eat
V-PRES-SG1	V-INF1	V-INF3-ILL

'I want to go to eat.'

Location or movement can be coded by using an adposition (iii) instead of case endings. Adpositions include postpositions (PSP) and prepositions (PRE):

(iii)

Lapse-t	ajoivat	kaupunki-a	kohti.
child	drive	city	toward
		N-SG-PTV	PSP

'The children were driving towards the city.'

The lower-level constituents, e.g., noun phrases, are predominantly head-final, the order being modifier-before-head. Adjectives agree in number and case with the head-noun when they occur as attributes (iv). Due to this redundancy, the function of a dependent noun phrase can be inferred before hearing or seeing the head of the phrase.

(iv)

Emmi	ajaa	punaise-lla	auto-lla.
Emmi	drive	red	car
N-PROP	V-SG3	A-SG-ADE	N-SG-ADE

'Emmi drives in a red car.'

At the clause level, the basic, or default, word order is subject-verb-object. According to (Hakulinen, Karlsson & Vilkuna, 1980), subjects precede finite verbs in 61% of all sentences in standard written prose. However, as Vilkuna (1989) points out, clause-level word order in Finnish shows great freedom; for example, in a simple sentence consisting of a subject, an object, a verb and one or two adverbials, all permutations are at least grammatically possible.

To sum up, we used a set of 21 mostly non-overlapping morphosyntactic properties, collected from three different textual positions relative to the verb. Each verb was thus encoded using a 63-dimensional feature vector. Averaged over a large number of samples, the value of a dimension in the vector is the conditional probability of a particular morphosyntactic feature in a particular contextual position given that verb.

Individual nouns as features. In the second experiment, instead of morphosyntactic properties individual nouns were used as features. The feature set consisted of the 10,000 most frequent nouns, returned to their base forms.

In order to keep the size of the feature vector reasonable, random projection was applied: Each of the 10,000 nouns was represented as a 500-dimensional vector, where 5 randomly selected positions were set to 1 and the values of the other dimensions were 0. The correlations thus introduced between words are in general negligible: random projection has been shown to roughly preserve distances between vectors, if the dimension of the projected vectors is sufficiently large. A theoretical treatment is presented in (Kaski, 1998); for empirical results on the use of random projection in the representation of documents, see (Kohonen et al, 2000).

Table 1: Sample noun categories.

Finnish nouns	Translations	Characterization
Matti, Jukka, Riitta, ...		first names of persons
maanantai, tiistai, ...	Monday, Tuesday	week days
kirja, levy, kokoelma, näytelmä, ...	book, record, collection, play	products of art
syy, pakko, tarkoitus, taipumus, ...	reason, obligation, intention, inclination	modalities

As in the previous experiment, each contextual position was encoded as a separate part of the feature vector. The resulting dimensionality of the vector was 1,500.

Noun categories as features. In the third experiment, noun categories were used as features. The categories were obtained by using the SOM algorithm to cluster the set of the 10,000 nouns from the previous experiment. The nouns were clustered based on verbs appearing in the same sentence at a maximum distance of five words. The position of verbs within the window was not taken into account. Verbs occurring at least 20 times in the corpus were considered as features, which yielded a total number of 3,089 verbs. Again, random projection was applied to reduce the dimensionality of the vectors.

A noun map consisting of 160 units was constructed. Each map unit was regarded as a noun category. Some examples of the resulting noun categories are shown in Table 1.

Next, the feature vectors for the 600 verbs were created. The encoding was identical to that of the first experiment except that instead of morphosyntactic properties, noun categories were used as features. The resulting feature vectors were 480-dimensional.

Creation of verb maps

In each of the three experiments, the feature vectors representing our selection of 600 frequent verbs were organized on a map of 140 units using the SOM ToolBox (Vesanto et al., 2000). As a consequence, verbs having similar feature vectors, and hopefully similar semantic representations, can be found close to each other on the map.

Results

The verb maps generated using different features were evaluated in two ways: by comparing to an

Table 2: Quantitative comparison with Pajunen's classification.

Exp. no.	Type of features	Precision
1	Morphosyntactic	35.7%
2	Individual nouns	23.6%
3	Noun categories	27.5%

existing classification, and by exploring the ordering of the verbs on the visualized maps.

Comparison to an existing verb classification

We compared the obtained clustering to Pajunen (2001), which is the most comprehensive semantic classification of Finnish verbs available. The semantic classification Pajunen presents is based both on conceptual classes, that is abstract schemas of states of affairs, and the theory of semantic fields (about field-theory, see e.g. Lyons, 1977). If compared e.g. to Levin's (1993) large-scale classification of English verbs, it can be seen that both Pajunen and Levin rely on the notion of semantic determination, i.e., the assumption that semantics determines syntax. However, while Pajunen explains form in terms of meaning, Levin (1993: 5) assumes that 'verbs that fall into classes according to shared syntactic behavior would be expected to show shared meaning components.'

Only 200 of our 600 verbs are mentioned in Pajunen's classification. These are divided into 54 classes, with 1-13 verbs per class. The comparison was carried out as follows: The set of verbs that formed a class by Pajunen were considered as correct hits for each other. For each of the 200 verbs, the map unit of that verb was examined, and the *precision* was the number of hits divided by the total number of verbs in that map unit[5].

For each experiment, the precisions, averaged over the set of 200 verbs are reported in Table 2. For the experiments 2 and 3, in which random projection was used in feature encoding, the results are averages of five runs with different random seeds. A paired t-test showed significant differences for each pair of experiments ($p=0.9997$ between 1 and 2, $p=0.9706$ between 2 and 3, and $p=0.9853$ between 1 and 3). The experiments, ordered by similarity to Pajunen's classification, were: morphosyntactic properties, noun categories, and individual nouns.

Visual inspection of the maps

The map obtained in the first experiment is shown in Figure 1. The organization of the map seems to highlight the importance of cultural, social and emotional aspects in lexical organization. Dimensions of social interaction, wielding of power, the will of an individual person, and manipulative behavior between people all occupy rather strong regions on the map.

[5]The verb itself was excluded, as well as all the verbs that were not mentioned by Pajunen.

Table 3: Sample verb categories based on noun categories (Experiment 3).

Finnish verbs	Translations	Topic
myydä, ostaa, tuottaa, palkata, työllistää, kattaa, vuokrata	sell, buy, produce, hire, employ, cover, rent	business
nousta, laskea, kasvaa, pudota, vähentyä, kohota, pienentyä, supistua, noutaa, kallistua	rise, decrease, grow, fall, diminish, rise, get smaller, contract, fetch, go up in price	stock rates
kuolla, hukkua, ampua, surmata, ammua, hyökätä, loukkaantua, menehtyä	die, drown, shoot, kill, moo, attack, get hurt, pass away	dying

The maps based on the distribution of individual nouns and noun categories (not shown) seem to be organized more according to a continuum from verbs describing subjective cognitive events, e.g., the states of mind of an individual, to the verbs expressing mostly social actions. There were clear categories that appeared to be clustered together on grounds of a common topic in the text. Some examples are shown in Table 3.

Note that the verb 'to moo' has apparently fallen into the 'dying' category due to an incorrect morphological analysis made by the FDG parser. The verbs *ampua*, 'to shoot', and *ammua*, 'to moo', have many common forms that are homonyms, e.g. *ammun*, 'I shoot' or 'I moo'. In fact, the FDG is reported to make 2–7 % of disambiguation errors for words in general.

Discussion

Based on visually examining the map and the clusters and a quantitative comparison to Pajunen's categories, the results are promising. It is nevertheless possible that some improvement may be achieved by (1) correcting the errors in preprocessing, (2) trying yet different types of features, feature windows, or feature encodings. Moreover, when the purpose is only to obtain individual clusters and not a visualization or projection (as done by the SOM), different clustering methods chould be examined as well.

Morphosyntactic properties appear to correspond most closely to the grounds by which Pajunen's classification is formed: both categorizations emphasize

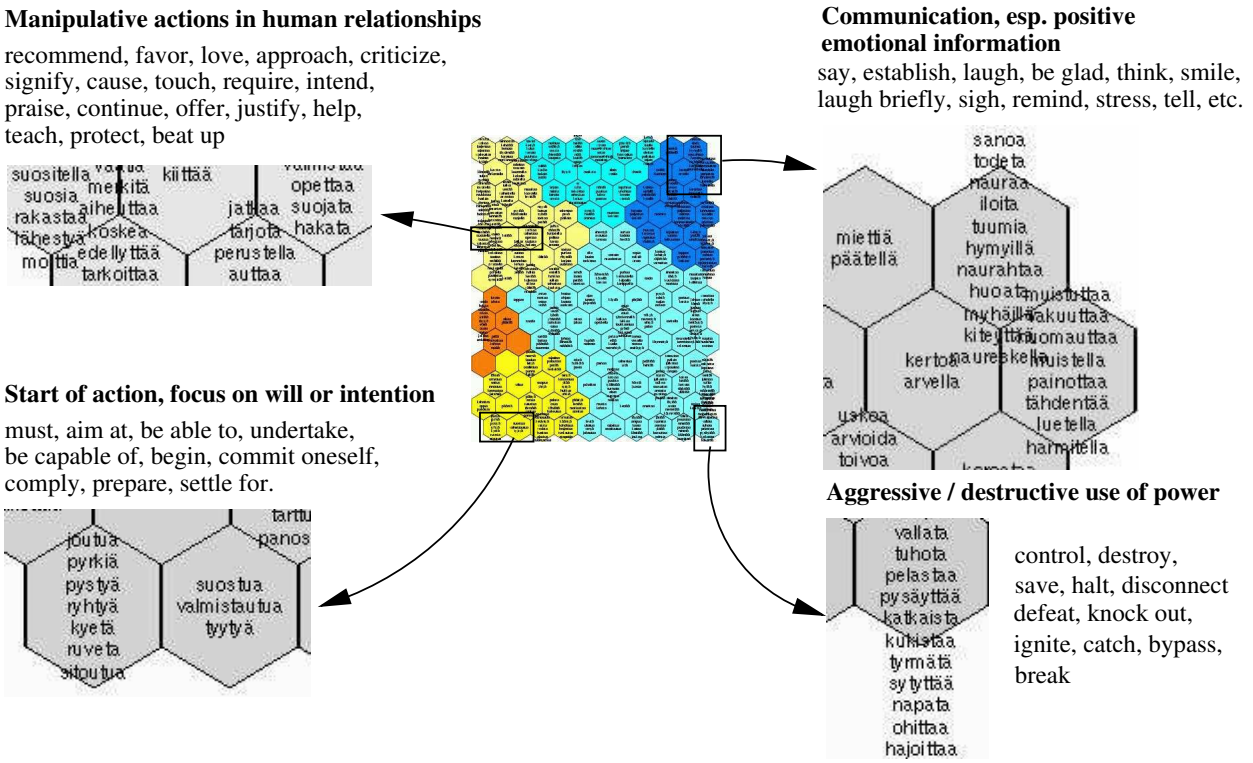


Figure 1: A map of the 600 verbs organized based on the distribution of morphosyntactic properties (non-finite forms of the verb were excluded). Properties of the preceding word and the two following words were considered. The contents of four sample map regions are shown in the insets. Many of the obtained categories correspond to categories defined by Pajunen. However, in Pajunen's classification the verbs in the lower right corner indicating 'destructive use of power' are further divided into two specific categories, namely (1) break verbs (*tuhota* 'destroy', *katkaista* 'break', *hajoittaa* 'break down') and (2) fight verbs (*pysäyttää* 'stop', *kukistaa* 'defeat', *tyrmätä* 'knock out'). Similar categories can be found in (Levin, 1993) for English verbs.

the semantic roles a noun phrase may bear in its clause. This seems reasonable, since in Finnish, the primary means to express semantic roles (e.g. AGENT and PATIENT) is the case system. Even though semantic roles cannot be simply derived from morphosyntactic cases, a strong correlation can still be assumed.

The other feature types, i.e. individual nouns and noun categories, exhibit fairly similar information regarding the verbs. In general, morphosyntactic properties seem to push the categorization towards the direction of linguistic semantics, while categorization based on nouns or noun categories is more a reflection of subject matters communicated through texts.

In some cases, the morphosyntactic map distinguishes verbs based on the *kind* of the patient (upper left corner with human→{human} relationships vs. lower right corner with human→{nation, abstraction} relationships). This result confirmed our expectations, and the understanding that the type of consequence of the action for the patient is not reflected in the morphological features. On the other hand, the maps based on noun features seem to make distinctions based on both the topic, the consequences for the patient {dying, creation, change of possession, change of state}, and the kind of patient {human, food, artefact}. In this way, the different types of features highlight different relevant aspects of categorizing verbs.

Conclusions

Different feature selections correspond to different assessments of what is important in the categorization of verbs. The categorization most similar to Pajunen's was obtained with morphosyntactic features. In contrast, it appeared that the noun features bring out the similarities between verbs in a richer and more useful manner.

It is interesting to consider whether the ordering qualities observed on the maps could count as quality dimensions of the conceptual spaces suggested by

Gärdenfors. The observed ordering qualities seem to reflect various higher-level cognitive, emotional, or social dimensions. These could be emergent properties of representations on some lower, more basic level. In fact, the emergent properties of one representation level or process are likely to be used as input features of another level or process.

By looking at the maps it seems clear that there are many relevant aspects for categorizing verbs, and any single categorization or ordering is of necessity reduced into considering only some of these. However, to obtain a more accurate representation, instead of a single categorization or projection, one should create several simultaneous categorizations (representations), induced using different kinds of features.

References

Charniak, E. (1993). *Statistical Language Learning.* MIT Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., and Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Finch, S., & Chater, N. (1992). Unsupervised methods for finding linguistic categories. In Aleksander, I., & Taylor, J. (Eds.), *Artificial Neural Networks, 2*, pp. II–1365–1368. North-Holland.

Gallant, S. I., Caid, W. R., Carleton, J., Hecht-Nielsen, R., Pu Qing, K., and Sudbeck, D. (1992). HNC's MatchPlus system. *ACM SIGIR Forum*, 26(2):34–38.

Gärdenfors, P. (2000). *Conceptual Spaces.* MIT Press.

Hakulinen, A., Karlsson, F., & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus.* Publications of the Department of General Linguistics, Yliopistopaino, Number 6, Helsinki.

Honkela, T. (1997). *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki University of Technology, Espoo, Finland.

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proc. COLING-00*, pp. 747–753.

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of IJCNN'98, Intl Joint Conference on Neural Networks*, vol. 1, pp. 413–418. IEEE Service Center, Piscataway, NJ.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biol. Cybern.*, 43(1):59–69.

Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin, 3rd edition 2001.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1996). SOM_PAK: The Self-Organizing Map program package. TR A31, Helsinki University of Technology, Laboratory of Computer and Information Science.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, V., & Saarela, A. (2000). Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585.

Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In Simoudis, E., Han, J., & Fayyad, U., (Eds.), *Proc. KDD-96*, pp. 238-243. AAAI Press, Menlo Park, CA.

Levin, B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation.* The University of Chicago Press, Chicago and London.

Li, H. and Abe, N. (1998). Word clustering and disambiguation based on co-occurrence data. In *36th Annual Meeting of the ACL, COLING-98*, pp. 749–755.

Light, M. (1996). Morphological cues for lexical semantics. In *ACL 34*, pp. 25–31.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Pajunen, A. (2001). *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* SKS, Helsinki.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of English words. In *30th Annual Meeting of the ACL*, pp. 183–190.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biol. Cybern.*, 61:241–254.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). Som toolbox for matlab 5. TR A57, Helsinki Univ. of Technology, Neural Networks Research Centre, Espoo, Finland.

Vilkuna, M. (1989). *Free Word Order in Finnish. Its Syntax and discourse functions.* SKS, Helsinki.

Multitasking as Skill Acquisition

Frank J. Lee (fjl@rpi.edu)
Department of Cognitive Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Niels A. Taatgen (niels@ai.rug.nl)
Artificial Intelligence, University of Groningen
Groningen, The Netherlands

Abstract

Multitasking allows people to cope with the ever changing and complex world that we live in. However, as much as cognitive scientists have learned about the details of human cognition, the question of how people acquire multitasking ability remains a mystery. In this paper, we argue that multitasking can be best understood as a product of skill acquisition. In particular, we describe *production composition*, a computational theory of procedural skill acquisition, which can account for the acquisition of multitasking skill. We explore this idea in this paper as part of our effort to develop a cognitive model of a simulated air-traffic controller Task.

Introduction

Multitasking is a critical ability that allows people to cope with and flourish in the complex world that we live in. However, as much as cognitive scientists have learned about the inner workings of human cognition, our ability to multitask remains a mystery. In this paper, we argue that we can best understand multitasking as a product of *production composition* (Taatgen & Lee, submitted), a computational theory of procedural skill acquisition that has been implemented within in the ACT-R framework (Anderson & Lebiere, 1998). Production composition has been used successfully to account for skill acquisition in a wide variety of domains including language learning (Taatgen & Anderson, submitted) and individual differences in complex skill acquisition (Taatgen, 2001). We believe that it can also be used to account for the acquisition of multitasking skill.

Multitasking

Multitasking is the ability to handle the demands of multiple tasks simultaneously. At the most basic level, this may involve executing multiple perceptual-motor actions at the same time, such as moving your attention to the next lane and turning the steering wheel. At a more complex level, this may involve interleaving the steps of many complex tasks, such as shifting down to a lower gear while navigating a curve and carrying on a conversation.

Important insights into people's ability to multitask come from the dual-task performance literature. One such insight is that while there is some interference between the two tasks that are being performed (with a caveat regarding the modality of stimuli and responses), people can consciously trade off performing one task for the other (Wickens & Gopher, 1977). Another is that people's performances in both tasks depend highly on their skill in the individual tasks (Allport, Antonis, & Reynolds, 1972). That is, being skilled in one task allows a person to perform it and other tasks with negligible impact on the overall performance of both tasks. For example, a skill driver might have little difficulty talking with a friend while driving, whereas a novice driver might find it difficult.

Skill Acquisition

Anderson (1982) proposed a theory of skill acquisition in terms of transitioning from declarative knowledge to procedural knowledge through a process called *knowledge compilation*. Initially, knowledge is in declarative form and is interpreted. Interpreting declarative knowledge is slow and may lead to errors, especially if the relevant knowledge cannot be retrieved when needed or erroneous knowledge is retrieved instead. With practice, declarative knowledge is compiled into procedural knowledge and is fast and free of errors. Newell and Rosenbloom (1981) proposed an alternate theory of skill acquisition called *chunking* that became an important component of the Soar cognitive architecture (Newell, 1990). Within Soar, skill acquisition is a function of combining multiple procedures into a single procedure and converting the current goal context into a more specialized procedure.

Production Composition

Production composition is a theory of skill acquisition that incorporates aspects of both Anderson's and Newell and Rosenbloom's account. It involves compiling declarative knowledge into procedural knowledge and combining multiple procedures into a new procedure. Consider the process of retrieving

information from declarative memory in ACT-R, which is usually done in two steps. In the first step, a production rule[1] issues a request to declarative memory for a certain piece of knowledge, while in the next step another production rule acts on the retrieved knowledge. Production composition eliminates the retrieval process and creates a single production rule out of the two original rules while substituting the retrieval into this new production rule. Through this process, general rules can be specialized into task-specific rules.

Figure 1 graphically illustrates this process. Before production composition takes place, a production rule requests an instruction from declarative memory on what to do next. Declarative memory returns with an instruction that the Enter key should be pressed in the current context of "Land plane 3". In response a production rule issues a motor command to press the Enter key, which initiates the motor system to actually press the Enter key. Production composition eliminates the retrieval from declarative memory and combines both of these production rules into one single rule, producing a task-specific production rule that issues a motor command to push the Enter key once it is in the context of "Land plane 3".

Production composition has been used successfully to model learning in a simulated air traffic controller task (Taatgen & Lee, submitted), inflection of the English past tense (Taatgen & Anderson, submitted) and the German plural (Taatgen, 2001), and strategy development in the balanced-beam task (van Rijn, van Someren & van der Maas, submitted). Perhaps reflecting the utility of production composition, it has been incorporated into the current version of the ACT-R cognitive architecture. In the next section, we discuss how production composition can be used to account for the acquisition of multitasking skill.

Multitasking and Production Composition

In the example given in Figure 1, all steps are executed serially. According to the ACT-R theory, however, each of the different subsystems, Hand, Declarative Memory and Production, as well as the Visual and other sensory-motor systems can work asynchronously and in parallel (Byrne & Anderson, 1998). This is not always possible: sometimes one subsystem must wait for information from another. The goal of multitasking in such cases is to exploit these gaps in processing by slipping in other useful processes. The production composition mechanism is capable of modeling this aspect of multitasking. Figure 2 gives a graphical example of this process, in which two tasks have to be carried out: Task

[1] Within ACT-R and other cognitive theories, procedural knowledge is often represented by *production rules* that have the form of "IF-THEN" condition-action rules.

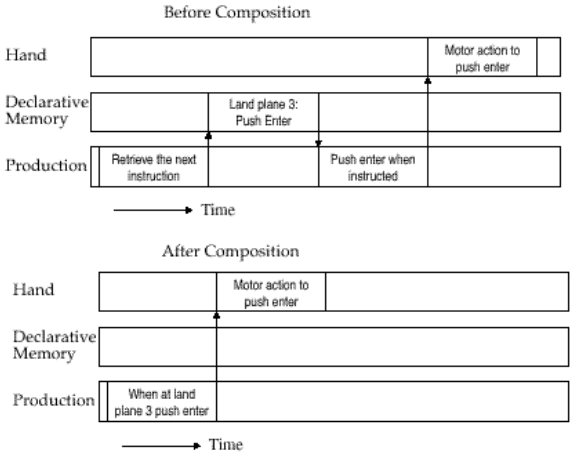


Figure 1: The production composition process.

A to push the Enter key and then check whether a light has gone on, and Task B to say "yes". Obviously these tasks do not make any sense without context, but consider them as part of a larger task, for example in context of the simulated air-traffic controller task that we will discuss later on.

The top left panel of Figure 2 depicts Task A for the novice, comparable to Figure 1. First, a production rule issues a request for the next instruction. Declarative memory produces the instruction to push the Enter key. Next, a production rule issues a motor command to the Hand to do this. Although it takes the motor system some time execute this command, a production rule immediately fires to retrieve the next instruction.

The retrieved instruction requests the visual system to check the light after the button has been pushed. The production rule that carries out this request has to wait for the instruction and the completion of the previous motor command. Only then can it issue a command to the visual system to check the light. Task B has a similar, although slightly simpler structure: an instruction is retrieved, after which the speech system is instructed to say, "Yes".

If both task A and B rely on declarative instructions, it is impossible to carry them out concurrently because declarative memory is busy almost all of the time. Once production composition has taken care of some of the declarative retrievals, multitasking is possible. Suppose task A has been composed into task-specific production rules but not task B. Now task A is carried out as in the top right panel of Figure 2: a rule issues the motor command, the motor command is carried out, after which a production rule issues the perceptual command. But now there is time left after the first production rule to do something else, for example slip in task B. The bottom left panel shows how this is done. After a rule has fired to initiate the Hand command, a new rule fires

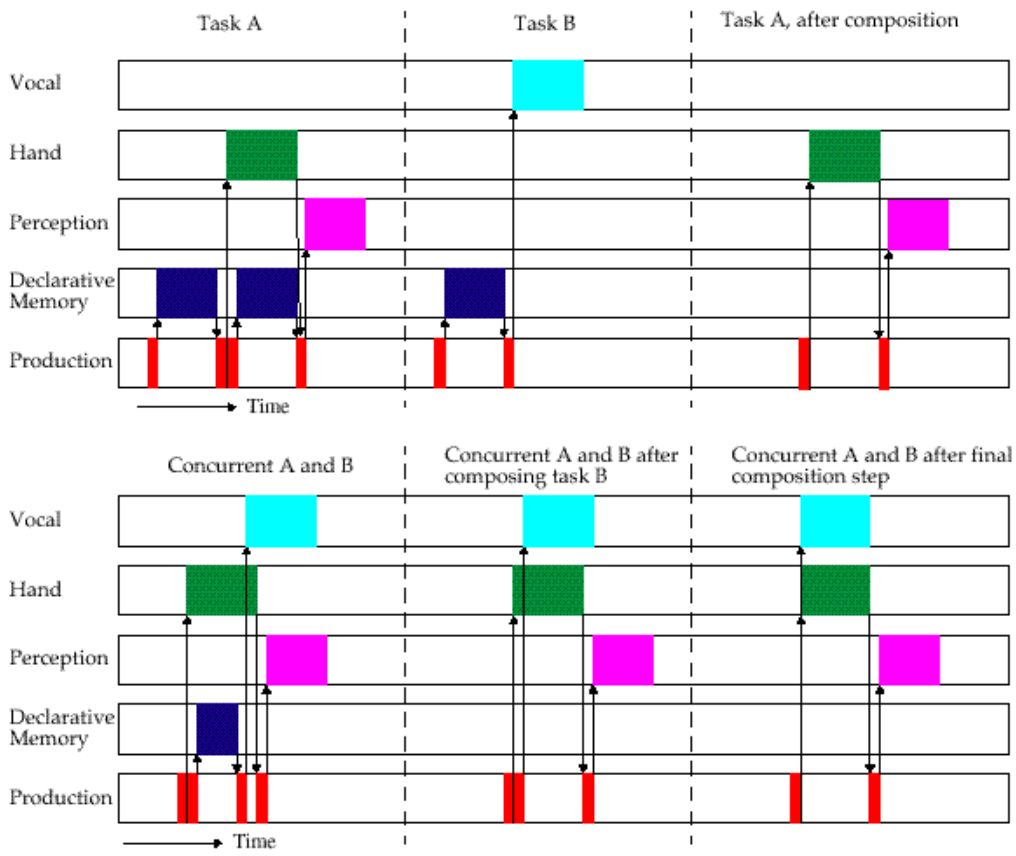


Figure 2: Development of multitasking production rules through production Composition.

to retrieve the instruction for task B. Depending on whether the Hand or retrieval from declarative memory is faster (we chose declarative memory in this case), the Vocal or the Perceptual command is issued. The other follows directly afterwards. The composition process does not stop here, because the retrieval in task B can also be eliminated, producing the situation in the bottom middle panel. Finally, the rule that initiates task A can be combined with the rule that initiates task B, producing the final state of the bottom right panel of Figure 2.

The Task

The task that we use in this paper to explore the concept of multitasking as skill acquisition is the Kanfer-Ackerman Air Traffic Controller (KA-ATC) Task (Ackerman, 1988; Ackerman & Kanfer, 1994). The KA- ATC task is composed of the following elements displayed on the screen: (a) 12 hold positions, (b) 4 runways, (c) information on current score, landing points, penalty points, conditions of the runways, and wind direction and speed, (e) a queue of planes waiting to enter the hold, and (f) 3 message windows, 1 for notifying of weather changes, 1 for providing feedback on errors, and 1 for displaying of the rules of the task in response to information requests by the participants. The 12 hold positions are divided into 3 levels corresponding to altitude, with hold level 3 being the highest and hold level 1 being the lowest. A typical display of the KA-ATC task is presented in Figure 3.

Six rules govern participant's actions in this task: (1) Planes must land into the wind, (2) Planes can only land from hold level 1, (3) Planes can only move 1 hold level at a time, but to any open position in that level, (4) Ground conditions and wind speed determine the runway length required by different plane types. In particular, 747's always require long runways, DC10's can use short runways only when runways are DRY or WET (i.e. not ICY), and wind speed is less than 40 knots, 727's can use short runways only when the runways are dry or wind speed is 0-20 knots, and PROP's can always use short runways, (5) Planes with less than 3 minutes of fuel remaining must be landed immediately, and (6) Only one plane at a time can occupy a runway.

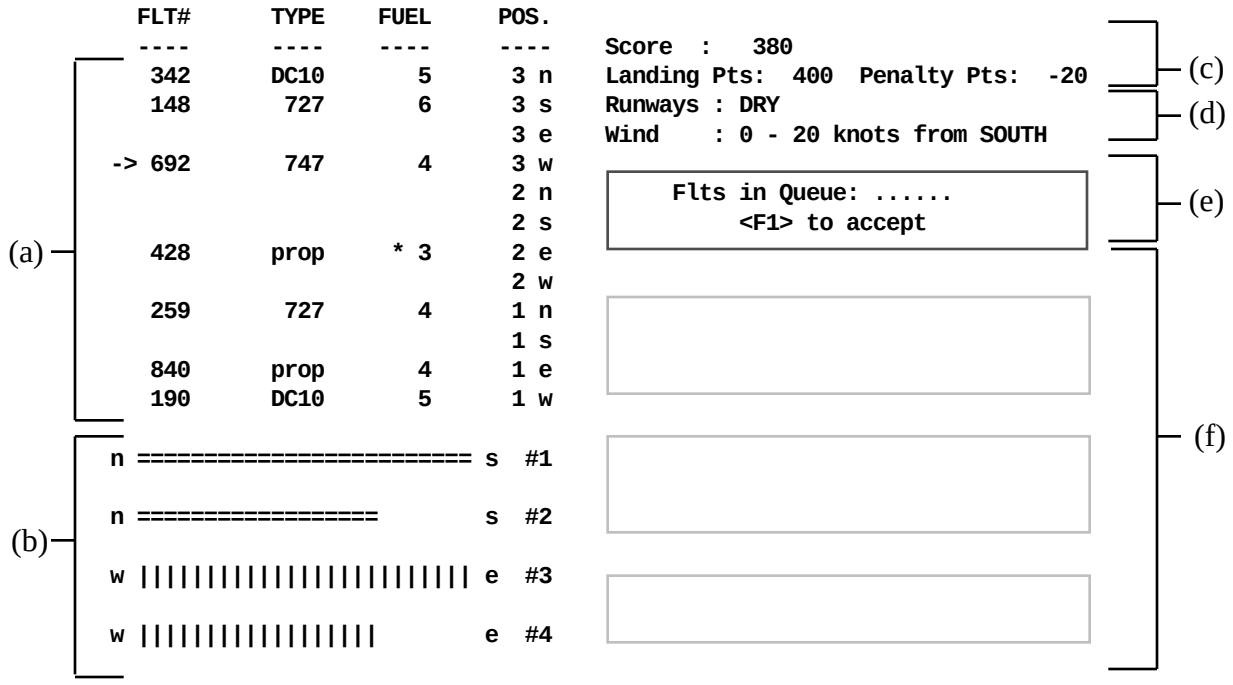


Figure 3: Kanfer-Ackerman ATC Task.

Participants can execute three actions in this task: (a) they can accept a plane from the queue into an open hold-position, (b) they can move a plane between the three hold-levels, and (c) they can land a plane on a runway. They can accomplish these actions by using four keys: the Up-arrow and the Down-arrow keys, ↑ and ↓; the F1 function key, **F1**; and the Enter key, ↵. They can move the cursor up and down the hold-positions and the runways using the ↑ key and the ↓ key. They can accept a plane from the queue into an open hold-position using the **F1** key. And, they can select a plane in the hold, place a selected plane in an open hold-position (either from the queue or from another hold-position), or land a plane on a runway using the ↵ key. In addition, participants can press the number keys **1** - **6** to examine the rules 1 - 6 any time during the task.

Participants are given 50 points for landing a plane, penalized 100 points for crashing a plane, and penalized 10 points for violating one of the six rules. A plane crashes when the fuel-level of a plane falls to 0 minutes. Planes are added to the queue approximately every 7 seconds and it takes 15 seconds for a plane to clear a runway. Once planes enter the hold position from the queue, they have between 4 - 6 minutes of fuel and begin to lose fuel in real time.

In Ackerman (1988), participants performed in the fair-weather condition where the wind speed was fixed to 0 – 20 knots and the runway condition was fixed to DRY. Under this condition, Rule 4 simplifies to the rule that all planes, except 747s, can land on a short runway.

The Model

Taatgen and Lee (submitted) have developed a model of the initial learning of the task. In this paper we describe a modification of this model to include some aspect of multitasking. The general idea of the model is that the participant in the experiment first encodes the instructions declaratively, forming a (often incomplete) plan on how to do the task. As interpretation of these instructions is slow, initial performance is also slow, resulting in poor performance. But due to production composition, a speed-up is realized that can account for the increase in performance.

The former model consisted of a fairly linear plan to land planes, to decide between the tasks to land planes, to move planes between hold levels and to get new planes from the queue. The model does not allow for much concurrency because of the linear structure of the plan that persists even after proceduralization. In order to test the new approach, we took out one aspect of the plan to land a plane, namely the checking of the wind direction. Wind direction has to be checked to see which of the runways can be used at the moment, and as it periodically changes, it has to be rechecked occasionally. In the original model, the wind was checked as one of the first steps in the landing procedure. We took this checking step out of the main plan, and made an alternative plan to check for the wind

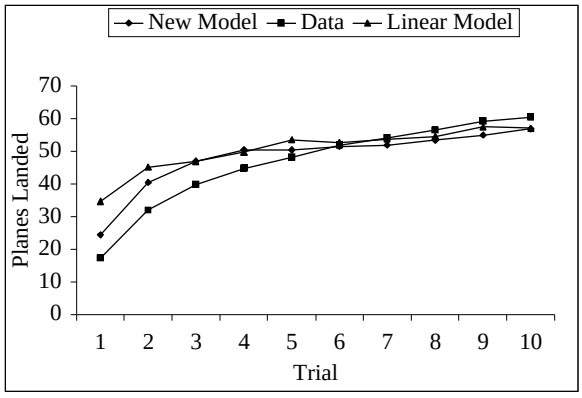


Figure 4: Number of planes landed.

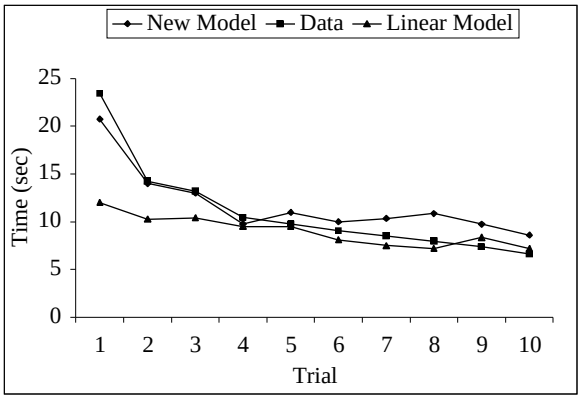


Figure 5: The elapsed time between a change in the wind direction and the first plane landed.

at moments of "slack time", for example when the arrow is moved to a plane or a runway. These arrow movements take multiple key-presses, allowing for some time for the model take a quick peek at the wind direction. This checking procedure will only succeed after the relevant steps have been proceduralized themselves, similar to the example in Figure 2.

The main question to be answered now is whether this change from a linear to a more parallel model improves the fit with the data. An interesting dependent measure in this respect is the time it takes for the participant (or model) to notice a change in wind direction. Although this cannot be measured directly (at least not in human participants), a measure (also used by Ackerman) is the elapsed time between a change in wind direction and the first landing of a plane on a runway in the new direction.

We compare the model predictions with data from Study 2 in the ONR data set (Ackerman & Kanfer, 1994), as reported in Ackerman (1988). The data from Study 2 were from 65 college undergraduates who completed 27 trials of the KA-ATC task with each trial lasting 10 minutes. For our model comparisons we only use trials 1 through 10, all in the fair-weather condition.

Figure 4 shows the overall score in terms of the number of planes landed in each 10-minute trial. Both the scores for the original "linear model" are shown (from Taatgen & Lee, submitted), and the predictions by the new model. Although the new model is more accurate than the old, linear model, the difference is slight. A larger difference can be seen in the time to notice a change in wind direction, the measure closely tied to the change in the model. Figure 5 shows the results. Although the linear model also predicts an improvement in this reaction time, because all processing is faster due to proceduralization, the new model matches the data much more closely especially in the first three trials where the model still has trouble interleaving checking the wind with other behavior.

The current model is not a full implementation of the principle of multitasking within a complex task like the KA-ATC. It just demonstrates one aspect of in the act of checking the wind direction. Fortunately this aspect can be verified empirically. The current model is not yet capable of explaining improvements in performance after trial 10, where human participants still gradually improve but the model does not. This can only be explained within the ACT-R theory by a more efficient schedule of perceptual, cognitive and motor processes (Lee & Anderson, 2000).

Discussion

Multitasking and Planning

It is worthwhile in our discussion to see how multitasking and production composition might be related to other areas of human behavior. Especially relevant is the area of planning that researchers in Artificial Intelligence have looked at closely. The mechanism of production composition is an automated process that is below a person's conscious control that automatically generates new procedural knowledge that are then tuned in their utility with use.

Planning, on the other hand, is largely seen as a deliberative and conscious process. In the context of multitasking, one can clearly imagine people reasoning about the structure of the multiple tasks that they must engage in, and explicitly devising a "plan" to interleave the tasks. This can happen at a larger time scale, such as when attempting to cook several dishes at the same time for a 7-course meal, or at a smaller time scale, such as trying to press the clutch and change the gear when learning to drive a car with a manual transmission.

From our perspective, planning and any other weak-method problem solving is completely consistent with production composition. Weak-methods, such as using instructions, examples, and planning, are all an aspect

of the declarative problem solving process that generates sequential actions that can be exploited by production composition to develop more efficient (i.e. multitasking) procedures.

We believe production composition can provide a resolution to the debate in the AI community between traditional planning versus reactive planning (c.f. Russell and Norvig, 1995). Traditional planning posits that agents reason over situations and actions in order to formulate a plan before taking the requisite actions. Reactive planning on the other hand posits that agents simply find the most applicable action in the current situation and executes them. Within ACT-R cognitive architecture, one can view weak-method problem solving as a mechanism for traditional planning and production composition as a mechanism for reactive planning. Weak methods generate sequential actions that are then used by production composition to generate reactive production rules. This makes perfect sense from the perspective of studying human behavior, since people display both types of planning.

Is Multitasking a Mystery?

Some may wonder, "Why is multitasking a mystery? Isn't it a well understood phenomenon?" Our response is that while the fact that people are able to multitask is well understood, it is less clear how multitasking skill is learned. In particular, most of the current models of multitasking, with very few exceptions, include explicit control structures to switch between multiple tasks. However, the question still remains as to *how* these control structures came into being. We have argued in this paper that multitasking is a product of production composition and is able to learn new control knowledge for performing multiple tasks.

Conclusion

The learning in the KA-ATC task can be conceptualized as retrieving instructions from memory and executing them (Taatgen & Lee, submitted). The two steps are then compiled by production composition into a single production rule. One result of production composition in the KA-ATC task is the acquisition of production rules that encapsulates the execution of a keystroke-level task (c.f. Lee & Anderson, 2001). However, another result of production composition is the combining of two (or more) learned keystroke-level production rules into a single rule. This second result, we argue, reflects the acquisition of multitasking skill.

References

Ackerman, P.L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288-318.

Ackerman, P.L., & Kanfer, R. (1994). *Kanfer-Ackerman air traffic controller task© CD-ROM database, data collection program, and playback program.* Office of Naval Research, Cognitive Science Program.

Allport, D.A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology, 24*, 255-265.

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369–406.

Anderson, J.R. (1995). *Cognitive Psychology and Its Implications.* NY: Freeman.

Anderson, J.R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Byrne, M. D., & Anderson, J. R. (1998). Perception and Action. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 167-200). Mahwah, NJ: Erlbaum

Fitts, P.M. (1964). Perceptual-motor skill learning. In A.W. Melton (Ed.), *Categories of human learning*. New York, NY: Academic Press.

Lee, F.J. & Anderson, J.R. (2000). Modeling Eye-Movements of Skilled Performance in a Dynamic Task. In N.A. Taatgen, & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling*. Veenendaal, The Netherlands: Universal Press.

Lee, F.J. & Anderson, J.R. (2001). Does learning a complex task have to be complex? A Study in Learning Decomposition. *Cognitive Psychology, 42*, 267-316.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, M.A.: Harvard University Press.

Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Taatgen, N.A. (2001). Extended the past tense debate: a model of the German plural. In K. Stenning & J. Moore (Eds.), *Proceedings of the Twenty-Third Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Taatgen, N.A., & Anderson, J.R. (submitted). Why do children learn to say "Broke"? A model of the past tense without feedback.

Taatgen, N.A., & Lee, F.J. (submitted). Production composition: A simple theory of complex skill acquisition.

van Rijn, H., van Someren, M, & van der Maas, H. (submitted). Modeling developmental transitions on the balance scale task.

Wickens, C.D., & Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. *Human Factors, 19*, 249-366

Using Cognitive Decision Models to Prioritize E-mails

Michael D. Lee, Lama H. Chandrasena and Daniel J. Navarro
{michael.lee,lama,daniel.navarro}@psychology.adelaide.edu.au
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

E-mail prioritization involves placing all of the 'useful' or 'good' unread e-mails at the top of the inbox, and all of the bad ones at the bottom. We use two cognitive decision models—a rational model, which considers all of the available information, and a fast and frugal model that uses one reason decision making—to prioritize e-mails. Experimental results, using real data obtained by unobtrusively logging e-mail user behavior, show that the fast and frugal model is just as effective as the rational model. The results also show that a Bayesian approach to learning is superior to the standard frequentist approach, because it balances the competing demands of exploration and exploitation in finding good e-mails. We use the results to draw some applied conclusions about the development of an e-mail prioritization system, and note some theoretical implications of the results for the cognitive modeling of human decision making in general.

Introduction

Anybody who has returned from holidays to be confronted with 600 unread e-mails appreciates the need for prioritization. Ideally, we would like an unread inbox to rank the e-mails, putting those that are the most 'important', 'urgent', 'useful' or 'good' at the top, and those that are less important at the bottom.

While machine learning methods have been applied to the problem of e-mail prioritization (e.g., Macskassy, Dayanik, & Hirsh 1999; Mehran, Dumais, Heckerman, & Horvitz 1998), it has typically not been treated as a cognitive modeling problem. Clearly, however, prioritizing requires an ability to predict whether or not a user is likely to evaluate a message as a good message, and so requires an effective model of human decision making to be successful.

Using cognitive models for prioritization does not only promise to provide an answer to an applied problem, but also has theoretical benefits for the more general study of human decision making processes. This is because, in the form of real-world e-mails, it deals with a richly structured stimulus domain. There are, of course advantages in studying decision making with artificial stimuli, as is often done in the categorization and classification literature (e.g., Shepard, Hovland, & Jenkins 1961), because of the experimental control that is achievable. A central argument of ecological approaches (e.g., Simon 1956; Gigerenzer & Todd 1999), however, is that it is also important to consider the role of non-arbitrary stimulus environments in supporting (or confounding) human decision making.

In this paper, we develop and evaluate two cognitive models for prioritization. One is a 'rational' model, that performs exhaustive calculations, while the other is a 'fast and frugal' model, that requires only limited time by making assumptions about the nature of its environment. In the next section, we describe how e-mails are represented by these models, and how information about them is learned. We then describe the two models in detail, before presenting the results of an experiment in which both are evaluated on real-world data. Finally, we draw some conclusions regarding the theoretical implications of the results for understanding human decision making, and the applied implications for building an e-mail prioritization system.

E-mail Representation and Learning

Cues and Cue Validities

We follow previous research in assuming e-mails are represented in terms of a set of binary features, which we call cues. These cues may relate to the content of the e-mail, such as a keyword in the message text, or metadata associated with the e-mail, such as the name of the sender. In this way, each e-mail may be defined by the set of cues that it contains.

Following Gigerenzer and Todd (1999), we associate a cue validity with each cue, which measures the probability that an e-mail will be regarded as good, given that it has the cue. Formally, this means that the validity, v_i of the i-th cue, c_i is defined as $v_i \equiv p(\mathrm{G} \mid c_i)$, where G denotes good. Notice that, because each e-mail is assumed to be either good or bad, $1 - v_i$ gives the $p(\mathrm{B} \mid c_i)$, the probability that an e-mail will be bad when it has the i-th cue.

Learning Cue Validities

Where the cues constitute the representational component of our decision making models, the way in

which the validities are specified constitute the learning processes, in the sense that different validities apply in different environments, and are formed on the basis of information observed in those environments. We consider two methods for learning cue validities, arising from the alternative frequentist and Bayesian statistical approaches. In both cases, we assume that (in ways described later) every e-mail that has previously been processed by a user has been classed as a good e-mail or a bad e-mail. This means that the raw data for the i-th cue take the form of a count g_i, giving the number of good e-mails with the cue, and a count b_i, giving the number of bad e-mails with the cue.

Under the frequentist approach, the validity of a cue is estimated simply as the proportion of good e-mails with the cue:

$$\hat{v}_i = g_i / \left(g_i + b_i + \epsilon\right),$$

where ϵ is a small positive number that ensures cues have a defined validity of zero before they have been observed in any e-mail (i.e., the case $g_i = b_i = 0$).

Under the Bayesian approach, prior beliefs regarding the validity of the i-th cue are modified using the data provided by the counts g_i and b_i. As a cue becomes associated with more good e-mails, higher values for its validity become more likely. Conversely, as a cue becomes associated with more bad e-mails, lower values for its validity become more likely. Bayes' theorem describes the way in which the prior beliefs are modified by data to give a probability distribution over the range $[0, 1]$ of possible validities. Defining the validity of a cue as the mean of this distribution, and assuming a uniform prior, gives the result (see Gelman, Carlin, Stern, & Rubin 1995, p. 31):

$$\hat{v}_i = E\left[p\left(v_i \mid g_i, b_i\right)\right] = \frac{g_i + 1}{g_i + b_i + 2}.$$

As more e-mails with the i-th cue are processed the counts g_i and b_i increase, and the frequentist and Bayesian approaches converge towards the same value. When few data are available, however, we later show that the Bayesian approach has advantages for prioritization.

Decision Models for Prioritization

The 'Rational' Approach

Under the rational approach to decision making used here, the evidence provided by every cue associated with an e-mail is integrated to give an estimate of the overall log odds that the e-mail is good, as opposed to bad. Assuming that the evidence provided by each cue is independent, and that the prior probabilities of an e-mail being good or bad are equal, then Bayes' theorem gives:

$$\ln \frac{p\left(\mathrm{G} \mid c_1, ..., c_n\right)}{p\left(\mathrm{B} \mid c_1, ..., c_n\right)} = \sum_{i=1}^{n} \ln \frac{p\left(c_i \mid \mathrm{G}\right)}{p\left(c_i \mid \mathrm{B}\right)}.$$

The required evidence ratios $p\left(c_i \mid \mathrm{G}\right) / p\left(c_i \mid \mathrm{B}\right)$ can be estimated from the data in the same way as cue validities, or (with some manipulation) written in terms of the validities themselves. The rational approach has the attraction of considering all of the data, in the sense that it considers the evidence provided by every cue associated with every stimulus. For this reason, it is often considered a normative account of decision making, and has been used extensively (in one form or another) to model human decision making. As an (arbitrary) example, consider Kruschke's (1992) well known ALCOVE model, which uses a weighted sum of the evidence provided by each dimension of a stimulus in deciding whether or not that stimulus belongs to a category. The rational approach is also widely used in machine learning, and has been applied in previous research (Macskassy *et al.* 1999; Mehran *et al.* 1998) on prioritizing e-mails.

The 'Fast and Frugal' Approach

In developing their 'fast and frugal' approach to modeling human decision making, however, Gigerenzer and Todd (1999) challenge the rational approach. They argue that because human decision making processes evolved in a competitive environment, they need to be fast, and because they evolved in a changeable environment, they need to have the robustness that comes from simplicity. To meet these challenges, the fast and frugal approach adopts Simon's (1982) notion of 'bounded rationality', and models human decision making using simple algorithms that rely on an assumed structure in the stimulus environment to function effectively.

For example, in an environment where the validity of one stimulus cue is highly predictive of the validities of the remaining cues, and the examination of additional cues is an effortful process, it is sensible to consider only the first cue. Similarly, in an environment of diminishing returns, where the examination of each successive cue provides less information than previous cues, it makes sense to base decisions on a small number of cues. Gigerenzer and Todd (1999) show that many real-world stimulus domains have these sorts of structures, and develop a number of cognitive models—including the 'Take the Best' model of forced choice, the 'QuickEst' model of value estimation, and the 'Categorization by Elimination' model of categorization—that make inferences by assuming environmental regularities.

Unfortunately, none of these models is directly applicable to the problem of e-mail prioritization, and so

we developed a new model using the basic 'fast and frugal' modeling approach. Gigerenzer and Todd (1999) argue that their models of human decision making are based on simple mechanisms that answer three fundamental questions:

- How should a stimulus environment be searched for information?
- When should this search for information be terminated?
- Once the search has been terminated, what decision should be made given the available information?

In the context of finding good unread e-mails, as required for priorization, it is not difficult to provide answers to these questions:

- Unread e-mails should be searched in terms of cues, looking for e-mails with high validity cues.
- The search should be terminated as soon as at a candidate good e-mail has been identified. Since users process e-mails serially, there is no benefit in seeking to sort the unread e-mails, beyond attempting to ensure that at any time the top-most e-mail is the one most likely to be good.
- The best available e-mail should be placed at the top of the inbox, as the next one to be read by the user.

These answers suggest a simple fast and frugal decision model for prioritization. The cues are ordered in terms of their estimated validity and, starting with the highest validity, a search is made for an unread e-mail that has this cue. If this search is successful, the process terminates without considering any further cues. If no e-mail is found, the search continues using the next highest validity cue, and this process is repeated until an e-mail is found. This model is closely related to Take the Best, and belongs to the class of what Gigerenzer and Todd (1999) term 'one reason decision making' models. Only one reason, in the form of the presence of a high validity cue, is all that is required to find the next e-mail for presentation.

Experiment

Data Collection

We developed a macro for the Microsoft Outlook e-mail application that unobtrusively logged the behavior of one user for a period of 76 consecutive days. This logging involved recording the actions made by the user in reading, responding to, and organizing the e-mails in their inbox. Every time the user replied to, forwarded, saved, moved or deleted an e-mail in their inbox, an entry in a log file was made. Often, a particular e-mail was subjected to several processing actions over time, such as being forwarded, and then deleted. We used an operational definition of what makes an e-mail a good or bad one, based on these processing sequences, to enable each e-mail to be labeled as either good or bad. If an e-mail was *ever* printed, forwarded, replied to, copied to another folder, or saved, it was regarded as a good e-mail. Otherwise, when the e-mail was only deleted, it was regarded as bad. Of course, this sort of operational definition is not unproblematic, but we believe it represents a reasonable first-order approximation for identifying those e-mails that are more 'important', 'useful' or 'urgent'.

Table 1: The logged e-mail properties used to generate cues, together with a sample cue.

Property	**Sample Cue**
Attachments	"AttachmentCount=2"
CC list	"CC=mark@adelaide.edu.au"
Flag status	"FlagStatus=0"
Importance	"Importance=1"
Sender's e-mail	"SendersEmail=jdl@mbox.com"
Sender's name	"SendersName=John Lee"
Subject keyword	"Subject=upgrade"
Addressee list	"To=Ben Stamley"

The properties of the e-mails being manipulated were also recorded, providing the subject text of the e-mail (for ethical reasons no message content was recorded), as well as metadata identifying the sender of the e-mail, whether it had an attachment, and so on. Those properties taking discrete values were used to generate the cues for representing e-mails by pairing the property with each possible value. Table 1 details the eight properties used in this way, together with an example of a cue for each. The only pre-processing used in generating these cues was to remove common English words from the subject text using a 'stopword' list. The final data set contained 886 e-mails, 362 of which were good, defined in terms of 3,112 binary cues.

Effectiveness of Prioritization

Both the rational and the fast and frugal models were applied to the e-mail data, using the Bayesian learning approach. Each day's e-mails were prioritized in sequence, to simulate the effect that prioritization would have if it were implemented on-line. Figure 1 summarizes the results of 10 independent applications of each method using an effort-reward graph. The performance curves relate hypothetical levels of 'effort', which describe the proportion of available e-mails processed by the user, to the resultant level of reward, as measured by the proportion of available good e-mails that are found. Mean performance levels are shown by the curve, with best- and worse-case perfor-

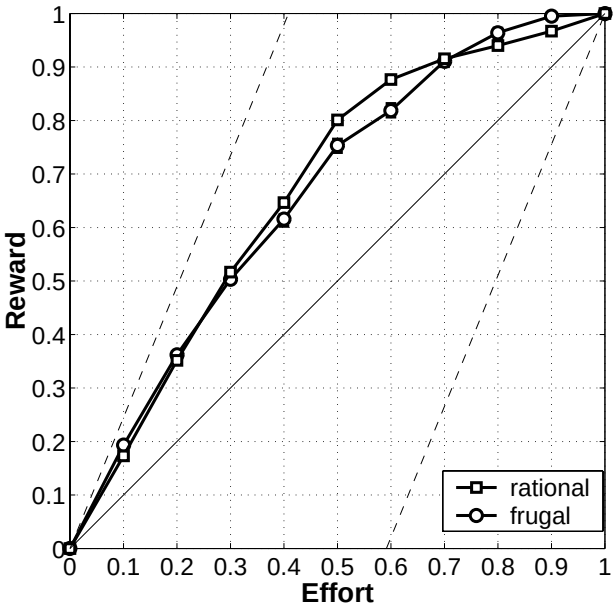


Figure 1: Effort-reward performance of the rational and 'fast and frugal' decision models.

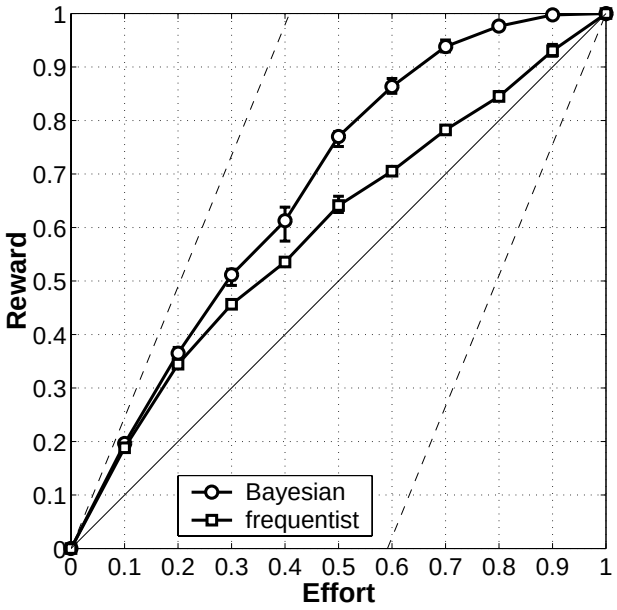


Figure 2: Effort-reward performance of the 'fast and frugal' model using Bayesian and frequentist learning.

mance, arising from the stochastic process of breaking ties, indicated by error bars (where large enough to be visible).

Without prioritization, good e-mails are evenly distributed according to their base-rate of occurrence, which corresponds to the diagonal line in Figure 1. The best- and worse-case possible effort-reward performance of prioritization are shown by the dotted lines, which correspond, respectively, to the cases where all good e-mails are presented first, and where all bad e-mails are presented first. Figure 1 shows that the rational and the fast and frugal model perform very similarly. They are close to optimal for the first 10-20% of good e-mails, but then perform less impressively, although they continue to provide a significant advantage over non-prioritized presentation. Reading the first 50% of e-mails, for example, results in finding approximately 75-80% of the good e-mails available.

Figure 1 suggests two important conclusions. Firstly, it shows that prioritization is effective, which suggests that human decisions in processing the e-mails have some level of systematic relationship with the various cues by which the e-mails are represented. Second, the fast and frugal approach is approximately as effective as the rational approach, which suggests that the human decision making process can be understood in terms of the identification of key features of the e-mails, rather than the exhaustive integration of all of their properties.

Bayesian and Frequentist Learning

An important theoretical problem for prioritization relates to the balance between exploitation and exploration processes. In the context of e-mail prioritization, exploitation involves using cues that are known to have some validity to find good e-mails, while exploration involves learning more about cues for which little or nothing is known, in the hope of find new sources of good e-mails. Prioritization algorithms are of limited use if they achieve their results by exploitation at the expense of exploration, particularly in dynamically changing environments. For this reason, there has been some considerable effort in the machine learning literature (see Sutton and Barto 1998) to balance the competing demands of exploitation and exploration, usually by introducing some stochastic element into the search process.

As it turns out, the Bayesian approach to learning validities addresses this problem. Figure 2 shows the effort-reward performance of 10 runs of the fast and frugal model using both the Bayesian and the frequentist approaches. To assist in the exposition of our subsequent analyses, only a limited set of cues, consisting of all of those generated from the easily understood 'Senders Name' field were used. As Figure 2 shows, the Bayesian approach performs better, particularly for effort levels greater than about 0.5.

The reason for the superiority of the Bayesian validity estimate can be demonstrated through a concrete example. On day 43, a (small) total of five e-

Table 2: Sender cues, good (G) and bad (B) counts, and estimated Bayesian and frequentist validities for day 43.

Sender's Name	G	B	Bayes	Freq.
ABC News Online	1	140	0.01	0.01
Scott Brown	1	0	0.67	1.00
Tapes Subliminales	0	0	0.50	0.00
Virtual Florist	0	1	0.33	0.00
W. Paul Malcolm	0	0	0.50	0.00

mails required prioritization, coming from five different senders. Of these senders, three had previously sent e-mails: "Scott Brown" had sent one good e-mail, "Virtual Florist" had sent one bad e-mail, and "ABC News Online" had sent 141 e-mails, only one of which had ever been good. These patterns of good and bad counts, together with their Bayesian and frequentist cue validity estimates, are shown in Table 2.

Under the frequentist approach, the "Scott Brown" e-mail will be presented first, because it has been associated with the highest proportion of good e-mails. The next e-mail presented will be the "ABC News Online" e-mail, because it has the next highest estimated validity, by virtue of being the only other sender ever to provide a good e-mail. The remaining two unknown senders have estimated validities of zero, and so their e-mails will be presented in random order. As it happens, one of these e-mails, from the new sender "W. Paul Malcolm" is a good one, and so prioritization will be ineffective. Fundamentally, this is because frequentist validity estimation favors the exploitation of sources with very limited returns over the exploration of unknown sources.

Using the Bayesian approach, the "Scott Brown" e-mail will again be presented first, because it has the highest estimated validity. However, "Virtual Florist" and (especially) "ABC News Online" e-mails will not be presented until after those from the senders about whom nothing is known, because their validities are below the 0.5 prior. In this way, the potential new sources of good e-mails will be explored before those that are known to have limited returns are exploited. Notice also that the "Virtual Florist" e-mail will be presented before the "ABC News Online" e-mail, because less data are available for estimating the validity of the former, and so it has more scope to achieve a higher estimate as more observations are made (i.e., it is more worthy of further exploration). Finally, we note that the situation with many 'spam' e-mails is naturally handled within the Bayesian approach by changing the prior on good e-mails.

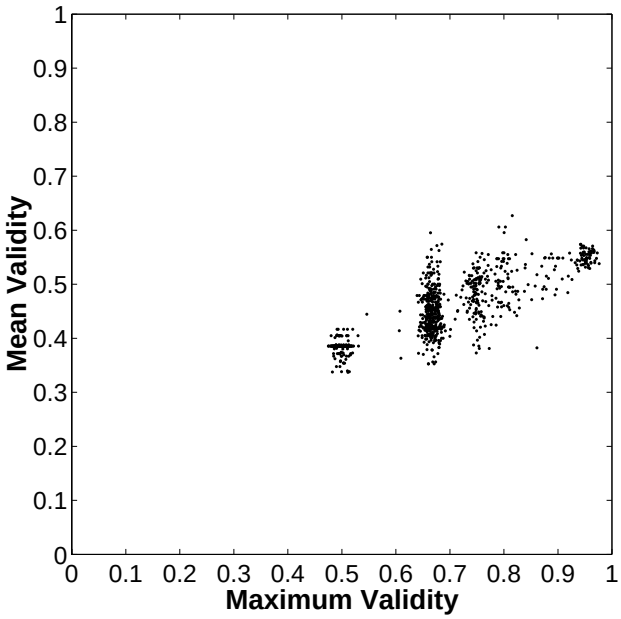


Figure 3: The relationship between the maximum cue validity for an e-mail, and its mean cue validity.

The Structure of the Environment

An analysis of the e-mail stimulus domain explains why the fast and frugal approach performs similarly to the rational approach. Figure 3 shows the relationship between the mean estimated validity of the cues associated with each message (using Bayesian learning), and the maximum estimated validity. There is a positive correlation of $r = 0.80$ between these measures, indicating that the maximum cue validity, as used by the fast and frugal method. is highly predictive of the validities of the remaining cues considered by the rational method. This environmental regularity is the reason for the success of the fast and frugal model: By finding the unread e-mail with the greatest cue validity, it does not need to consider further cues, because their validities are largely already determined by the maximum value.

Future Work

The outstanding problem relates to adaptation. If the characteristics of the external e-mail environment change (e.g., people send different types of e-mails), or the user changes the way they regard e-mails as good or bad, prioritization needs to reflect the new situation. The learning processes used in our study will be slow to adapt to these sorts of changes, as demonstrated for the Bayesian approach by the pattern of change of the five cues shown in Figure 4. Validities for the "To=Mike Lee" and "Subject=newmail" cues are learned effectively, because

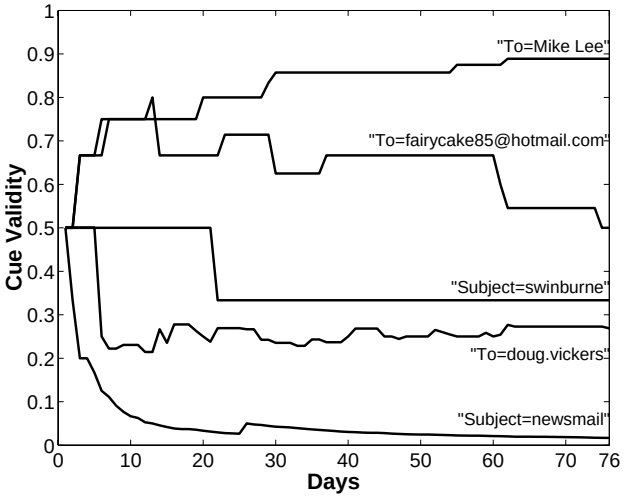


Figure 4: The pattern of change over all processing days for five cues, using Bayesian learning.

they are consistently evaluated by the user. The "To=fairycake85@hotmail.com" cue, however, is evaluated as good in the first two weeks, but its change to a bad cue is learned slowly. Meanwhile, the cues "Subject=swinburne" and "To=doug.vickers" have similar estimated validities at day 76, yet there are grounds to be more confident about the accuracy of the latter, since it is based on a significant volume of recent data, while the former has not been seen since about day 22.

The ability to adapt requires that memory processes be introduced into the cognitive decision models. By replacing old information in the counts g_i and b_i with new information, giving greater weight to new information, or forcing information to decay over time, validity estimates will be based on data that reflects the current state of affairs. A variety of memory mechanisms have been developed for simple psychological decision models (e.g., Pietsch & Vickers 1997), and their detailed empirical evaluation is a priority for future research. The other necessary area of future research is to extend our evaluation to a larger number of users.

Conclusion

We argued in the Introduction that using cognitive decision models to prioritize e-mails provided a way to address an applied problem, and also advance our theoretical understanding of human decision making. We conclude by suggesting some implications of our results on both the applied and theoretical fronts.

In terms of developing an e-mail prioritization application, the fast and frugal model has significant potential. The data required to drive the algorithm, in the form of user evaluations of good and bad e-mails, is done entirely unobtrusively, does not require any additional user effort, and provides a continual on-line data source that should allow for adaptation. The balance between exploration and exploitation is handled naturally by the Bayesian approach to validity estimation, and the the fast and frugal algorithm scales well to large problems. Only one e-mail with one cue needs to be found at each stage of prioritization, as compared with the rational approach, which examines every cue of every e-mail at every stage.

Theoretically, our results suggest that human decision making in processing e-mails can be understood in terms of a one reason decision making process that is tuned to regularities in its environment, and so supports Gigerenzer and Todd's (1999) fast and frugal approach to cognitive modeling. The Bayesian approach to validity estimation also provides a theoretical tool for any learning or decision making situation where exploration must be balanced with exploitation, and could be used in other cognitive decision models.

Acknowledgments

This work was supported by the Australian Defence Science and Technology Organisation. We thank Helen Braithwaite, Brandon Pincombe, Kenneth Pope, Florian Sollich, Douglas Vickers, Michael Webb, and Chris Woodruff.

References

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gigerenzer, G., & Todd, P.M. (1999). *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Macskassy, S. A., Dayanik, A.A., and Hirsh, H. (1999). Emailvalet: Learning user preferences for wireless email. In *Proceedings of Learning about Users Workshop, IJCAI'99*.

Mehran, S., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In *AAAI-98 Workshop on Learning for Text Categorization*.

Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information processing mechanism. *Journal of General Psychology, 124*, 229–339.

Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classification. *Psychological Monographs, 75*(13), 517.

Simon, H.A. (1982). *Models of Bounded Rationality*. Cambridge, MA: MIT Press.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Is Concept Formation An Age-Independent Process?

Kenneth R. Livingston (livingst@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 479
Poughkeepsie, New York 12604

Janet K. Andrews (andrewsj@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 146
Poughkeepsie, New York 12604

Emily Kuschner (emkushner@vassar.edu**)**
Department of Psychology
Vassar College, 124 Raymond Avenue, Box 1471
Poughkeepsie, New York 12604

Abstract

We present the results of a study examining the effects of category learning on the performance of five year-old children and adults on similarity judgment and same-different tasks. Participants in the learning condition learned to distinguish two kinds of invented alien stimuli by hearing an interactive story over the course of two days, at the end of which they performed three tasks. A comparison of their performance with control participants revealed a marked expansion effect in both children and adults, with learning groups judging between-category pairs to be more different than control groups did. There was no compression effect (within-category pairs were not judged as more similar by learning than control groups). We hypothesize that expansion occurred because distinguishing pairs of stimuli was difficult, as indicated by a high error rate on the same-different task for both child and adult participants.

Introduction

Nearly a decade of research now suggests that the space of similarities within which we locate objects undergoes a systematic change in metric structure in the course of category learning (e.g., Beale and Keil, 1995; Goldstone, 1994a; Goldstone, Lippa, and Shiffrin, 2001; Livingston, Andrews, and Harnad, 1998). This result contrasts sharply with the view of similarity taken for granted in classical descriptions of the category learning, where it is assumed that the metric of psychological similarity is fixed, with the result that the locations of objects within that space, and thus their relationships to one another, are entirely determined by their perceptual properties (e.g., Bruner, Goodnow, and Austin, 1956; Hutchinson and Lockhead, 1977). The more recent work suggests that the process of category learning itself may actually alter the similarity space and thus the representational structure of our categories.

Two different kinds of changes to psychological similarity space have now been documented in the literature. *Compression* occurs when one region of the n-dimensional space of similarities changes such that items falling within that region come to have more nearly equivalent encodings than they did prior to category learning. This pattern has been observed by Livingston, et al. (1998), and Kurtz (1996), for example, and manifests as (1) an increase, following category learning, in similarity ratings among items drawn from the same category as compared with items drawn from different categories, or, (2) as greater confusability among items drawn from the same category than among those drawn from different categories. In neural network simulations, the change has been measured directly as an increased similarity in activation patterns on hidden units in a simple feedforward network (Harnad, Hanson, and Lubin, 1995).

The other pattern of change in similarity space following category learning, called *expansion,* occurs when a region of the space of similarities changes such that items falling within that region are judged to be more different after category learning than prior to it, or are less confusable in a same-different task. This pattern has been extensively documented by Goldstone (1994a; 1994b; 1996; see also Goldstone, et al. 2001). In neural network simulations, the change has been

measured directly as an a greater *dis*similarity in activation patterns on hidden units in a feedforward network (Harnad, et al. 1995; Tijsseling and Harnad, 1997).

In theory, both kinds of changes could occur in the course of category learning, but in general only one pattern is typically observed for a given set of stimuli. Research is currently ongoing to establish the conditions under which one observes compression versus expansion. One hypothesis under active investigation is that expansion is observed in those cases where the discrimination among exemplars in the training set is perceptually difficult, which results in discrimination learning. Compression, on the other hand, occurs when no difficult perceptual discrimination is required. What is important to note, regardless of the ultimate fate of this hypothesis, is that either compression or expansion is sufficient to produce the effect necessary for the psychological distinctiveness that characterizes concepts: a set of similarity relationships that sets the members of the category apart from non-members by its *relatively* greater degree of intra-category similarity (or, alternatively, inter-category dissimilarity).

It has been suggested by many of the researchers who have studied compression-expansion effects that the process may be so fundamental to category learning that it constitutes a basic mechanism by which abstract and universal representations (concepts) are formed (Damper and Harnad, 2000; Goldstone, 1996; in press; Livingston, et al, 1998). If this contention is correct, then evidence for the operation of this process should be found among young children as well as in adults. To count as truly fundamental to the process by which perceptual categories are built, it should not turn out that compression-expansion effects reflect a strategy acquired late in life or taking a long time to develop. Indeed, it does not appear that there is anything consciously strategic about the process at all; it seems to reflect the operation of an automatic recalibration of psychological similarity space in response to the discovery, during category learning, that a set of items needs to be partitioned in a consistent way. Nevertheless, evidence that this process operates in young children as well as in adults would strengthen the claim that it constitutes a basic mechanism of category learning.

Certainly there is little doubt that children and young infants can learn to make category distinctions, at least among perceptual categories of the kind at issue here (e.g., Quinn, Slater, Brown, and Hayes, 2001). There is also a growing, if still controversial, body of literature concerning the ability of young children to make use of information about function (e.g., Rakison and Cohen, 1999) or internal, inferred features (Gutheil, Vera, and Keil, 1998) when learning new categories or assigning novel objects to existing ones. There seem to be many similarities between the processes of concept formation in children and adults. To date, however, there has been no successful demonstration that children's category learning is characterized by compression-expansion effects (but see Katz, 1963 for suggestive findings).

The major purpose of the research reported here is to test the hypothesis that the category learning of children will show patterns of compression and/or expansion similar to those already observed among adults. In addition, the present study presents an opportunity to compare performance on similarity judgments with performance (errors and response times) on a same-different discrimination task. The similarity task may be more sensitive to the effects of category learning than the same-different task, but its conceptual complexity makes it difficult to use with children younger than five. Evidence that the same-different task can capture the effects of category learning would clear the way for future work with younger children.

The limited attention spans of young children necessitated the development of a more elaborate training and testing procedure than is needed with adults. Extensive pre-testing was required to design a story-based category learning task and engaging tasks for the testing process. Pilot studies revealed that the procedures are too demanding for children younger than five years of age, and even for older children must be spread across sessions on two consecutive days. Rather than rely on an implicit comparison to the adult literature, we included an adult sample that followed the same procedures used with the children.

Method

Participants

Participants were 27 kindergarten children between the ages of five and six, and 23 Vassar College students participating through an introductory psychology research requirement. Participants in each age group were randomly assigned to the learning or control conditions.

Stimuli

The stimuli were designed to resemble friendly-looking alien creatures and varied on the dimensions of torso width and arm length. Figure 1 shows stimuli with extremes on these dimensions; intermediate values were defined at equal intervals between extremes. All stimuli had yellow bodies, green feet, blue hands, and a pink nose.

For the learning condition, two categories were created and identified by the nonsense labels *Fip* and *Zug.* The Fips had longer arms and narrower torsos, while the Zugs had shorter arms and wider torsos. For each category there were three possible values on each dimension, for a total of nine possible members of each category. Of the eighteen different possible stimuli, fourteen (seven in each category) were used in the experiment. Stimuli were printed out on yellow paper, laminated, and glued onto felt with a black oval-shaped background. A 155-cm X 74-cm board covered in black felt served as the background for the story. To enhance the interaction of the children with the materials and make the story more interesting, felt props were also used. These props represented various objects and devices described in the story. For instance, when it was explained that the Zugs trained by lifting moon rocks and eating a diet of fuzzy pickles and purple pretzels, participants would be asked to place moon rocks, fuzzy pickles, and purple pretzels alongside the Zugs.

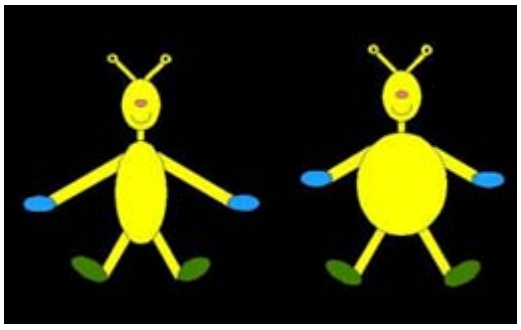


Figure 1. Examples of the stimuli.

Procedure

The learning condition. Those who learned to distinguish Fips and Zugs did so over the course of two, one-on-one sessions with an experimenter. Learning occurred in the context of a story, told using the large felt board and felt props, about two teams of four aliens, the Fips and the Zugs, who compete in an alien Olympics competition. The story is designed to hold a young child's attention so as to allow the experimenter to highlight the differences between the two categories in the context of an interactive dialogue rather than by direct instruction. For example, the long-armed Fips more easily get tangled in a cargo net while climbing, but their narrow torsos make them less likely to get stuck in an obstacle course. Interactivity is introduced by inviting the participant to help construct each scene. At several points in the script the participant is also asked to sort the aliens into categories, to allow provision of feedback as learning progressed.

On the second day of the learning condition, the story was continued. It concludes with a final competition, which results in a tie. The participant is told that he or she will get to stage one last game to settle the tie, but that first there are some other games to be played. These other games are the three primary data gathering tasks.

The participant gave similarity judgments for all fifteen possible pairs of six aliens, which include four that the participant had learned to categorize during the story telling (two from each category) and two not seen before (one from each of the two categories). Pre-testing indicated that fifteen judgments is an upper limit on five-year-olds' attention. The novel stimuli provide a check on whether what has been learned is a generalizable category. In the similarity judgment procedure one picture is placed at the left end of a long felt strip marked off into distinct intervals that allowed scores from 0.5 to 8.5 in 0.5 intervals. The participant is asked to place the other item according to how similar it was to the first item, with more proximal placement indicating greater similarity. The participant is trained on the task using pictures of different breeds of dogs. Pre-training continued until the judgments were being made reliably and with confidence. Once the system was understood, we presented the fifteen pairs of aliens. The experimenter recorded the judgment by reading the position of the center of the stimulus in relation to the marks on the strip.

In the second task, participants viewed twenty-one pairs of stimuli, presented simultaneously on a Macintosh Powermac G3 or Powerbook G3 using SuperLab Pro 1.75 software. The same six stimuli used in the similarity judgment task were used here as well. In addition to the fifteen pairs presented in that task, an additional six pairs were presented, comprised of each of the six stimuli presented with its identical twin. Careful training using pictures of flowers ensured that participants understood that a "same" response

required the stimuli to be identical. Participants answered by pressing clearly marked keys on a keyboard. A colorful feedback screen indicated whether the response was correct. These screens were designed during pre-testing so as to assure that the children wanted to produce the "correct" screen and did not like the "incorrect" screen.

For the third and final task each participant was asked to sort a slightly larger set of fourteen stimuli into two groups, the Fips and the Zugs. To the eight stimuli used in the story, and the two added during testing, we added four more, two from each category. Pilot studies suggested child sorting becomes unreliable when more than fourteen stimuli were included. This final task provides data concerning whether participants in the "learning" condition actually did learn the category distinction, and if so how well they extend the concept to new instances.

The control condition. Participants in the control condition performed the same three tasks as those in the learning condition but did not learn the story or receive any information about categories or types of aliens. Because they had not learned to categorize them, we could not refer to them by name. The only change in the tasks required by this difference was to the sorting task instructions, which simply asked that the aliens be put into two groups according to which ones seem to go together.

Results

Similarity judgments. A 2 (age: child vs. adult) by 2 (group: learning vs. control) by 3 (pair type: Fip-Fip, Fip-Zug, Zug-Zug) analysis of variance with repeated measures on the third variable yielded a highly significant main effect of pair type ($F(2, 90) = 81.140$, $MSE = 1.004$, $p < .0001$) and a highly significant interaction between condition and pair type($F(2, 90) = 10.473$, $MSE = 1.004$, $p < .0001$). The between-category pairs were judged overall to be less similar than the within-category pairs, and the between-category pairs were also judged to be less similar by the learning groups than by the control groups, a clear case of expansion at the category boundary following learning. There is no interaction with age (see Figure 2.) No other effects were statistically significant.

Same-different judgments. This task yielded two dependent measures, proportion of errors and mean response time. Single-sample t tests demonstrate that all four groups performed the same-different task better than chance. A 2 (age:

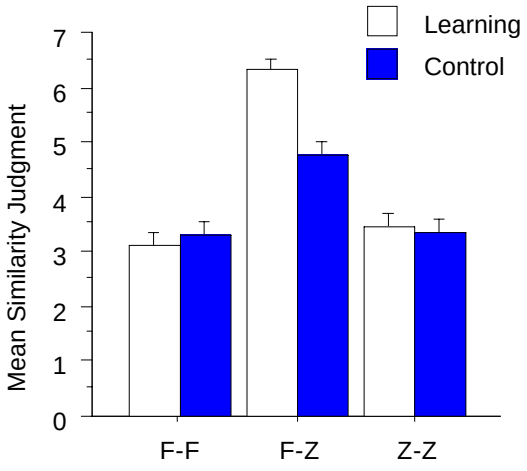


Figure 2. Mean similarity judgments by both learning and control groups for each of three pair types (FF, FZ, ZZ). Higher numbers indicate greater dissimilarity

child vs. adult) by 2 (group: learning vs. control) by 3 (pair type: identical, same category non-identical, different category) analysis of variance with repeated measures on the third variable on the proportion of errors yielded a significant main effect of age ($F(1, 46) = 4.611$, $MSE = .045$, $p < .04$) and a highly significant main effect of pair type ($F(2, 92) = 23.095$, $MSE = .027$, $p < .0001$). Children made more errors than adults (.221 vs. .146) and different category pairs produced fewer errors (.057) than identical pairs (.218) or same category non-identical pairs (.275). No other effects were statistically significant.

A 2 (age: child vs. adult) by 2 (group: learning vs. control) by 3 (pair type: identical, same category non-identical, different category) analysis of variance with repeated measures on the third variable on the response times yielded significant main effects of age ($F(1, 45) = 8.823$, $MSE = 12859202$, $p < .005$) and pair type ($F(2, 90) = 5.399$, $MSE = 1932184$, $p < .007$), and a significant interaction of age and pair type ($F(2, 90) = 7.748$, $MSE = 1932184$, $p < .001$) . Adults were significantly faster than children overall (2679 msec. vs. 4577 msec.), but this difference was due entirely to the non-identical pairs (both same and different category), which were also faster overall than the identical pairs. No other effects were statistically significant.

Sorting task. An item was considered correctly sorted if it was placed with the majority of the items of its category. This allows characterization of sorts by control participants as correct or incorrect. If one of the groups was

sufficiently larger than the other at the completion of the sorting, it might contain a majority of items from both categories. In that case, the larger majority was said to define the category and thus what counted as correct and incorrect in the two categories. A 2 (age: child vs. adult) by 2 (group: learning vs. control) analysis of variance on the number of items incorrectly sorted yielded significant main effects of age ($F(1, 45) = 9.032$, $MSE = 2.603$, $p < .005$) and condition ($F(1, 45) = 13.683$, $MSE = 2.603$, $p < .001$). The interaction approached significance ($F(1, 45) = 3.846$, $MSE = 2.603$, $p < .06$). Children made more errors than adults and control groups made more errors than learning groups, with control children making by far the most errors. Single-sample t tests demonstrate that only the children in the control condition performed this sorting task no better than chance.

Discussion

The finding, based on similarity judgments, that both adults and children in the learning condition show the same pattern of expansion at the category boundary when compared with participants in the control condition is consistent with the idea that changes to the metric of similarity space may mediate concept formation in an age-independent fashion. The results thus provide encouragement to seek similar evidence from work with still younger children, and to pursue that idea that adjustments to the metric properties of similarity space constitute a general phenomenon in category learning. Unfortunately, the failure to find evidence for expansion using the same-different task suggests that this procedure is not a good candidate for extension to younger ages. We had hoped that there would be differential changes in speed of responding between experimental and control groups, even in the absence of differences in errors, but found none of the necessary interaction effects for that measure either. Clearly, other task candidates, like the match-to-sample technique (e.g., Smiley and Brown, 1979), will have to be explored. At least one finding from the same-different task bears noting, however. The fact that identical pairs and different pairs from within the same category produced the same high level of errors for both adults and children (over 20%) suggests just how difficult the discriminations were between items, and is at least consistent with the hypothesis that expansion effects at the boundary reflect perceptual discrimination learning rather than solely higher-order cognitive changes (Livingston, et al., 1998).

One of the more interesting theoretical -- and empirical, for that matter -- questions going forward will be how the operation of a similarity metric modification process like the one described here maps onto other patterns observed in the development of the child's system of concepts. We earlier highlighted the similarities between the concept learning of adults and children, but interesting differences have been noted and discussed in the developmental literature. For example, how does one square a compression-expansion mechanism with variations in criteria for classification, which have been said to shift from thematic to taxonomic (Smiley and Brown, 1979), or perhaps from basic-level taxonomic to thematic and then to superordinate-taxonomic (Gelman, Coley, Rosengren, Hartman, and Pappas, 1998). To address this issue more fully would require a more detailed analysis than is possible here, but two possibilities are immediately apparent. The first is that there is an important difference between perceptual categorization and conceptual categorization (Mandler, 2000), and that the processes we are describing apply only to the former. This is a highly controversial distinction (see the numerous commentaries that follow Mandler's paper), but if correct it would make it all the more important to find ways to pursue evidence for compression-expansion effects in toddlers and infants, for whom high-level conceptual processes are still poorly developed. The other possibility is that there is but a single process, mediated by changes in similarity metrics, and that variations in organizational structure, whether identified as thematic, taxonomic, holistic, analytic, or what-have-you, reflect shifts in the pattern of attention given to objects and events in the world, shifts that establish the basic dimensionality of the similarity space into which objects are sorted on a given occasion.

Conclusion

The successful demonstration of learned expansion in children shows that the modification of psychological similarity space that occurs in adult category learning operates very early in life and may indeed constitute a fundamental mechanism in concept acquisition. We suggest that further work is needed to extend these results to still younger children, and to resolve important theoretical issues about how

the compression-expansion process is related to known developmental changes in concept learning during the childhood years.

Acknowledgments

Our thanks to the Wimpfheimer Nursery School of Vassar College, the Poughkeepsie Day School, the Vassar Undergraduate Research Summer Institute, Krista Garver, Delia Hom, Maria Jalbrzikowski,Elizabeth Kappler, Jennifer Mason, and Erika Strochlic for their assistance with this study.

References

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition, 57*, 217–239.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: Wiley.

Damper, R.I. & Harnad, S. (2000). Neural Network Modeling of Categorical Perception. *Perception and Psychophysics, 62*, 843-867

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development, 63* (1, Serial No. 253).

Goldstone, R. L, Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27-43.

Goldstone, R. L. (1994a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178-200.

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125-157.

Goldstone, R. L. (in press). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, and C. Olson (Eds.), *Perceptual Organization in Vision: Behavioral and Neural Perspectives.* Mahwah, NJ: Erlbaum.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27-43.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum

Gutheil, G., Vera, A., & Keil, F. (1998). Do houseflies think? Patterns of induction and biological beliefs in development. *Cognition, 66,* 33-49.

Harnad, S., Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* Boston: Academic Press, pp. 191–206.

Hutchinson, J. W., & Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 660–678.

Katz, P. A. (1963). Effects of labels on children's perception and discrimination learning. *Journal of Experimental Psychology, 66,* 423-428.

Kurtz, K. J. (1996). Category-based similarity. In G. W. Cottrell (Ed.) *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, 290.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 732-753.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development, 1*, pp. 3-36.

Quinn, P., Slater, A., Brown, E., & Hayes, R. A. (2001). Developmental change in form categorization in early infancy. *British Journal of Developmental Psychology 19*, 207-218

Rakison, D. H. & Cohen, L. B. (1999). Infants' use of functional parts in basic-like categorization. *Developmental Science, 2*, 423-431.

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: A nonmonotonic trend from preschool to old age. *Journal of Experimental Child Psychology, 28,* 437–458

Tijsseling, A. & Harnad, S. (1997). Warping Similarity Space in Category Learning by Backprop Nets. In: Ramscar, M., Hahn, U., Cambouropolos, E. & Pain, H. (Eds.) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization.* Department of Artificial Intelligence, Edinburgh U.: 263 - 269.

Theories and Similarity: Categorization under Speeded Conditions

Christian C. Luhmann (christian.luhmann@vanderbilt.edu)
Department of Psychology, Vanderbilt University
301 Wilson Hall, Nashville, TN 37203 USA

Woo-kyoung Ahn (woo-kyoung.ahn@vanderbilt.edu)
Department of Psychology, Vanderbilt University
301 Wilson Hall, Nashville, TN 37203 USA

Thomas J. Palmeri (thomas.j.palmeri@vanderbilt.edu)
Department of Psychology, Vanderbilt University
301 Wilson Hall, Nashville, TN 37203 USA

Abstract

A largely accepted view in the categorization literature is that similarity-based reasoning is faster than theory-based reasoning. In the current study, we explored whether theory-based categorization behavior would continue to be observed when people are forced to make category decisions under time pressure. As a specific test of the theory-based view to category representation we examined the causal status hypothesis, which states that properties acting as causes are more important than properties acting as effects when categorizing an item. Subjects learned four categories of items composed of three features and learned causal relations between those features. In two experiments we found that participants gave more weight to cause features than to effect features even under rapid response conditions. We discuss implications of these findings for categorization.

Introduction

When posed with categorization tasks in everyday life people recruit information from a variety of sources. In general, previous work on categorization has focused on two sources of information: similarity and theories. One family of categorization theories has centered on the notion of similarity (e.g., Kruschke, 1992; Nosofsky, 1986; Smith & Medin, 1981; Rosch & Mervis, 1975). On this view concept learning and use is based on computing the similarity between an object to be categorized and a stored representation of a category (e.g., exemplars, Nosofsky, 1986; or prototypes, Hampton, 1995).

An alternative view assumes that people have theories that embody relations between properties and influence categorization behavior (Carey, 1985; Keil, 1989; Murphy & Medin, 1985; Rips, 1989). An illustrative example comes from Keil's (1989) discovery experiment. When presented with an animal that had the appearance and behavior of a horse but the insides and lineage of a cow, adults would categorize the animal as a cow. This behavior suggests that lineage has a special status above and beyond perceptual features, presumably reflecting the importance of lineage in our lay theory of biology. Similarly, Medin and Shoben (1988) showed that people would rather accept a square cantaloupe than a square basketball, presumably because "being round" is more central in naïve theories of physics (i.e., the domain in which basketballs are grounded) than in naïve theories of biology (i.e., the domain in which cantaloupes are grounded).

The similarity-based and theory-based views are not necessarily incompatible (e.g., Sloman & Rips, 1998). In fact, many proponents of either view allow for, or even advocate, the operation of both kinds of processes (e.g. Sloman, 1996; Smith & Sloman, 1994). However, these proposals typically put the two views on unequal footing. A persistent bias present in these 'hybrid' models is that similarity-based categorization is primary. For instance, in the developmental literature, it has been argued that theory-based mechanisms cannot precede similarity-based mechanisms in development because theories must be acquired through similarity-based mechanisms (Quine, 1977; Vygotsky, 1962; but see Keil, Smith, Simons, & Levin, 1998). Thus, only after sufficient experience has been obtained may theories be developed and used, amending (or supplanting) similarity-based information.

In addition to the idea that similarity-based categorization is developmentally primary, there is a notion that similarity-based information is accessed more rapidly, and perhaps more automatically, than theory-based information. This assumption may be motivated by the observation that novices (e.g., children) use similarity-based reasoning and thus it is a somehow simpler mode of reasoning (cf. Keil et al, 1998). Smith and Sloman (1994) make this argument explicit by assuming that theory-based reasoning is a type of rule-based reasoning, arguing that rule-based reasoning is, "more analytic and reflective than similarity-based categorization" (pp. 377-378).

To test this assumption Smith and Sloman designed a study to examine the effect of time constraints on theory-based (or rule-based) categorization. Smith and Sloman's subjects performed a forced-choice task in which each item consisted of a description of an object paired with two possible categories (task and stimuli adapted from Rips, 1989). Each item had one response that corresponded to a rule-based (i.e. theory-based) decision and one that corresponded to a similarity-based decision. For example, "Circular object with a 4 inch diameter" could be categorized as a pizza or a quarter. Calling this object a pizza would signify theory-based understanding of the minting process whereas calling this object a quarter would signify a similarity computation because a circular 4-inch object is more similar to quarters that to most pizzas (but see Nosofsky & Johansen, 2000). Rips (1989) found that people tended to choose the theory-based response. However, when asked to respond as quickly as possible, subjects in Smith and Sloman's study failed to reproduce this result. Only when instructed to talk aloud while categorizing did subjects tend to answer in accordance with the theory-derived rules. Thus, Smith and Sloman concluded that a "…possible constraint…is that the situation encourage people to articulate and explain their reasons for categorization, rather than encourage rapid judgments" (p. 383).

One problem with this interpretation is that the use of either similarity or theories resulted the in subjects accepting bizarre objects as category members. For instance, participants had to decide whether a circular object with a 4-inch diameter that is silver colored is a pizza or a quarter. This is a rather strict test of theory-use and may not represent a naturalistic situation in which to test the influence of speed.

More recent studies have suggested that theory-based categorization may be at least as fast as (and perhaps as automatic as) similarity-based categorization. For example, Lin and Murphy (1997) pitted perceptual similarity against knowledge of an object's function during speeded categorization. For instance, an object with a loop was either described as a tool used to hunt animals where the loop is placed around the animal's neck, or a pesticide sprayer where the loop was used to hang it when not in use. Thus, the loop should have been viewed as central to the category in the former condition and more peripheral in the latter condition. Even when category responses had to be made within a one-second deadline, or when the picture of the object to be categorized was presented for only 50ms and then masked, subjects continued to be influenced by domain knowledge (e.g., the object's described function).

Palmeri and Blalock (2000) reported a similar pattern of results. Extending the findings of Wisnewski and Medin (1994), they had subjects categorize drawings supposedly drawn by children described as either "creative" or "non-creative." Subjects were able to categorize using this background knowledge (e.g., by the amount of emotional expression) even when the pictures were shown for only 200ms. Their study demonstrated that theory-use did not require lengthy periods of reflection.

The main goal of the current study was to build upon these recent findings and to examine speeded theory-based categorization at a finer level. As a specific test of the theory-based view we have chosen to examine the causal status hypothesis (CSH; Ahn, Kim, Lassaline, & Dennis, 2000; Ahn, 1998). This hypothesis was developed in response to the valid criticisms that specific mechanisms underlying theory-based categorization had not been explicated. CSH states that features of an object that act as causes in one's domain theory are more important than features that act as effects, ceteris paribus. This measure, referred to as *causal depth*, makes explicit why some features are more central to one's theory than others.

As a test of the CSH, Ahn et al. (2000) provided subjects with novel categories that possessed features at different causal depths. When asked to classify possible category members, each of which was missing a single characteristic feature, subjects rated those missing an effect feature as a better category member than those missing a cause feature. In the current study we employed a similar methodology to test the effects of time pressure on the causal status effect.

Experiment 1

Method

Adapting the paradigm from Ahn et al. (2000), our stimuli consisted of four fictional animals (see Fig. 1). Each animal was described as possessing three features. The features were described as having a causal chain structure such that feature A causes feature B, and feature B causes feature C.

It is crucial to ensure that, in the absence of causal information, the three features did not vary in salience. Otherwise, any obtained causal status effect could not be solely attributed to the causal background knowledge but could instead be attributed to some other factor (e.g., physiological feature versus behavioral feature). To eliminate this possibility we pre-tested the stimuli on a separate set of subjects, using the animal descriptions without the explicit causal information. Subjects were then asked to rate the likelihood of category membership of items missing a single feature (see Fig. 2). The results of this pre-test showed no significant differences between the ratings of items missing the first feature, items missing the second feature, and items missing the third feature (all p's$>.4$). Thus, we concluded that the features were equated for a priori strength.

In the categorization tasks used by Ahn et al. (2000), subjects were allowed to view the animal descriptions (along with the causal structure information) while they were making their category judgments. To allow for

speeded responses, subjects in our study were instead required to learn and memorize the four animals, their features, and the causal relations between the features. First, subjects were given the opportunity to study the description of each animal at the beginning of the experiment. While studying each description subjects were instructed to "write about how you think each feature causes the next," in an attempt to force subjects to think causally about the features (instead of as a simple ordered list). To help subjects further learn the items, they were then presented with 6 blocks of trials, during which they were prompted with the name of one of the animals and were required to select (using a mouse-click) the features of that animal from an array containing the features of all 4 animals. They were required to select those features in the appropriate causal order. Successfully responding to the entire set of animals twice allowed subjects to move on to the next block. In the first two blocks responses were unspeeded, while in the last four blocks responses had 5-second deadlines (any response not meeting the deadline was counted as incorrect). This speeded-learning procedure was added so that the novel causal background knowledge would be sufficiently internalized, thereby approximating real-life lay theories. In addition, on half of the blocks, subjects were asked for the causal relations in the forward order (e.g. A, B, C) and on the other half in the backward order (e.g. C, B, A). The order manipulation alternated across blocks and subjects, always beginning with a forward block.

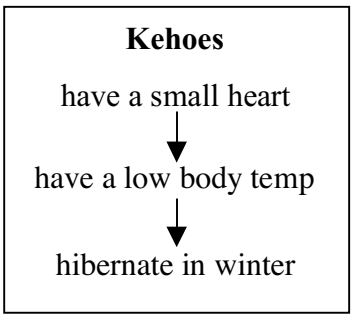


Figure 1: A sample animal with causal links

Once subjects completed these six blocks they proceeded to the experimental transfer task. Subjects were presented with items missing a single feature and were asked to rate the likelihood that the item belonged to its target category on an 8-point scale (with 1 being "Definitely Unlikely" to 8 being "Definitely Likely"). Features of each transfer item were presented in a triad as shown in Figure 2, with the position of the features randomized.

There were 4 blocks of trials in the transfer task. In two of the blocks, subjects were instructed to answer as quickly as possible. In the other two blocks, they were told to take as much time as needed. The speed condition alternated across blocks and was counterbalanced across subjects.

For the unspeeded trials we expected to find results similar to those of Ahn et al. (2000). That is, items missing the terminal effect feature should be rated as more likely category members than those missing the initial cause feature. The critical question was whether this causal status effect would disappear during the speeded trials, as suggested by Smith and Sloman (1994).

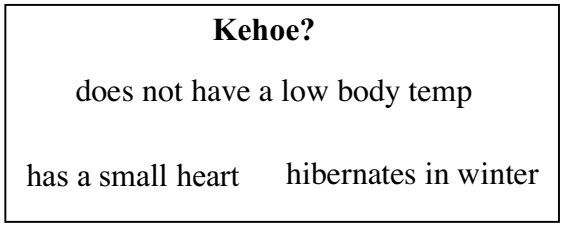


Figure 2: A sample transfer item from Experiment 1

Results and Discussion

Before analyzing the subjects' category ratings, we verified the instructional speed manipulation. The RTs in the speeded blocks (M = 1560ms) were indeed significantly faster than the RTs in the unspeeded blocks (M = 3202ms), $p<.05$, Tukey's HSD.

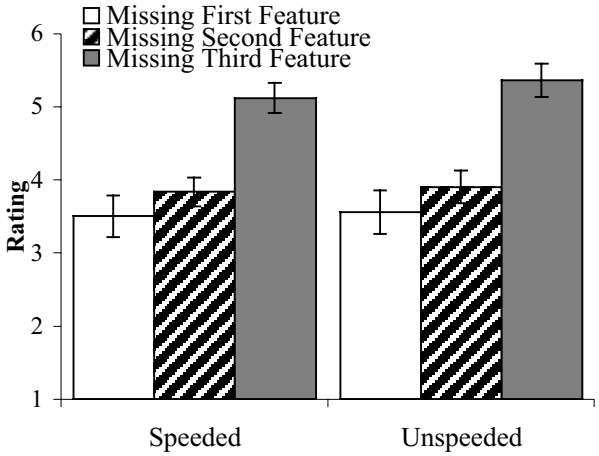


Figure 3: Results from Experiment 1

The results for subjects' categorization responses are summarized in Figure 3. A 2 (speed condition: speeded vs. unspeeded) X 3 (item type: missing first feature vs. missing second feature vs. missing third feature) repeated measures ANOVA was performed on the data. We observed a significant main effect of item type, $F(2, 56)=22.69$, $p<.0001$, demonstrating that subjects categorized in accordance with the causal status hypothesis. In addition, we observed no main effect of speed, $F(1,28)=2.89$, $p>.05$, and the speed X item type interaction was also not significant, $F(2,56)=1.03$, $p>.05$.

Planned comparisons were carried out to examine differences between item types. In both the speeded and unspeeded conditions items missing the third feature were rated significantly higher than those missing the first or second features (p's<.05, Tukey's HSD). The difference between items missing the first feature and those missing the second feature was not significant (p's>.05, Tukey's HSD), possibly because the second feature also served as a cause of another feature, making the difference between the first and the second feature less pronounced (see also Kim and Ahn, 2002).

Overall, these results demonstrate that it is possible to categorize using causal knowledge even when time for lengthy reflection is not allowed. It is tempting to contrast our findings with those of Smith and Sloman (1994). In their experiment, subjects needed unspeeded conditions and to talk aloud while making the judgment in order to demonstrate theory-based behavior. It is possible that the paradigm used by Rips (1989) and Smith and Sloman (1994) created a situation in which theory-use was more difficult to apply than our situation (see above). Nevertheless, our findings clearly question the assertion that theory-use is relegated to situations in which reflection and analytic thought is permitted.

Experiment 2

In Experiment 1, subjects were simply asked to respond as quickly as possible to the "speeded" items. Given this freedom, some subjects responded very quickly but others responded significantly more slowly. Although the speed manipulation we used in Experiment 1 is naturalistic, in that participants carried out what they thought to be a rapid decision making process, forcing participants to respond within a specific deadline would ensure uniform time pressure across all subjects. Therefore, in Experiment 2, we imposed stricter control over subjects' response times by enforcing deadlines on their category decisions.

One methodological complication with establishing appropriate response deadlines is that it is difficult to determine beforehand whether a particular deadline is short enough to challenge the categorization system but not so short as to make accurate responses impossible. That is, if the speeded condition does not show the causal status effect, it can be because theory-based reasoning does not take place during rapid categorization or because the deadline is too short to produce any reasonable responses.

For this reason, we also tested whether similarity information could be used under similar deadlines. By testing both kinds of knowledge, the casual status effects can be compared to similarity-based categorization at each deadline. In this way it can be inferred whether any breakdown of the causal status effect is due to the inability to complete the processes necessary for theory-based categorization or if reasonable responses at that deadline are impossible for both kinds of categorization.

Similarity is frequently calculated based on how many attributes an item has in common with other members of the category (e.g., Tversky, 1977). Therefore, as a similarity-based determinant for feature weighting, we manipulated the relative base rates of each feature within a category (i.e., what percentage of category members possess a feature), a measure also known as *category validity*. In fact, category validity has been shown to be positively correlated with typicality ratings (Rosch & Mervis, 1975).

Experiment 2 contains a similarity condition that provides category validity information in much the same way causal information was provided in Experiment 1. Using this condition as a point of comparison, and with the addition of strict response deadlines, we hope to provide a more rigorous test of the causal status effect under speeded conditions.

Methods

Subjects in the Causal condition were given the same stimuli and accompanying causal information as used in Experiment 1. Subjects in the Base-Rate condition were given the same stimuli but were instead given information about the relative base rates of each feature. Thus, each category was described as having three features (e.g. A, B, and C) such that 100% of category members possessed feature A, 80% of possessed feature B, and 60% possessed feature C. In our parlance, the Causal condition represents a theory-based situation whereas the Base-Rate condition represents a similarity-based situation. Paralleling the results of Experiment 1, items in the Base-Rate condition missing the third (60%) feature should be rated as better category members those missing the first (100%) feature. This is because those missing the third feature share more features with more category members than those missing the first feature.

The learning phase for the Causal condition was identical to that used in Experiment 1. Subjects in the Base-Rate condition did not have to generate explanations but instead categorized exemplars into one of the four animal categories. For this task, each exemplar always possessed the first feature of its category, possessed the second feature on 80% of the trials, and the third feature 60% of the time (thus mirroring the stated base rates). When a given feature did not appear in an exemplar, a feature from one of the other animals was substituted. Feedback was given after each trial.

Blocks of 30 such trials alternated with blocks of the "selection task" used in the Causal condition. The only difference was that features were selected in an order (forward or backwards) dictated by their base rate rather than their position in the causal chain.

The transfer phase for both conditions was nearly identical to that used in Experiment 1 except for a modified speed manipulation. Instead of an instruction to respond quickly, Experiment 2 employed a signal-to-respond technique (Lamberts, 1998). Thus, each trial presented the feature triad (Fig. 2) for a specified duration (see below). When the triad was removed from the screen subjects made their response. If a response was made more than 300ms after the disappearance of the triad, subjects were told to respond more rapidly.

There were four blocks of trials. Each block used one of four presentation durations (1500ms, 750ms, 500ms, and 300ms). These blocks were ordered randomly for each subject.

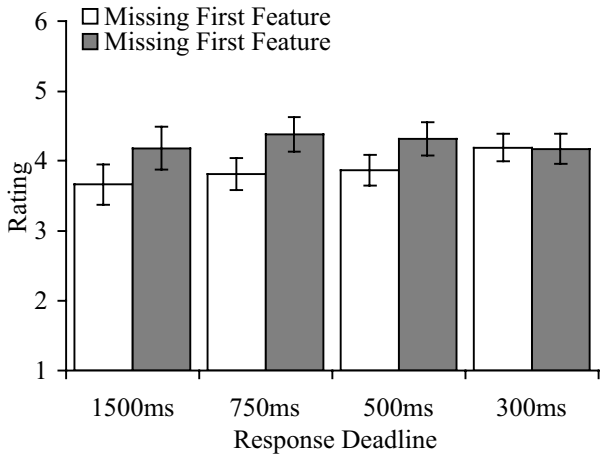


Figure 4: Results from Causal Condition

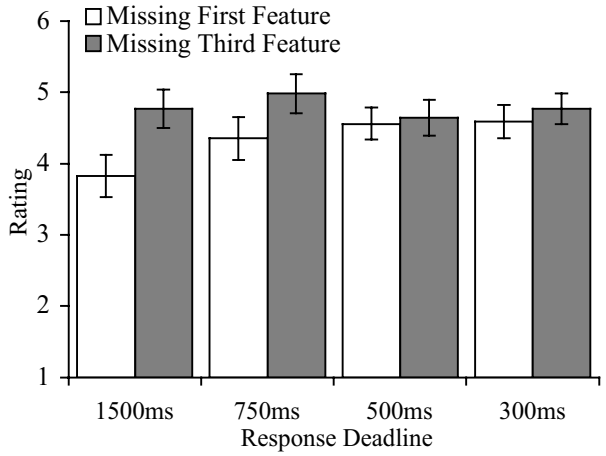


Figure 5: Results from Base-Rate condition

Results and Discussion

The results from the categorization task can be seen in Figures 4 and 5. A 2 (knowledge condition: Causal vs. Base-Rate) X 4 (speed condition: 1500ms vs. 750ms vs. 500ms vs. 300ms) X 2 (item type: missing first feature vs. missing third) ANOVA was performed with repeated measures on the latter two factors. We observed a significant main effect of item type, $F(1, 58)=15.14$, $p<.0005$, that did not interact with knowledge condition, $F<1$, demonstrating that both background conditions had the predicted effect on categorization behaviors. No significant main effect of speed was observed, $F(3, 174)=1.88$, $p>.05$, but this main effect must be interpreted in light of a significant interaction between speed and item type, $F(3, 174)=6.26$, $p<.001$, which will be further examined below. No significant main effect of background condition was observed, $F(1, 58)=3.43$, $p>.05$, and this factor did not interact with either of the other two factors. The three-way interaction between background condition, speed, and item type also failed to reach significance, $F(3, 174)=1.93$, $p>.05$.

Planned comparisons were carried out to determine at what response deadlines the background information had a significant effect on categorization. For simplicity, we only report comparisons between items missing the first feature and those missing the third, the differences that CSH predicts to be the largest. For the Base-Rate condition, items missing the first (100%) feature significantly differed from items missing the third (60%) feature in the 1500ms condition, $t(29)=3.43$, $p<.005$, and the 750ms condition, $t(29)=2.41$, $p<.05$, but not in the 500ms, $t(29)=.3$, $p>.05$, or 300ms, $t(29)=1.59$, $p>.05$, conditions. In the Causal condition, items missing the first (initial cause) feature differed from those missing the third (terminal effect) feature in the 1500ms, $t(29)=2.22$, $p<.05$, the 750ms, $t(29)=2.86$, $p<.01$, and the 500ms conditions, $t(29)=2.06$, $p<.05$, but not the 300ms condition, $t(29)=.81$, $p>.05$.

In light of these results, it is clear that theory-driven causal knowledge can be utilized to categorize stimuli under even faster conditions than those created in Experiment 1. Subjects in this experiment were able to categorize according to their theory even when allowed only 800ms to view the exemplar and make a response. This is impressive considering that this condition was 200ms condition faster than the speeded condition used by Lin and Murphy (1997; Experiment 4) – their stimuli were pictures whereas ours were verbal descriptions. Furthermore, our results indicate that the base rate information, which has been considered a key determinant of similarity (Rosch & Mervis, 1975), did not result in differential responses under this deadline. The results taken together provide strong evidence that theory-based categorization cannot be slower than similarity-based categorization.

Conclusion

The current experiments demonstrated two important findings: First, theory-based categorization, as measured by the causal status effect, did not necessarily require excessive periods of deliberation in order to exert an influence on

behavior. Second, the latencies at which theory-use is possible are comparable to those of similarity-use. These findings bolster the idea that use of similarity is not necessarily more primary than theory-use. Instead, it appears that people's use of theories and similarity may be inexorably intertwined, not just during development (Keil, et al., 1998) but during the course of a single category decision as well.

One open question concerns the type of reasoning that takes place when subjects actually categorize transfer items. Was any causal reasoning taking place during the transfer tasks, or were subjects simply retrieving pre-compiled notions about feature importance derived during learning? Future studies on this issue will help us develop more detailed processing accounts of theory-based categorization. The current results, at the very least, clearly demonstrate that previously acquired causal knowledge influences later categorization judgments even when rapidly categorizing objects.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition*, 69, 135-178.

Ahn, W., Kim, N. S., Lassaline, M. E, & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41, 361-416.

Carey, S. (1985). *Conceptual change in childhood*, MIT Press, Cambridge, MA.

Hampton, J. A. (1995). Testing prototype theory of concepts. *Journal of Memory and Cognition*, 34, 686-708.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, 65, 17-49.

Kim, N. S., & Ahn, W. (2002). The influence of naïve causal theories on lay concepts of mental illness. *American Journal of Psychology*, 115, 33-65.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 695-711.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 1153-1169.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20, 158-190.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 510-520.

Nosofsky, R.M., & Johansen, M.K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375-402.

Palmeri, T. J., & Blalock, C. (2000). The role of background knowledge in speeded perceptual categorization. *Cognition*, 77, B45-B57.

Quine, W. V. (1977). Natural kinds. In Schwartz, S.P. (Ed.), *Naming, Necessity, and Natural Kinds*. Cornell University Press, Ithaca, NY.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity analogical reasoning* (pp. 21-59). New York: Cambridge Univ. Press.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.

Sloman, S. A. & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, 65, 1-15.

Smith, E. E. & Medin, D. L. (1981). *Concepts and Categories*, Harvard University Press, Cambridge, MA.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory and Cognition*, 22, 377-386.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Vygotsky, L. S., (1962). *Thought and language* (E. Hanfmann and G. Vakar, Trans.), MIT Press, Cambridge, MA.

Wisnewski, E. J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18, 221-281.

Case, Word Order, and Language Learnability: Insights from Connectionist Modeling

Gary Lupyan (il24@cornell.edu)
Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology
Cornell University
Ithaca, NY 14853 USA

Abstract

How does the existence of case systems, and strict word order patterns affect the learnability of a given language? We present a series of connectionist simulations, suggesting that both case and strict word order may facilitate syntactic acquisition by a sequential learning device. Our results are consistent with typological data concerning the frequencies with which different type of word order patterns occur across the languages of the world. Our model also accommodates patterns of syntactic development across several different languages. We conclude that non-linguistic constraints on general sequential-learning devices may help explain the relationship between case, word order, and learnability of individual languages.

Introduction

In language acquisition, children are faced with many formidable tasks, yet they normally acquire most of their native language within the first five years of life. One of the most difficult of these tasks involves mapping a sequence of words onto some sort of interpretation of what that sequence is supposed to mean. That is, in order for the child to understand a sentence, she needs to determine the grammatical roles of the individual words so that she can work out who did what to whom. Although the children appear to bring powerful statistical learning mechanisms to bear on the acquisition tasks (e.g., Saffran, Aslin, & Newport 1996), the existence of linguistic universals common across radically different languages (Greenberg 1963) points to the presence of innate constraints on such learning. Without such constraints, it becomes difficult to explain why there are few, if any, Object-Subject-Verb (OSV) languages (van Everbroeck, 1999) even though in principle such a language appears to be as good as any other. In this paper, we propose that these constraints may arise from non-linguistic limitations on the sequential learning of statistical structure, and examine how this perspective may shed light on how children learn to map the words in sentences onto their appropriate grammatical roles. There are two major ways in which languages signal syntactic relationships and grammatical roles—word order (WO), and case markings. In a strict WO language like English, declarative sentences follow a Subject-Verb-Object (SVO) pattern. It is the occurrence of the subject in the first position, and the object in the second, that allows the hearer to comprehend who did what to whom. In contrast, languages such as Russian or Japanese allow multiple word orders and rely on case markings to disambiguate subjects from objects. For instance, *Masha lubit Petyoo* (SVO), *Petyoo lubit Masha* (OVS), and *Lubit Petyoo Masha* (VOS) are all grammatical in Russian and all mean *Mary loves Peter* (albeit with different emphases on the constituents), due to the nominative *-a*, and accusative *-u* case markers.

While long-standing theories describe acquisition of language through an innate language acquisition device (e.g., Pinker, 1995), an alternative approach that is gaining ground is the adaptation of linguistic structures to the human brain rather than vice versa (e.g., Christiansen, 1994; Kirby, 1998). On this account, language universals may reflect non-linguistic cognitive constraints on learning and processing of sequential structure, rather than constraints prescribed by an innate universal grammar. Previous work has shown that sequential-learning devices with no language-specific biases are better able to learn more universal aspects of language as compared to aspects encountered in rare languages (e.g., Ellefson & Christiansen, 2000; Christiansen & Devlin, 1997; Van Everbroeck, 1999, 2001).

Here, we examine the ways in which case markings and word order may function as cues for a sequential learning device acquiring syntactic structure. In simulation 1, we model different word orders, and hypothesize that typologically common languages should be easier to learn by a sequential-learning device than the more rare ones. We expand on this idea in simulation 2 by studying the performance of networks trained on languages of varying degrees of case markings and flexibility. Finally, in simulation 3, we establish that our trained networks are able to mimic syntactic performance of children learning English, Italian, Turkish, and Serbo-Croatian (Slobin and Bever, 1982).

Acquisition of Word Order

Generative linguists have long relied on parameter setting to explain how children acquire the distinct patterns of their native language. For instance, it has been assumed that the way a child knows to generate SVO and not SOV English sentences is through the setting of a VO/OV parameter (Neeleman, 1994). This account has been unsatisfactory because it does not account for many observed correlations; for instance, OV languages typically have flexible word orders (Koster, 1999). More generally, parameter theory has been largely unable to account for the asymmetries and patterns in the distribution of world languages. Why, for instance, are the most common word orders SOV, SVO, and VSO (Greenberg 1963: Universal 1)? Why do verb-final languages almost always have a case system (Greenberg 1963: Universal 41)? And even more fundamentally, why do case languages have flexible word orders to begin with? It is our position that these observations can be at least partially accounted for by examining the learnability of languages from the viewpoint of sequential learning.

Generative linguistics also leaves largely unexplained the process children use to actually set the parameters. With regard to word orders, an explanation espoused by Pinker (e.g., 1995) involves the so-called Subset Principle. According to the Subset Principle, children take the most conservative strategy and so, by default, assume a fixed order. Alternative word orders are only accepted if a child is exposed to these orders, at which time a free word order parameter gets switched on. Under this assumption, FWO languages are predicted to be more difficult to learn. Although the idea that all languages are initially approached as having strict word-order (SWO) was popular in the sixties and seventies (Slobin, 1966), Slobin and Bever (1982 conclude that the primacy assigned to word order was unduly influenced by languages such as English.

There is ample evidence that children learning a strict word-order language such as English never leap to the conclusion that it is a free-word order language (Pinker 1995). While Pinker has used this evidence for reinforcing parameter-setting—the reason children never leap to such conclusions is because a word-order parameter has been set—we suggest an alternative explanation. Simply, children learning English generally do not produce non-SVO sentences because non-SVO sentences are incomprehensible in English. In the absence of case markings, *Kicked John Bill* is ambiguous as to who did the kicking. Children learning English, use the statistical properties of the language to learn that word order is a reliable cue to syntactic relationships. Children learning a case-based language such as Russian, make a similar observation about case markings. This view obviates the need for a default strategy. What is important is that there exist some set of cues to indicate syntactic relationships—there is nothing inherently special about word order or case markings. In short, we posit that a major reason for the observable asymmetries among the world's languages is that certain patterns make a language more easily learnable by a general sequential-learning device, ensuring the proliferation of such a language in the human population.

Simulation 1: Exploring the Learnability of Case and Word Order

In the view that the frequency of certain WOs is correlated with their learnability, we hypothesized that typologically rare languages will be more difficult to learn by a sequential-learning device than the more common languages. To test this prediction, we trained simple recurrent networks (SRNs: Elman, 1990) on a total of 14 artificial grammars, reflecting the 6 possible strict word orders (SWO) and a flexible word order (FWO), with or without the presence of case markings.

Method

Networks Ten SRNs were used in each condition. The networks were initialized with random weights in the interval [-0.1, 0].[1] Each input to the networks consisted of a distributed representation of a word, spliced with a case marker. Words were represented by 20-unit randomly generated bit-vectors. Although some vectors were bound to be close in the representation space, random assignment to words assured that any such interaction would not bias the results. Having words represented by random vectors may seem odd considering the complex phonology that underlies human languages. However, for present purposes such a representation seems to work just as well as phonological (e.g., van Everbroeck 2001), while dramatically decreasing training time. Case markings (nominative, accusative, dative, and genitive) were represented by a four-bit vector appended to the word vector. This made for a total of 24 input units. There were seven output units, corresponding to the grammatical roles the network was supposed to predict: subject, direct object, indirect object, genitive noun, verb, or end-of-sentence. In all simulations, the learning rate was set to 0.1, and momentum to 0.01. Each SRN had 30 hidden units and 30 context units.

Materials The lexicon contained 300 nouns and 100 verbs. This noun-to-verb ratio is generally consistent with human languages (e.g., British National Corpus). The verbs were evenly divided into intransitive, transitive, and ditransitive categories. As illustrated in Table 1, each grammar included three

[1] It was found that the slightly inhibitory starting weights provided for better performance across the board. A similar conclusion was reached by van Everbroeck (2001).

Table 1: A Sample SOV Grammar Used to Generate Training Corpora

S → Intransitive [.35] | Transitive [.35] | Ditransitive [.3]
Intransitive → NP-nom V-intrans
Transitive → NP-nom NP-acc V-trans
Ditransitive → NP-nom NP-acc NP-dat V-ditrans
NP → N | N N-gen [.25]

types of sentences: intransitive, transitive, and ditransitive. A sentence consisted of noun phrases (NP) and one of three verb classes. Twenty-five percent of noun phrases contained a noun in the genitive form (e.g., *John's brother*. The simplest sentence generated by such a grammar was a simple intransitive: e.g., *John walks*. The most complex sentence contained 7 words: *Mary's friend gave Peter's key [to] John's brother*. A fully flexible grammar was identical to the strict WOs except the order within each element was randomly varied from sentence to sentence. In an effort to model the languages as naturalistically as possible, we modeled genitives based on Greenberg's (1963) universal 2: in typically prepositional languages (SVO and VSO) genitives generally follow the governing noun, while in postpositional languages (SOV), the reverse is true. We modeled the remaining three word orders with genitives following the noun. We also added a genitive case-marking to SWO-no case languages. Without this, it was impossible for the networks to discern governing nouns from genitives. This addition is motivated by the observation that even normally case-less languages have some form of genitive case markings (e.g., in English: *Mary's house*) (van Everbroeck, 2001).

Procedure We used a crossover design of 7 word orders (6 strict, and one flexible), by two case conditions (with or without case) resulting in 14 training corpora. For each condition, we generated 3,000 random sentences of the appropriate order. Such a corpus occupies a very small part of the possible sentences that can be generated by the corpus. For instance, 9 million different sentences are possible for a transitive SOV configuration (300 x 300 x 100).

The networks were trained for 100,000 sweeps (input/output pairs), corresponding to about 7 passes through the corpus. During each training sweep, the network was presented with a word, and depending on the condition, a case marking. A corpus of 200 novel sentences was created for testing. In the testing corpus, 50% of words were completely new—ones to which the network has never been exposed. Performance was measured by assessing the network's ability to map a given word to its correct grammatical role. During testing, the network's highest-activated output unit was compared to the expected output. If the units matched, the word was marked as being correctly mapped.

Table 2: Network performance and Language Distributions

Word Order	Words Correct – No Case Condition (%)	Attested Frequency (%)
SOV	90	51 (most w/cases)
VOS	85	8
OVS	80	0.75
OSV	74	0.25
Flexible	65	0 (all w/cases)

Note. Attested language frequencies taken from Van Everbroeck (1999).

It may seem that providing the networks with direct mapping from word to grammatical category is not ecologically valid. After all, it has long been recognized that kids are not given sufficient ostensive cues to syntactic relationships and word meanings. No one explains to the child after each encountered sentence that word *A* referred to the "do-er" and word *B* to the "do-ee". However, Tomasello and colleagues have shown that children are able to use pragmatic cues such as eye gaze to help figure out which object is being referred to (Tomasello & Akhtar, 1995). Twenty-four month olds show understanding of adult intentions in inferring meanings of novel verbs (Tomasello & Barton, 1994), and 18-month old children are able to learn new words in non-ostensive contexts (Tomasello, Strosberg, & Akhtar, 1996). Such use of pragmatic cues enables children to map words onto meanings and correctly infer who did what to whom. Considering that our networks live in a purely linguistic world, the method of direct mapping seems reasonable.

Results

All networks trained in the case condition were able to map 100% of the words to the correct categories. When case was not available, the network performance roughly corresponded to attested language frequencies (Table 2). Only two caseless WOs obtained nearly perfect performance: SVO and VSO (99%). There, however, appears to be a discrepancy. While SOV is the most common WO, it is outperformed by both SVO and VSO. According to Greenberg's Universal 41, however, the great majority of SOV languages have case, and most caseless languages turn out to be either SVO or VSO (e.g., English, Welsh). This finding supports our learnability hypothesis: verb-final languages presumably have a case system because reliance on WO results in suboptimal learnability.

The likely reason SOV-no case grammars did not achieve perfect accuracy is because they contained two unmarked nouns prior to the verb. Since the networks learn to map different types of verbs to different argument constructions, verb-final grammars are at a disadvantage—in these grammars the

grammatical role that provides the most information about what is to come, is received last (van Everbroeck, 2001). The poor performance of VOS is due to the intervening indirect subject in ditransitive sentences. We should also note that even though FWO-no case languages perform poorly, their performance is consistently above chance. This can be explained by the networks' learning to map familiar verbs to intransitive, transitive or ditransitive word schemas.

These simulation results confirm the idea that FWO-case languages are no more difficult to learn than common SWO-no case languages such as SVO and VSO, counter to predictions of the subset principle. The difficulty associated with learning a FWO language without case markings is underscored by typological evidence, suggesting that such languages use case markings to signal grammatical relationships (Payne 1992).

Simulation 2: The Impact of Case on Word Order Flexibility

In natural languages, case markings are not wholly deterministic. For instance, Slavic languages such as Russian and Serbo-Croatian, contain a number of nouns which, perhaps for historical reasons, do not take case markings. Additionally, because these markings often take the form of suffixes, they change the phonology of words. This results in potential phonological ambiguity. For instance, in Russian *stali* is either the genitive form of *steel* or a conjugated verb meaning *we stopped*. By examining the effects of varying cases on different word orders, we hoped to show that (1) even probabilistic case markings improve performance for FWO languages, and (2) case markings do not improve performance for languages that already rely on WO.

Method

Networks Ten SRNs were used for each condition. The initial conditions and training details were the same as in Simulation 1.

Materials We generated five artificial grammars varying on the salience of case markings—from only genitive markings, to full case markings. A grammar with case marked 50% of the time corresponded to a language in which 50% of case markings are possibly phonologically ambiguous, or a language in which certain of nouns do not take on case markings. Five more grammars varying on strictness of word order—from a completely flexible order, to a completely strict one (SVO). The word orders approximated distributions found in natural languages (Italian, and Turkish: Slobin & Bever 1982); Polish: Jacennik & Dryer 1992). The two conditions were crossed, for a 5x5 matrix. As in simulation 1, 3000 sentences were generated for each condition.

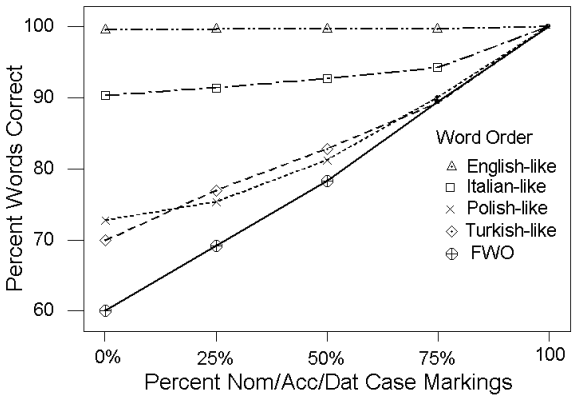


Figure 1: Network performance in simulation 2 for increasing degrees of case markings as a function of word order.

Procedure Each group of networks had case cues added to the sentences based on case condition. The testing proceeded as in Simulation 1.

Results

As expected, SWO languages such as English and Italian were little-benefited by case (Figure 1). In contrast, the probabilistic addition of case markings to FWO languages consistently improved performance. The slightly lower performance of Italian is due to it having a more flexible word order than English (see Table 3). To compensate for possible ambiguities, Italian relies heavily on prosodic and contextual information (Slobin & Bever, 1982) which was not available to our networks. In summary, the precise nature of the cue: case, or WO, does not seem to matter. Neither needs to be primary.

Simulation 3: Interactions between Case and Word Order Flexibility in Development

In this simulation, we demonstrate that networks trained on corpora similar to those used in simulation 2 are able to mimic syntactic performance of children learning English, Italian, Turkish, and Serbo-Croatian. Slobin and Bever (1982) tested 48 children divided into 8 age groups (24-52 months). Each child was tested on their ability to demonstrate familiar actions (e.g., *scratch*, *bump*, *pick up*) using familiar toy animals after hearing a transitive language in their native language. The authors hypothesized that Turkish, English, and Italian-speaking children would have the easiest time due to the consistent, unambiguous case markings available in the case of Turkish, and the consistent word-order information available in English and Italian. Children

Table 3: Word Order Distributions for Simulation 3

Language	Words Order	Cases
English	100% SVO	Genitive only
Italian	82% SVO, 2% SOV, 11% VSO, 5% OVS	Genitive only
Serbo-Croatian	55% SVO, 16% SOV, 16% VSO, 3% VOS, 2% OVS, 8% OSV	Full for non-SVO For SVO: 55% nom, 55% acc, 100% dat, 70% gen
Turkish	48% SOV, 25% SVO, 13% OVS, 8% OSV, 6% VSO	Full

acquiring Serbo-Croatian would have a more difficult time due to its more ambiguous case markings, requiring them to pay attention to word-order as well as cases.

Method

Networks The networks and training details were identical to simulation 2. We used 12 SRNs in each condition, mirroring the number of subjects used by Slobin and Bever (1982).

Materials We created 4 types of grammars motivated by the four languages used in the study. English was modeled as being 100% SVO, and having only genitive case markings. The word orders for the remaining languages were modeled based on the data provided by Slobin and Bever's (1982) corpus of adult speech, reflecting the linguistic input available the children.

Although Turkish does not have an explicit nominative case, it was found that such a marker was necessary in this simulation. In the absence of semantic information and case markings, the networks must rely on the syntactic position of a word to correctly identify its category. However, in a relatively FWO language such as Turkish, this information is ambiguous. Without a nominative case, both verbs and subjects are unmarked, and the network naturally has trouble telling them apart. In contrast to these networks, children rely on semantic information, in addition to syntax, to tell apart verbs and nouns. In other words, a Turkish child knowing the meanings of "dog" and "sniff", will not confuse the two even when "dog" is an unmarked agent in the sentence.

Serbo-Croatian has all four of the cases we were modeling, however, only masculine and feminine nouns take on accusative and nominative markings. Sentences containing one or more neuter nouns are typically ordered as SVO. We did not have data on the proportion of neuter nouns in Serbo-Croatian or the percentage of SVO sentences containing such nouns. It was estimated that about 55% of SVO sentences would contain one such noun, therefore case

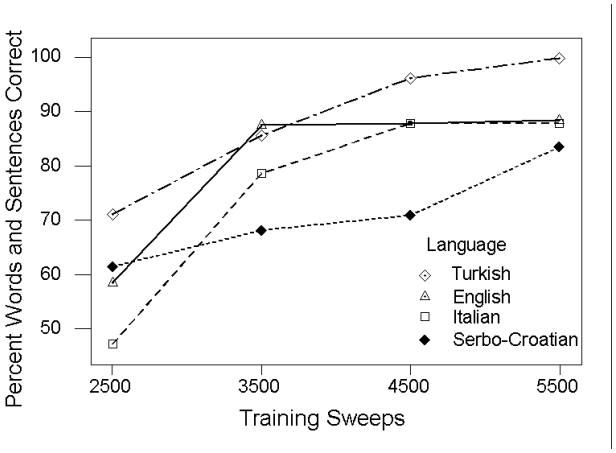


Figure 2: The pattern of performance across training for Turkish, English, Italian, and Serbo-Croatian in simulation 3

Table 4: Percentage correct performance for grammatical sentences in a given language in the Slobin and Bever (1982) study.

	Language			
Age	English	Italian	Serbo-Croatian	Turkish
24-28	58	66	61	79
32-36	75	78	58	80
40-44	88	85	69	82
48-52	92	90	79	87

markings were deleted from 55% of nouns in SVO sentences. Serbo-Croatian neuter nouns do have dative case-markings, hence the datives are marked 100% of the time. However, plural neuter nouns are not declined in genitive constructions. If plural genitive nouns are used an estimated 30% of the time, then 70% of SVO sentences will have genitive case markers.

Procedure Training proceeded as in simulation 1. The extent of training was varied for networks corresponding to different age groups. Testing was done following the procedure employed by Slobin and Bever (1982). We used transitive sentences using only words which the networks have seen during training. Performance was quantified by measuring the percentage of subjects and objects the network identified correctly, and averaging the data with the overall percentage of words correctly identified.

Results

The networks' performance (Figure 2) closely matched Slobin and Bever's (1982) data (Table 4). As predicted, networks trained and tested on Turkish had the easiest time mapping words onto grammatical roles. Networks trained on Serbo-Croatian,

had the most difficult time, highlighting the higher processing-cost associated with having to pay attention to WO and case markings. This pattern of results runs counter to the subset principle since the latter predicts case-languages to be more difficult to acquire. Performance on Italian was slightly worse than on English, reflecting the more flexible WO of Italian. It is predicted that with the addition of prosodic and semantic cues, the performance of Italian would more closely parallel that of a fully SWO language such as English.

Conclusion

Our findings confirm that learnability of languages may be a major factor in the frequency of certain language types. In the view of language as an organism (Christiansen, 1994), languages which are easily learnable by the human sequential-learning device proliferate, while languages not easily learnable die out or never come into existence. Our simulations suggest that all that is needed to learn syntactic relations is a reliable cue: case, or word order—neither needs to be primary. As such, no parameter-setting or subset principle is needed to account for the data. These results also provide added support for a connectionist approach to studying acquisition and evolution of language.

The simulations described here have several notable limitations. The sentences used for training were admittedly simple. Although simple intransitive, transitive, and ditransitive sentences are very frequent in speech, natural languages are rife with more complex structures such as relative clauses and embedding. Offsetting the simple grammars, however, were the limited cues available to the networks, which relied solely on distributional information of grammatical categories. In contrast, children routinely use semantics and prosodic cues, and even more subtle cues such as differential word length of nouns and verbs (Cassidy & Kelly, 1991—see Christiansen & Dale, 2001, for a review). It is therefore quite remarkable that relying only on word order or case, the performance of the networks was near-perfect for common language types.

References

British National Corpus. Located at: www.hcu.ox.ac.uk/BNC

Cassidy, K.W. & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*: 348-369.

Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum.

Christiansen, M.H. & Devlin, J.T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. (pp. 113-118). Mahwah, NJ: Erlbaum

Christiansen, M. H. (1994). Infinite languages, finite minds: Connectionism, learning and linguistic structure. Unpublished doctoral dissertation, Centre for Cognitive Science, University of Edinburgh, U. K.

Ellefson, M.R. & Christiansen, M.H. (2000). Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. (pp. 645-650). Mahwah, NJ: Erlbaum

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Greenberg, J. H. (1963) Some universals of grammar with particular reference to the order of meaningful elements. In: *Universals of language*. Greenberg, J.H. (ed.). Cambridge, MA.: MIT Press.

Jacennik, B. & Dryer, M.S. (1992). Verb-subject order in Polish. In *Pragmatics of word order flexibility*, ed. Payne, D. L. Amsterdam: John Benjamins.

Kirby, S. (1998). Language evolution without natural selection: From vocabulary to syntax in a population of learners. *Edinburgh Occasional Paper in Linguistics*, EOPL-98-1.

Koster, J. (1999) The word orders of English and Dutch: Collective vs. individual checking. In: *Groninger Arbeiten zur germanistischen Linguistik*, Abraham, W. (ed.), University of Groningen, Groningen. 1-42.

Neeleman, A. (1994) *Complex predicates*. PhD Dissertation Utrecht University, Utrecht.

Payne, D. L.(1992). Verb-Subject Order in Polish. In *Pragmatics of Word Order Flexibility*, Amsterdam: John Benjamins.

Pinker, S. (1995). Language acquisition In *An invitation to cognitive science. Vol 1: Language* Gleitman, L.R. & Liberman, M. (Eds.) Cambridge, MA.: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Slobin , D. I. (1966). The acquisition of Russian as a native language. in F. Smith and G. A. Miller (eds.), *The genesis of language: A psycholinguistic approach*. Cambridge, MA: MIT Press

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition 12*: 229-265

Tomasello, M. & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*, 201-224.

Tomasello, M. & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology*, *30*, 639-650.

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, *23*, 157-176.

Van Everbroeck, E. (1999). Language type frequency and learnability. A connectionist approach. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 755-760). Mahwah, NJ: Erlbaum

Van Everbroeck, E. (2001). *Language type frequency and learnability from a connectionist perspective*. Submitted Manuscript.

On Understanding Discourse in Human-Computer Interaction

Paul P. Maglio (pmaglio@almaden.ibm.com)
Teenie Matlock (tmatlock@psych.stanford.edu)
Sydney J. Gould (sydneygould@hotmail.com)
Dave Koons (dkoons@almaden.ibm.com)
Christopher S. Campbell (ccampbel@almaden.ibm.com)
IBM Almaden Research Center
650 Harry Rd, B2-NWE
San Jose, CA 95120 USA

Abstract

We report on an experiment that investigated how people naturally communicate with computational devices using speech and gaze. Our approach follows from the idea that human-human conversation involves the establishment of common ground, the use of gaze direction to indicate attention and turn-taking, and awareness of other's knowledge and abilities. Our goal is to determine whether it is easier to communicate with several devices, each with its own specialized functions and abilities, or with a single system that can control several devices. If conversations with devices resemble conversations with people, we would expect interaction with several devices to require extra effort—both in building common ground and in specifying turn-taking. To test this, we observed participants in an office mock-up where information was accessed on displays through speech input only. Between groups, we manipulated what participants were told: in one case, that they were speaking to a single controlling system, and in the other, that they were speaking to a set of individually controlled devices. Based on language use and gaze patterns, our results suggest that the office environment was more efficient and easier to use when participants believed they were talking to a single system than when they believed they were talking to a several devices.

Introduction

One approach to human computer interaction is to improve the usability, user experience, and intuitiveness of technology by creating natural user interfaces. Here, *natural* refers to interactions that are like those people have with one another. Such is the goal of multimodal or attentive systems (Maglio, Matlock, Campbell, Zhai & Smith, 2000; Oviatt & Cohen, 2000), and speech and conversational interfaces (Maybury, 1997). Understanding cues in conversation, language use, perceptual abilities, and expectations is vital to building systems that can be used with little training.

Advances in technology are resulting in smaller, cheaper, and more pervasive computational systems than ever before. But are we ready for this surge of electronics and information? No longer confined to desktop or laptop machines, computational systems will soon extend across numerous "information appliances" that are specialized for individual jobs and embedded in the everyday environment (Norman, 1998). If point-and-click graphical user interfaces (GUI) have enabled wide use of PCs, what will be the paradigm for interaction with pervasive computing systems? As natural human-computer interfaces and pervasive systems converge, what form will technology take?

To address these questions, we explored the design of a pervasive system with speech input in an office setting. We were concerned specifically with conversational cues that people rely on when interacting with the system. Some evidence suggests that people can attribute human-like or social qualities to computers with which they interact; for instance, networked computers described as physically close to the user are judged as more helpful than those described as physically distant (Reeves & Nass, 1996). Although people do not treat computers as true conversational partners (Yankelovich, Levow & Marx, 1995), these sorts of results suggest that people apply natural ways of interacting to situations in which the conversational partner is a computer or other computational device.

Our main concern is whether it is easier for people to talk to a single system or to a collection of devices. In a previous study of a speech-controlled office, we found behaviors and attitudes depended on whether users received simple command recognition feedback (a blinking light) from the various devices that performed tasks or from a single, central location (Maglio, Matlock, Campbell, Zhai & Smith, 2000; Matlock, Campbell, Maglio, Zhai & Smith, 2001). In that study, users were faced with simple office tasks (such as looking up information, dictating a letter, and printing a letter) to be completed using speech input only. To do this, users were given a set of physical displays dedicated to various functions (such as address book, calendar, and so on). Between groups of participants, we manipulated whether feedback was associated with individual displays or with the room as whole. This feedback manipulation was meant to suggest either central control or distributed control. Behaviorally, we found that regardless of condition, participants rarely addressed individual devices verbally, but they looked at the devices that they expected to display the results

before they spoke (Maglio et al., 2000). In a questionnaire aimed at uncovering attitudes toward the office, we found that participants in the central condition were more likely to rate their interactions with the office as being similar to interactions with people than were those in the distributed condition (Matlock et al., 2001). The results show that although people judge the central controller to be more like a person, they interact with devices individually in both cases, looking at devices when they speak. One design implication is that the feedback provided by blinking lights enables natural user-computer interactions. But the question of whether it is easier to speak to a single system or to multiple devices remains.

Let us first consider how people use language to communicate. There are many theories. A popular view is that discourse is a shared activity whereby two or more individuals cooperate to build and achieve understanding (Clark, 1996). This joint activity view implies that the meaning of an utterance is determined not only by what the speaker wishes to say to the listener, but also by context. This includes speaker's beliefs about the situation, (e.g., what speaker assumes the listener knows about the context), common ground (e.g., shared history), and listener's ability to accurately interpret the speaker's message (e.g., listener is paying attention). For example, imagine that it's early afternoon and you have just come back from a favorite lunch spot. A friend looks at you and asks, "Was it crowded?", where *it* refers to the restaurant. It is no problem to use the indexical *it* because the friend can assume that you know which restaurant is being asked about. The question can even be reduced further by simply asking, "Crowded?", and you are still likely to understand what is meant. This type of coordinated interaction is so common and natural that people do not think twice about it.

Given context's role in understanding, the joint activity view implies that the process of conversation also involves verbal (e.g., prosodic) and non-verbal (e.g., gaze) cues to convey meaning. Feedback is critical for supporting user interactions with computational systems (Perez-Quinones & Sibert, 1996); for instance, appropriate acknowledgments (e.g., "uh-huh") based on prosodic cues in users' speech can improve user evaluation of the system (Tsukahara & Ward, 2001). Likewise, gaze provides important cues to attention and turn-taking in group interactions (Kendon, 1967; Argyle & Cook, 1976).

Hypotheses

Following the joint activity view of conversation, our main hypothesis is that given complex tasks and their dependencies on one another, participants who interact with a single system will be more likely to establish context and then assume the system shares it than will those who interact with several devices. In human-human communication, the number of words used by participants in a conversation decreases over time, suggesting that the more common ground, the easier the communication (Clark & Wilkes-Gibbs, 1986). Thus, if people treat the system in the central condition in some ways like a single other person and they treat each of the devices in the distributed condition in some ways like several other people, participants may feel that they need to establish common ground with the single system only once but that they need to establish common ground with each of the devices individually, and we can measure this by counting words used—specifically, words that refer to things mentioned previously. Moreover, if establishing common ground is easier in the central condition than it is in the distributed condition, many predictions follow; for instance, participants should make fewer mistakes in the central, and participants should be more engaged with the central system, as shown by gaze.

An alternative to the joint activity view is that discourse is simply a process of transferring information without reference to context or to those involved. On this view, meaning is derived from what is spoken without concern for who the speaker is or what the situation is. If this is the case, maintaining separate functions in separate devices might make it easier for users to keep the various functions straight, as each device naturally conveys its own range of available options (e.g., email device for email). Thus, on this view, common ground is not constructed over time but is established once by what the device can do. If this is the case, talking to a single system ought to be more difficult than talking to multiple devices, as the single system does not make the options apparent.

Experiment

The goal of the experiment was to investigate whether and how language use and gaze would differ between participants interacting with a central system and those interacting with a distributed system. Our study was done in a mock office in which participants completed office tasks under the illusion that they were controlling what was displayed on four specialized screens. This sort of mock-up or *Wizard-of-Oz* method is often used to investigate user expectations and performance with speech-based systems (Dählback, Jönsson & Ahrenberg, 1993; Gould, Conti & Hovanyecz, 1983). The Wizard-of-Oz method relies on human controllers behind the scenes to create the appearance of an intelligent system, mocking up displays and interaction results to collect performance data.

In one condition, the experimenter instructed participants to speak to a system that controlled all the devices, and in the other, to speak to the individual devices. Unknown to participants, two other

experimenters in a separate room watched and listened, controlling what was displayed from a palette of many possible screens. The rules of the game were for the experimenters to simply behave intelligently: if what the participant was trying to do was clear from speech and other context, the system was to respond appropriately. The experimenters controlling the system were blind to which participants were given which instructions.

Method

Between two groups of participants, we manipulated *only* the instructions. In the central condition, participants were repeatedly told that they were to talk to a single computer system that displayed information on four displays. In the distributed condition, participants were repeatedly told that they were to talk to four separate information devices. Tasks were identical in both cases: sending and receiving email, updating address information, scheduling appointments, arranging a flight, and registering for a conference. Information was displayed the same way in both cases.

Participants Eighteen participants (13 females and 5 males) were recruited from summer student interns and office staff at our research lab, and paid for their time.

Materials and Apparatus Our office mock-up contained four 15-inch liquid crystal displays (LCDs) arranged on an L-shaped desk. Embedded in the bezel of two of the LCDs were pinhole video cameras, which enabled eye gaze and body position to be easily recorded. A third camera mounted on the wall above the room recorded an overview of the scene. Each display was dedicated to a different function or task: email, calendar, travel planning, and address book.

There were two sets of instructions, one for each condition. Instructions for the central condition told participants to talk to the "BlueSpeak system", a single computer system that controlled four displays. Instructions for the distributed condition told participants to talk to a set of "BlueSpeak devices", four separate devices that ran autonomously. In both cases, the script was written as a memo from a fictitious manager named Bob Wilson. The memo told the participant that he or she was to be his temporary assistant (or temp) for the day. It asked the temp to register Bob for a conference, add a new address to his address book, get a flight from San Jose to New York, reschedule a meeting, and request a vegetarian meal on the flight. The script purposely did not specify how to issue commands. It included statements such as "You will need to arrange my travel to New York and from San Jose", "I need to register for the XYZ Conference," and "Make sure my calendar is updated". Such language made for many possible ways of making requests. In addition, a few tasks were given in email messages sent to the temp during the course of the session. Table 1 shows the set of tasks and subtasks each participant was expected to carry out.

Table 1. Tasks completed by participants.

Update Address Book
- open new address form
- dictate name, address, city, state, zip, phone, email

Register for XYZ Conference
- find XYZ information screen
- obtain Bob's personal information
- compose new email to XYZ
  - dictate name, email, phone, credit card
- add XYZ to calendar

Find and Reserve Flights
- find airline reservation screen
- interact with reservation "system"
  - dictate cities, dates, non-stop, under $400
  - reserve/book itinerary
    - obtain Bob's personal information
    - dictate name, email, credit card
  - add flights to calendar

Reschedule Meetings
- obtain Kathy's personal information
- compose new email message to Kathy
- read Kathy's response
- modify Kathy's meeting

Notify Bob of Status
- compose new email to Bob
- read Bob's response
- adjust calendar , cancel meeting
- modify airline reservation
  - find reservation for Bob
  - specify vegetarian meal

Procedure Participants read the instruction sheet and were told their input was valuable because it would help ensure that the "BlueSpeak system" or the "BlueSpeak devices" would be tested on a wide range of voices. Participants were then taken into the "Office of the Future", and asked to test out their voices by reading a short passage to the system in the central condition or to the devices in the distributed condition. Participants were then told to carefully read the memo left by Bob Wilson. In all cases, participants were instructed to speak naturally and to do the best they could. They were told that there was no right or wrong way to speak to the system or the devices, and that if they were not understood, to try speaking differently. After issuing a command, the system did not give any feedback other than displaying the result of the request.

Results

All participants successfully completed the session. Few problems arose and on average it took participants 13 min 43 sec to complete all tasks. Data analysis targeted language-use and eye-gaze during the session. Only reliable differences are reported, except as noted.

Language Qualitatively, participants spoke to the system in a variety of ways. For instance, requests to send email included, "Let's send an email to Kathy Webster," "I need to send an email now – I would like to send it to k webster at ibm dot com," "Email k webster at ibm dot com," and "Write an email to Kathy Webster". Requests to get Bob a vegetarian meal on included, "Request vegetarian meal," "Vegetarian meal," "Let's make this a vegetarian meal", and "Special request, vegetarian meal for this flight".

More precisely, transcribed utterances in both conditions were examined for certain characteristics of language use. First, requests were placed into four categories: imperative, elliptical, first person, and question (cf. Maglio et al., 2000). Imperative requests are commands, such as, "update the address book", "view addresses", "register for conference". Elliptical requests contain no a verb, such as, "XYZ conference", "new entry", and "Kathy Webster". First person requests include either a singular or plural first person subject, such as "let's read this email", and "I want a vegetarian meal". Question requests include queries such as, "can I check my email?" and "are there any other flights available?". Figure 1 shows the break down of requests for both conditions. There were more imperatives (central, 70.4%; distributed, 80.3%; χ^2 = 9.75, $p < 0.01$) and more ellipticals (central: 11.8%; distributed: 16.9%; χ^2 = 4.16, $p < 0.05$) in the distributed condition, and there were more first persons (central, 13.3%; distributed, 2.8%; $\chi^2 = 26.9$, $p < 0.01$) and questions (central, 4.5%; distributed, 0.0%; χ^2 = 16.4, $p < 0.01$) in the central condition.

Second, we examined how participants verbally addressed individual devices. Specifically, we counted the number of times a device was specifically addressed by name, such as, "Address book, what is Kathy Webster's address?". The proportion of requests containing an addressee was greater in the distributed condition than in the central condition (central, 1.7%; distributed, 14%; $\chi^2 = 40.87$, $p < 0.01$).

Third, we examined the way participants recovered from errors. We were interested in how requests were reformulated after an initial attempt had failed. To take one example, the most problematic part of the script was ordering a vegetarian meal for the flight (last task in Table 1). Overall, in the central condition, the meal request was restated 9 times. In the distributed condition, it was restated 13 times. Three participants in the distributed condition were unable to complete this task at all and eventually gave up.

Fourth, we looked at the way participants relied on previously established context. For example, when interacting with the system, a participant might say "register Bob Wilson for XYZ conference" and then a short time later say, "add event to calendar", referring implicitly to the conference. In this case, the participant assumes the system is following the discourse, and that once established, the context (the conference event) need not be repeated. We calculated the proportion of statements that assumed context available in a previous statement across all participants in each condition (see Figure 2). Overall, participants in the central condition assumed that the system would understand the context more often than participants in the distributed condition (central, 7.0%; distributed, 1.4%; $\chi^2 = 14.50$, $p < 0.01$).

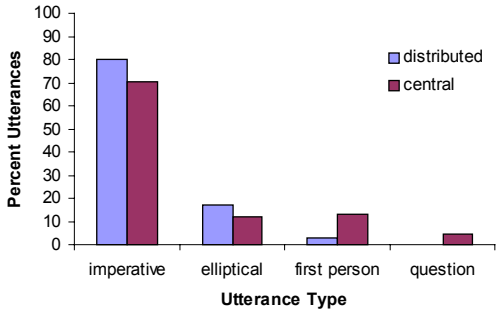


Figure 1. Percentage of the time participants used different types of requests.

Finally, we charted how language use changed during the course of a session. Each participant's discourse was cut in half, based on the task breakdown in Table 1. The number of times each participant relied on established context (as defined previously) was tallied separately for the first half and for the second half (see Figure 2). No difference between central and distributed conditions was found for the first half (central, 5.5%; distributed, 2.2%; $\chi^2 = 2.45$, NS), but a reliable difference was found for the second half (central, 8.5%; distributed, 0.6%; $\chi^2 = 13.59$, $p < 0.01$).

Behavior and Gaze Behaviors—including actions participants took and where they looked—were analyzed in terms of the task breakdown in Table 1. Specifically, all overt physical actions taken by participants were transcribed from the videotapes and time-stamped. From these data, we extracted number of tasks, time taken per task, number of gazes or looks to task-relevant and to task-irrelevant locations, and number and kind of the errors made. For all results, scores falling outside two standard deviations from the mean were removed and replaced by mean scores; these outliers constituted 8% of the scores.

Mean completion time was 13 min 17 sec for the central condition and 14 min 10 sec for the distributed condition. To calculate the time taken for each individual action for each participant, the time taken for each task (e.g., "Update Address Book," "Register for Conference," etc.) was divided by the number of actions (e.g., "open address book", "find Bob's personal information", etc.) actually taken to complete

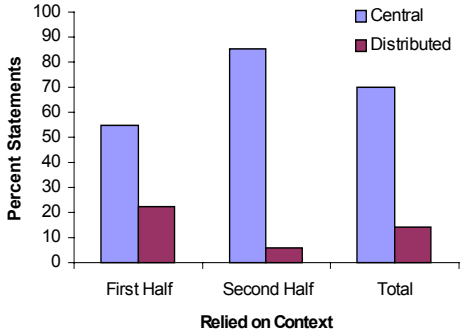


Figure 2. Percentage of time participants relied on established context.

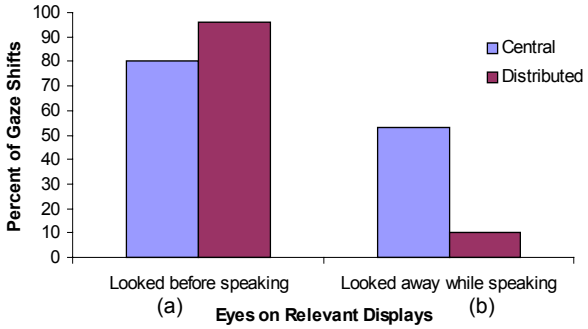


Figure 3. Percentage of time participants (a) shifted gaze to display before speaking, and (b) away from display while speaking.

the task. A difference was found between central and distributed conditions on the mean time to take an action in the address book task (central, 21.1; distributed, 27.1; $t(16) = 3.18$, $p < 0.01$), and a marginal difference was found between the mean times to take an action in the flight reservation task (central, 54.0; distributed, 73.4; $t(16) = 1.71$, $p < 0.11$).

Unnecessary actions and omitted actions constituted errors. Number of errors was calculated for each participant. Percentage of errors was determined by dividing the total number of errors by the total number of actions taken for each participant. Overall, there were 20% errors in the central condition and 16% in the distributed condition. This difference was not reliable.

Finally, we examined where people looked and when they altered their gaze. In particular, we counted the total number of times a participant looked at a specific display when making a request for which that display would be expected to show a result (see also Maglio et al., 2000). For instance, we counted times when a participant looked at the address book and then said "Michael Smith's address," but not times when the participant would say "Michael Smith's address" before looking at the address book. The percentage of the time each participant looked at the appropriate display when taking action was calculated by dividing the number of appropriate looks by the number of actions. As shown in Figure 3a, a difference was found between the two conditions (central, 80%; distributed, 96%; $t(16) = 2.79$, $p < 0.05$). In addition, we counted the number of times a participant looked away from a display they were using to complete an action (again, normalizing with respect to total number of actions). As shown in Figure 3b, a difference was found between the two conditions (central, 53%; distributed 10%; $t(16) = 2.39$, $p < 0.05$).

Discussion

In summary, participants interacting with the single system had an easier time than those interacting with multiple devices. Specifically, the data show

1. dominant use of the imperative in both cases, and more use of first-person and question forms in the central case
2. less verbal addressing in the central case
3. less reformulation—and more successful reformulation—of requests in the central case
4. more reliance on context in the central case, which increased over time
5. slightly faster overall completion time for the central case, and significantly faster for certain tasks (updating address book, reserving a flight)
6. less of the time, gaze shifted to the appropriate display before speaking in the central case
7. more of the time gaze shifted away from the appropriate display in the central case

Returning to our hypotheses, we can conclude that participants in the central condition relied more heavily on context they had established previously, shown by the number of times they implicitly referred to objects and information. This is what we expected given the joint activity view of discourse (Clark, 1996). It follows that participants in the central condition behaved more like they were speaking to a single entity than those in the distributed condition. The way participants addressed displays and shifted gaze also supports this conclusion. Participants in the central condition addressed individual devices less frequently, suggesting that they were less likely to be speaking directly to devices. Moreover, participants in the central condition shifted gaze to individual devices when starting a task less frequently, also suggesting that they were less likely to be speaking directly to individual devices. Finally, participants in the central condition more frequently shifted gaze from device to device while engaged in a task, suggesting that they were unconcerned with keeping eye contact with a specific device. Taken together, these results suggest that participants in the central condition behaved more like they were engaged with a single entity than those in the distributed condition.

Conclusion

The present study was intended to investigate how people speak to computational systems. Controlling whether users believed they were speaking to a single centralized system or to several separate devices, we found a centralized system was more efficient and easier to use than separate devices in several ways. Not surprisingly, the main difference was that users of the central system treated the system as a single entity whereas users of the separate devices treated the devices as independent entities. By relying on a single controller, users in the centralized condition were more likely to reuse conversational context than users in the distributed condition. Moreover, because they interacted with a single entity, users did not need to divide attention across several conversational partners.

What are some implications for Clark's joint activity view of conversation? It may initially seem misguided to apply this theory to human-computer interaction, for it was intended to deal with human-human interaction only. And afterall, computers and other devices are not *true* conversational partners because they are controlled by their users and cannot really engage in conversation. Nonetheless, Clark's theory was in fact predictive of behavior in this study, demonstrating that in this human-computer interaction context, many of the same assumptions about human-human interaction apply.

What are some implications for the design of future computing environments? First, for the sorts of tasks considered here, it is clear that a single controller is to be preferred over multiple devices. Thus, when designing a system that requires a user to coordinate information and activities among a set of distinct displays or information sources, it would be appropriate to provide the user a single point of contact with the overall system, as this would allow the user to establish an ongoing relationship with a single entity. Second, because maintaining context seems critical for efficiency (and possibly for ease of use as well), providing users with appropriate state information would likely encourage them to rely on established context. Third, because users tended to fix their gaze on individual devices in the multiple device condition, gaze cues (in addition to language cues) might be useful in helping the system determine level of engagement and to disambiguate referential statements, but cannot be relied on completely in the single controller case. Fourth, if assumptions about common ground can be manipulated by instructions, then the physical design of a system should be carefully considered. For example, putting several screens in the environment with the same physical size and characteristics might suggest multiple devices, whereas one large display and a few smaller displays might appear to be a single system with one point of contact and several output monitors. Fifth, giving devices obvious ways of collecting information from the room, such as a visible camera, allows users to understand what kind of common ground the system is likely to have. For instance, the camera may make users more likely to use gestures (e.g. pointing) to reference information. Finally, providing users with a single point of control need not have consequences for implementation; it might simply be enough to *tell users* to speak with a single device.

References

Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze*. London: Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge England: Cambridge University Press.

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how, in *Proceedings of the Workshop on Intelligent User Interfaces '93.*

Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM, 26(4),* 295-308.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32*, 1-25.

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces, in *Proceedings of the International Conference on Multimodal Interfaces 2000*.

Matlock, T. Campbell, C. S., Maglio, P. P., Zhai, S., & Smith, B. A. (2001) Designing feedback for an attentive office, in *Proceedings of Interact 2001*.

Maybury, M. T. (1997). Conversational multimedia interaction. In Y. Wilks, (Ed.) *Machine Conversations.* Kluwer Academic, Norwell, MA.

Norman, D. A. (1998). *The invisible computer.* Cambridge, MA: MIT Press.

Oviatt, S. & Cohen, P. (2000). Multimodal interfaces that process what comes naturally. *Communications of the ACM, 43(3)*, 45-53.

Perez-Quinones, M. A. & Sibert, J. L. (1996). A collaborative model feedback in human-computer interaction, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96*.

Reeves, B. & Nass, C. (1996). *The media equation.* Cambridge, England: Cambridge University Press.

Tsukahara, W., & Ward, N. (2001). Responding to subtle, fleeting changes in the user's internal state, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2001,* 77-84.

Yankelovich, N., Levow, G. A., & Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces, in *Proceedings of the Conference on Human Factors in Computing Systems, CHI '95*.

On the Potential of Epistemic Actions for Self-Cueing: Multiple Orientations Can Prime 2D Shape Recognition and Use

Paul P. Maglio (pmaglio@almaden.ibm.com)
IBM Almaden Research Center
San Jose, California

Michael J. Wenger (mwenger1@nd.edu)
Department of Psychology
University of Notre Dame

Abstract

Epistemic actions are physical actions people take more to simplify their internal problem-solving processes than to bring themselves closer to an external goal. Consider how when playing the video game Tetris, experts routinely rotate falling two-dimensional shapes more than is necessary to place the shapes. One reason for such apparently unnecessary actions is that they actually help the player make placement decisions. Such actions might facilitate placement decisions if additional previews of the shape afforded by rotating it provide information about the board, particularly when there is no direct perceptual match between the shape and the board at the time of decision. The study presented here tests the hypothesis that several distinct previews of a two-dimensional shape can improve a person's ability to recognize and use that shape when it is not correctly oriented at the time of decision. Results show that indeed task performance and recognition are faster with *two different* orientations than with only one. Thus, it is possible that Tetris players rotate two-dimensional Tetris shapes manually to see them in more than one orientation, as this can lead to faster decisions.

Introduction

People playing the video game Tetris often take actions that are not strictly necessary but that serve to simplify or speed up internal cognitive or perceptual operations (Kirsh & Maglio, 1994; Maglio & Kirsh, 1996). Playing Tetris involves maneuvering falling two-dimensional shapes into specific arrangements on the computer screen (see Figure 1). Even as players become faster with practice, they tend to over-rotate falling shapes, leading to backtracking as these over-rotations are corrected. To make sense of such backtracking, Kirsh and Maglio (1994) argued that sometimes physical rotation can serve the same purpose as mental rotation, effectively offloading mental computation onto the physical world (see also Clark, 1997; Kirsh, 1995; Maglio, Matlock, Raphaely, Chernicky & Kirsh, 1999). Such physical actions—taken to simplify internal cognitive computation rather than to move closer to the external goal state—are called *epistemic actions*.

Because shape identification can be facilitated when primed with orientations different from the target orientation (Cooper, Schacter, Ballesteros & Moore, 1992; Srinivas, 1995), and because numerosity judgments can be facilitated even when test stimuli are not presented at the same orientation as the originally learned patterns (Lassaline & Logan, 1993), memory for a target pattern might not require the retrieval cue be specifically oriented. Thus, the epistemic function of physical rotation in Tetris might be far more complex than is suggested by the simple idea that physical rotation can substitute for mental rotation, for instance, serving the function of cueing retrieval (Kirsh & Maglio, 1994). Because physically rotating a Tetris shape (which we call a *zoid*) provides the player two views of it (i.e., in each of two orientations), it is possible that seeing two different views makes retrieval of relevant information easier than does seeing just one. In fact, we found previously that when participants in a Tetris-like task are presented with two views of a zoid, the time taken to decide whether it fits a particular board is faster than when participants are presented with only a single view, but this does not depend on the orientation of the previews relative to one another (Maglio & Wenger, 2000).

There are at least three potential functions of rotation for self-cueing in Tetris. First, seeing the falling zoid in several different orientations may provide helpful information about the board, particularly when there is no direct perceptual match between zoid and board at decision time. That is, if the orientation of the zoid floating above the board does not match the orientation in which the zoid actually fits the board when the player must decide to place it, then having previously seen the zoid in the matching orientation might help the process of mentally matching zoid and board. Such an effect might be the result of having recently seen the zoid in its fitting orientation, basing the decision as to whether the zoid matches the board on memory rather than on mental rotation. Let us call this potential epis-

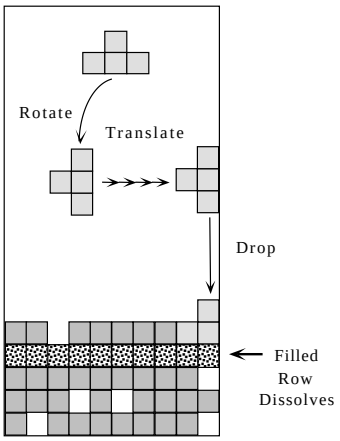


Figure 1: In Tetris, two-dimensional shapes fall one a time from the top of the screen, landing on the bottom or on top of shapes that have already landed. There are seven shapes, or *zoids*—, , , , , , . As a zoid falls, it can be rotated, and moved right or left. The object of the game is to fill rows of squares all the way across. Filled rows dissolve and all unfilled rows above move down.

temic function of rotation, the *board-match* function.

A second potential function of rotation for self-cueing might be to provide advance information about the zoid itself, particularly when the several previews coincide with the orientation of the zoid at the time of decision. That is, if the orientation of the zoid floating above the board matches the orientation in which the zoid fits the board when the player must decide to place it, then having seen it previously in that orientation might make recognition easier. In this case, such an effect might be the result of a complex memory retrieval process in which multiple views of a shape lead to faster or more reliable recognition of it (see Maglio & Wenger, 2000). Let us call this potential epistemic function of rotation, the *zoid-retrieval* function.

A third potential function of rotation in Tetris might relate to motor processes rather than to memory or perceptual processes. Because physically rotating objects can facilitate or inhibit mental rotation under certain conditions, it is possible that mental rotation and physical rotation share at least some internal processes (e.g., Wexler, Kosslyn & Berthoz, 1998). Thus, the specific motor act Tetris players take in rotating the falling zoid might serve the purpose of coordinating motor processes with other internal processes to facilitate zoid placement decisions. Let us call this potential epistemic function, the *motor-process* function.

These three epistemic functions of action—the board-match function, the zoid-retrieval function, and the motor-process function—are not mutually exclusive. All are possible reasons for the over-rotations observed in normal Tetris play. In this paper, we explore only the board-match function. Specifically, we test the hypothesis that seeing *several different* orientations of a falling zoid is better than seeing just one when the final orientation of the zoid does not match the region the zoid fits on the contour of the board. As noted, any such facilitation might result from matching the board to the memory of the previewed zoid rather than mentally rotating the zoid seen at test. Thus, our board-match hypothesis is a kind of memory-retrieval hypothesis.

Retrieval demands while playing Tetris can be thought of as *indirect* tests of memory in that they allow for effects of prior experience to be expressed without requiring explicit memory for the original experience (e.g., Richardson-Klavehn & Bjork, 1988). Tasks requiring explicit memory for the original event—such as old/new recognition or recall—are referred to as *direct* tests of memory. Because direct and indirect tests are differentially sensitive to orientation, object symmetry, and other physical aspects of visual objects (Srinivas, 1995; Srinivas & Schwoebel, 1998), the experiment presented here used both direct and indirect assessments of memory to determine how effective previews are under different retrieval demands. Because the effectiveness of memory cues generally depends on the time that elapses between presentation of cue and presentation of the item to be retrieved, we also investigated the effect of various delays between final preview and onset of test.

Method

To test whether two orientations of a falling zoid leads to faster performance in Tetris than one orientation does, we created a controlled experimental situation that shared many attributes with the game of Tetris but that allowed fine-grained control over the parameters of interest. In our experimental set up, a Tetris configuration (a Tetris board and zoid floating above it) is preceded either by none, one, or two previews of the zoid in either the same or different orientations. The participant's job is to quickly and accurately determine at the time of test whether the zoid floating above the board fits snugly on the board. Thus, the task creates situations similar to those faced by Tetris players during an actual game, and also requires responses similar to those required of players during an actual game. In all cases, in the final Tetris configuration, the zoid and the region it fits on the board contour (if it fits) are oriented

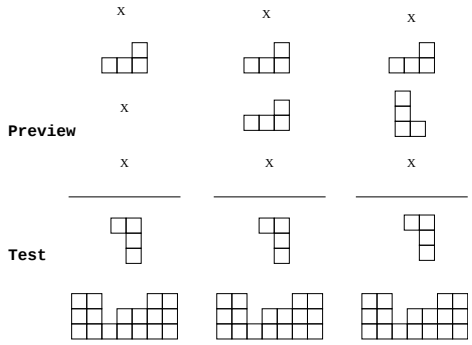


Figure 2: Three trial types used, from left to right: one preview, two previews in the same orientation, and two previews in different orientations. An "X" indicates display of an irrelevant zoid for the trial.

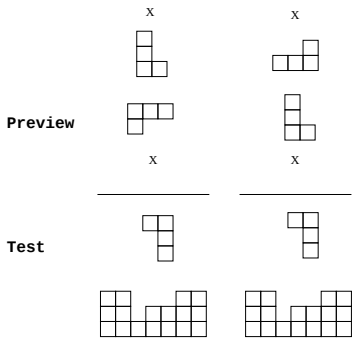


Figure 3: The second preview zoid is oriented properly relative to the board in the trial on the left but not in the one on the right.

differently, meaning there was no perceptual match between zoid and board at test (see Figure 2). In some cases, the last preview was oriented so as to fit snugly on the board contour without rotation, in which case memory for the previewed zoid might facilitate the fit/no-fit decision (see Figure 3).

Participants spent about three hours playing our experimental version of Tetris. Separate groups of participants were required either (a) to make judgments about whether a target zoid fit in an accompanying board (indirect test), or (b) to make this judgment *and* indicate whether they remembered seeing the test zoid in the set of zoids that were presented prior to the target (direct test). Between 0 and 2 previews of the target zoid were presented in a sequence of zoids prior to the target, and the orientation of these previews (when present) varied relative to the target. By placing the previews in a sequence of events prior to the test, we were able to manipulate the interval over which the preview would have to be retained in memory.

Participants

Twenty-nine participants were recruited from psychology courses and participated voluntarily in exchange for course credit: 15 in the indirect condition, and 14 in the direct condition. All participants reported normal or corrected-to-normal vision.

Design

The experimental design was fairly complicated so as to control as many factors as possible. As described, our main interest was in whether multiple previews of the zoid primed recognition and use better than a single preview when there was no perceptual match between zoid and board at test. In addition, we controlled whether the test zoid fit the test board, whether the preview zoids were in the same or different orientations, the time between preview and test, and whether memory was tested directly (asking whether the test zoid had been previewed) or indirectly (asking only for a fit/no-fit judgment).

More precisely, the experiment was conducted as a 4 (preview type: no previews, one preview, 2 previews same orientation, 2 previews different orientation) $\times$ 2 (orientation of the last preview relative to the board: same, different) $\times$ 3 (retention interval between last preview and target zoid, in frames: 0, 1, 2) $\times$ 2 (zoid type: ,) $\times$ 2 (status of target zoid relative to the board: fit, not fit) $\times$ 2 (type of memory judgment at test: direct, indirect) mixed factorial design. All factors except type of memory judgment were manipulated within participants.

Materials

All zoids and boards were constructed from 20×20 pixel squares. Squares were outlined by light gray lines, 1 pixel in width, and were filled in solid black. The background for all displays was also solid black. All zoid types were composed of four blocks. All boards were six blocks in height and width. Four "fit" boards were defined for each zoid type, corresponding to four ways in which the zoid could be snugly placed. Each such board was used with equal frequency. Materials were displayed on a 33 cm VGA monitor controlled by a PC-compatible computer. Onset and offset of each display was synchronized to the monitor's vertical scan. A standard keyboard was used to collect and time (to $\pm 1ms$) responses.

Procedure

Participants were tested on two consecutive days, at approximately the same time each day, with each session lasting approximately 90 min. All sessions were conducted in a darkened room, with participants seated an unconstrained distance from the

monitor, and began with a five min period for dark adaptation. Participants were told that, on each trial, they would see a sequence of zoids presented very rapidly. The zoids in the sequence would begin falling from a location near the top of the screen: each successive zoid would appear below the one before to create a sequence of falling zoids much as in the Tetris game. Each zoid was present for 250 ms, and each sequence consisted of between five and seven zoids, with the actual number determined randomly (and with equal likelihood) on each trial. At some random point in this sequence, participants would be presented with a combination of a test zoid and board, and would need to make one of two types of responses, depending on whether they were in the indirect or direct memory condition.

In the indirect condition, participants had to decide whether the zoid presented at test would fit snugly into the board. Participants responded in the affirmative using the index finger of the dominant hand, and in the negative using the index finger of the non-dominant hand, pressing either the "z" or "/" keys on the lower row of the keyboard. In the direct condition, participants had to indicate with a single key-press both judgment about whether the presented zoid fit snugly in the board and memory for any occurrence of the test piece (in any orientation) in the sequence that preceded the target. Participants responded with the index finger of the dominant hand if the target piece fit and they remembered seeing this piece in the preceding sequence, with the middle finger of the dominant hand if the target piece fit and they did not remember seeing it in the preceding sequence, and with the index finger of the non-dominant hand if the piece did not fit. Speed and accuracy were emphasized equally.

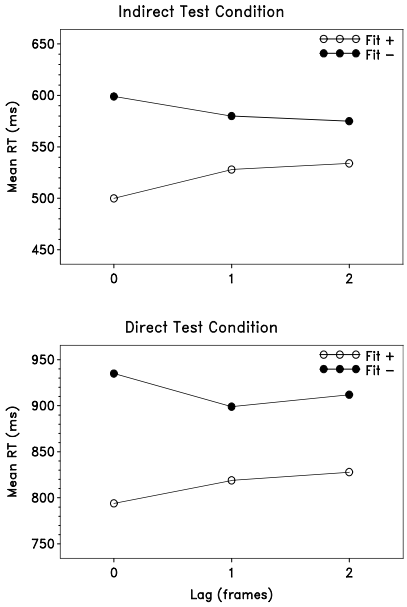


Figure 4: Preview-present trials: interaction of lag and fit status. Fit + indicates trials where the test zoid fit, and Fit -, where the test zoid did not fit. Lag is expressed in terms of the number of frames intervening between last preview and test zoid.

Results

Note that participants quickly became very good at this task; by the end of the first day, overall error rate was below 3%, indicating a high level of skill. Now, to determine whether primes had an effect, correct reaction times (RT) were analyzed using two 2 (zoid type: ,) × 2 (preview: present, absent) × 2 (status of the test zoid: does fit, does not fit) repeated measures ANOVAs, one for each test condition (direct, indirect). In both test conditions, presence of a preview speeded responses (indirect, 556 ms vs. 691 ms, $F_{(1,14)} = 11995.0$, MSE = 23.04; direct, 867 ms vs. 1008 ms, $F_{(1,13)} = 1095.75$, MSE = 254.10). Both conditions also showed faster responses when the test zoid fit relative to when it did not (indirect, 594 ms vs. 653 ms, $F_{(1,14)} = 43.67$, MSE = 1214.50; direct, 890 ms vs. 985 ms, $F_{(1,13)} = 120.27$, MSE = 1039.34). Finally, the effect of the presence of a preview was dependent on the status of the test zoid, with the preview effect being larger when the test zoid fit relative to when it did not (indirect, 140 ms vs. 131 ms, $F_{(1,14)} = 10.87$, MSE = 31.12; direct, 148 ms vs. 133 ms, $F_{(1,13)} = 5.08$, MSE = 181.28).

Having established that a preview made a difference, we next look to see whether having more than one preview made a difference, and whether the preview(s) had any interacting effects with other aspects of the design. The preview-present data were analyzed using two 2 (zoid type: ,) × 3 (number of previews: 1, 2) × 2 (orientation of the preview relative to the test piece: same, different) × 3 (lag, in frames, between the last preview and the test zoid: 0, 1, 2) × 2 (status of the test zoid: does fit, does not fit) repeated measures ANOVAs, one for each of the test conditions. A first result was an effect of number of previews: participants were faster with two previews than with one (indirect, 542 ms vs. 562 ms, $F_{(1,14)} = 101.02$, MSE = 353.96; direct, 857 ms vs. 872 ms, $F_{(1,13)} = 15.75$, MSE = 1172.30). The

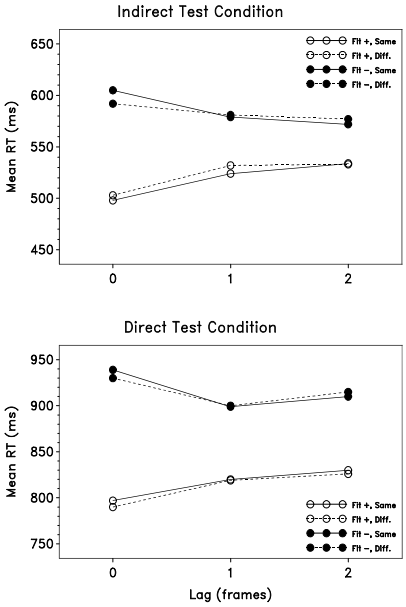


Figure 5: Preview-present trials: interaction of lag, fit status, and orientation of the test zoid.

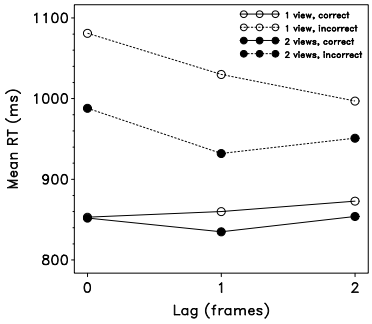


Figure 6: Preview-present trials, direct test condition: interaction of lag, number of previews, and accuracy of the memory judgment.

lag between the last of the previews and the status of the test zoid interacted in both test conditions (see Figure 4): decreases in lag produced faster RTs when the test zoid fit (indirect, $F_{(2,28)} = 73.19$, MSE = 379.39; direct, $F_{(2,26)} = 21.72$, MSE = 1488.90) but produced longer RTs when the test zoid did not fit. Finally, there was an interaction among lag, status of the test zoid, and orientation of the preview relative to the test zoid (see Figure 5), though this interaction was reliable only for the indirect condition ($F_{(2,21)} = 4.21$, MSE = 339.60).

We next examined trials on which there were two previews. Half of these involved previews in one orientation and half involved previews in two orientations. Analysis revealed that seeing two orientations led to faster responses than seeing one orientation in both indirect (616 ms vs. 686 ms, $t_{(28)} = 4.23$) and direct (932 ms vs. 980 ms, $t_{(26)} = 4.18$) conditions.

If epistemic actions serve the mnemonic purposes we have suggested, then there might be a certain awareness on the part of the player as to mnemonic state while playing, which in turn suggests that the ability to monitor memory state may modulate the benefits of epistemic action. To assess this possibility, we examined the preview-present trials of the direct test condition, treating accuracy of memory judgment as a random effect, using a 2 (memory accuracy: correct, incorrect) × 2 (zoid type: ,) × 3 (lag between last preview and test zoid: 0, 1, 2) × 2 (number of previews: 1, 2) repeated measures ANOVA. This analysis revealed the expected benefit of increasing the number of previews (899 ms for two vs. 946 ms for one, $F_{(1,24)} = 6.66$, MSE = 1194.80). It also revealed faster responses when participants accurately remembered the preview relative to when they did not consciously recall seeing a preview (855 ms vs. 997 ms, $F_{(1,24)} = 4.79$, MSE = 96227.58). Finally, there was an interaction among lag, number of previews, and the accuracy of the memory judgment (see Figure 6; $F_{(2,48)} = 4.02$, MSE = 1851.61).

Discussion

Our results show that when the test zoid and board are not oriented properly with respect to one another, one preview of the zoid in any orientation with respect to the board leads to faster responses than does no previews, and two previews lead to faster responses than does one preview. This was found for both indirect Tetris-like fit/no-fit judgment task as well as for direct memory recognition task. In addition, two previews with two different orientations produced faster responses than did two previews with the same orientation. Thus, under certain conditions, several orientations can prime two-dimensional shape recognition and use better than a single orientation can.

Note that response time was speeded up by a single preview in any of the three orientations relative to the test zoid and board. The benefit was not restricted to a preview that shared orientation with the test display. This finding is consistent with priming studies in which it was found that a prime need not be presented in the same orientation as the target to facilitate recognition or identification (e.g., Cooper, Schacter, Ballesteros & Moore, 1992; Srinivas, 1995). However, we have gone a step further than these by showing that *priming with several orientations is more effective than priming with a single orientation* under certain conditions.

The data also revealed that the benefit of previews, on trials in which the test zoid did fit, diminished as time elapsed between last preview and time of decision. This attenuation of the positive effects of previews suggests that the benefit of rotation for self-cueing may be restricted to a small window of time just prior to the final decision, which would be consistent with the reasonably rapid pace at which the game proceeds for skilled players. In contrast, on trials in which the test zoid did not fit, temporal proximity between last preview and judgment appeared to extract a cost, suggesting a strong specificity of the effect of previews to particular conditions of the game. Moreover, the data from the direct test condition revealed that accurate, conscious memory of a preview produced a benefit in responding, suggesting that players may have the ability to monitor their mnemonic state—as well as the state of the game—as play unfolds. One provocative idea is that epistemic actions may occur in response to players' assessment of a need for additional cueing.

Returning to the specific idea epistemic action in Tetris—the board-match function of rotation in particular—these results suggest that by rotating falling zoids, players may be able to effectively cue themselves, enabling quicker responses in a Tetris situation. Previous research has established various ways in which Tetris players take actions for their epistemic effects (Kirsh & Maglio, 1994; Maglio & Kirsh, 1996). The data reported here show that several previews of the falling zoid sometimes speeds up performance on a Tetris-like task, but the hypothesis that Tetris players over-rotate zoids in order to speed up performance is not directly tested. It remains to be seen whether actually taking the action of orienting the preview (i.e., physically rotating the falling shape) is a critical component of performance, independent of the presentation of the preview itself. It also remains to be seen whether the time-cost of making an extra move is more than compensated by the benefit in RT. We are exploring both questions.

References

Clark, A. (1997). *Being there: Putting body, brain, and world together again.* Cambridge, MA: MIT.

Cooper, L. A., Schacter, D. L., Ballesteros, S., & Moore, C. (1992). Priming and recognition of transformed three-dimensonal objects: Effects of size and reflection. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *18*, 43–57.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, *73*, 31–68.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Lassaline, M. E. & Logan, G. D. (1993). Memory-based automaticity in the discrimination of visual numerosity. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *19*.

Maglio, P. P. & Kirsh, D. (1996). Epistemic action increases with skill. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pages 391–396, Mahwah, NJ. LEA.

Maglio, P. P., Matlock, T., Raphaely, D., Chernicky, B., & Kirsh, D. (1999). Interactive skill in Scrabble. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*, pages 326–330, Mahwah, NJ. LEA.

Maglio, P. P. & Wenger, M. J. (2000). Two views are better than one: Epistemic actions may prime. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, Mahwah, NJ. LEA.

Richardson-Klavehn, A. & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology*, *39*, 475–543.

Srinivas, K. (1995). Representation of rotated objects in explicit and implicit memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *21*, 1019–1036.

Srinivas, K. & Schwoebel, J. (1998). Generalization to novel views from view combination. *Memory & Cognition*, *26*, 768–779.

Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, *68*, 77–94.

Immediate Integration of Syntactic and Referential Constraints on Spoken Word Recognition

James S. Magnuson (magnuson@psych.columbia.edu)
Department of Psychology, Columbia University
1190 Amsterdam Ave., MC 5501
New York, NY 10027 USA

Michael K. Tanenhaus (mtan@bcs.rochester.edu) and **Richard N. Aslin** (aslin@cvs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Rochester, NY 14627 USA

Abstract

We tested the hypothesis that syntactic constraints on spoken word recognition are integrated immediately when they are highly predictive. We used an artificial lexicon paradigm to create a lexicon of nouns (referring to shapes) and adjectives (referring to textures). Each word had phonological competitors in both form classes. We created strong form class expectations by using visual displays that either required adjective use or made adjectives infelicitous. We found evidence for immediate integration of form class expectations based on the pragmatic visual cues: similar-sounding words competed when they were from the same form class, but not when they were from different form classes.

Top-down constraints on word recognition

It is clear that we integrate top-down information when we interpret language. If someone tells us they put money in a bank, we understand that their money is in a vault and not buried next to a river. What is less clear is *when* and *how* we integrate top-down knowledge with bottom-up linguistic input.

One possibility is that language is processed in stages, with top-down information integrated after an encapsulated first-pass on the bottom-up input (e.g., Frazier & Clifton, 1996; Norris, McQueen & Cutler, 2000). The theory behind this genre of model is that optimal efficiency can be achieved by applying automatic processes that will almost always yield a correct result. In the rare event that the automatic result cannot be reconciled with top-down information, reanalysis would be required.

A second possibility is that top-down constraints are integrated immediately, with weights proportional to their predictive power (e.g., McClelland & Elman, 1986; MacDonald, Pearlmutter & Seidenberg, 1994; Tanenhaus & Trueswell, 1994). The theory behind constraint-based approaches is that a system can be made more efficient by allowing any sufficiently predictive information source to be integrated with processing as soon as it is relevant.

While a variety of results support constraint-based theories of sentence processing (see MacDonald et al., 1994), there is reason to believe that spoken word recognition is initially encapsulated from top-down constraints. Swinney (1979) and Tanenhaus, Leiman & Seidenberg (1979) provided the seminal results on this issue by examining whether all homophones are activated independent of context. Tanenhaus et al. presented participants with spoken sentences that ended with a syntactically ambiguous word (e.g., "they all rose" vs. "they bought a rose"). If participants were asked to name a visual target immediately at the offset of the ambiguous word, priming was found for associates both of the alternative suggested by the context (e.g., "stood" given "they all rose") and of homophones that would not fit the syntactic frame (e.g., "flower"). Given a 200-ms delay prior to the presentation of the visual stimulus, priming was found only for associates of the syntactically appropriate word. This suggests that lexical activation is initially based only on bottom-up information, and top-down information is a relatively late-acting constraint.

Tanenhaus & Lucas (1987) argued that this made sense given the predictive power of a form-class expectation. Knowing that the next word will be one of tens of thousands of nouns would afford virtually no advantage for most nouns (those without homophones in different form classes). Furthermore, expectations for classes like noun or verb might be very weak because modifiers can almost always be inserted before either class (e.g., "they just rose", "they bought a very pretty red rose"; cf. Shillcock & Bard, 1993).

Shillcock & Bard (1993) pointed out that there are form classes that should be more predictive than *noun* or *verb*, because they have few members: those made up of closed-class words. They examined whether /wʊd/ in a sentence context favoring the closed-class item, "would" (e.g., "John said that he didn't want to do the job, but his brother would, as we later found out"), would prime associates of its homophone, "wood", such as "timber" (compared with a context like "John said he didn't want to do the job with his brother's

wood, as we later found out"). They found priming for "timber" given the open-class context (favoring "wood") immediately after the offset of /wʊd/, but not given the closed-class context. The same result held when they probed half-way through the pronunciation of /wʊd/. This suggests the closed-class context was sufficiently constraining to bias the earliest moments of word recognition. A cloze test (in which participants were asked to supply the next word given the sentence contexts up to the word just prior to "would" or "wood", with the understanding that the word they supplied would not be the last in the sentence) confirmed that the closed-class context was much more predictive. Participants provided words of the same form class as the target most of the time for both cases, but were much more likely to provide the target given the closed-class context than the open-class context.

Shillcock & Bard's result is consistent with the constraint-based view that top-down information sources are integrated early in processing when they are sufficiently predictive. In the current experiment, we tested the hypothesis that even form class expectations for open-class words could constrain word recognition given a context with sufficient predictive power.

The Experiment

We hypothesized that form class could be sufficiently predictive to constrain initial activation if it were combined with strong visual and pragmatic expectations. For example, if there are four objects on a table – a brown purse, a purple book, a red ashtray, and a blue pen – and we ask you to pick one up, you would have strong expectations about how specific we would be in making reference to an item. For example, if we wanted the purse, you would expect to be asked, "pick up the purse" rather than "pick up the brown purse." Because of such conversational pragmatics (Grice, 1975), we would not expect subjects to experience strong competition between "purple" and "purse" as they hear "pur—," since if we wanted the book, we would ask for "the book," not "the purple book." But if there were brown and red purses, and purple and green books, given "pick up the pur–" we would expect little competition from *purse* – subjects would have a strong expectation to hear an adjective in this case.

Constructing such an experiment with real words poses significant problems. While there are many examples of cross-form class competitors in English, there are relatively few that are highly imageable and thus appropriate for our pragmatic manipulation. Even among these few, there is high variability in factors such as frequency and word length (e.g., purple-purse, dotted-dog, tan-tambourine, rough-rum).

Therefore, we extended an artificial lexicon paradigm that we previously developed to study the lexical neighborhoods of spoken words (Magnuson, Dahan, Allopenna, Tanenhaus & Aslin, 1998). An artificial lexicon allows precise control over such dimensions as phonological similarity and frequency of occurrence, as well as visual aspects of stimuli.

Table 1: The artificial lexicon.

	NOUN (shape)	ADJ (texture)	
1	pibo	pibʌ	1
2	pibe		
3	bupo	bupʌ	2
		bupe	3
4	tedu	tedi	4
		tedɛ	5
5	dotɛ	doti	6
6	dotu		
7	kagæ	kaga^{i}	7
		kagu	8
8	ga^{w}ku	ga^{w}kæ	9
9	ga^{w}ka^{i}		

We created a lexicon of nouns (referring to novel shapes) and adjectives (referring to textures). The lexicon (shown in Table 1) contained phonemic cohorts (e.g., /pibo/ and /pibʌ/) in different syntactic categories (e.g., /pibo/ was a noun and /pibʌ/ was an adjective) or the same category (e.g., another noun was /pibe/). Thus, the artificial lexicon allowed us to compare phonological competitors in same or different form classes with similarity precisely controlled. (Note that there are even fewer examples of real words with comparable phonological competitors in the same form class and another, and the possible sets are quite heterogeneous, e.g.: purple-purse-person, tattered-tan-tambourine.)

Participants learned the lexicon over two days of training. Instructions were given in an English context, with English word order (e.g., "click on the /pibʌ/ [adj] /tedu/ [noun]"). We created conditions in which the visual context provided strong syntactic expectations by constructing contexts in which adjectives were required (e.g., two examples of the shape associated with /pibo/, but with two different textures) or infelicitous (e.g., two different shapes, making the adjective superfluous, even if the shapes have different textures). If syntactic expectations in conjunction with pragmatic constraints embodied in the visual display can constrain word recognition early in processing, we should observe competition effects only between cohorts from the same syntactic form class.

Methods

Participants

Eight native speakers of English who reported normal or corrected-to-normal vision and normal hearing were paid for their participation. Participants attended sessions on two consecutive days.

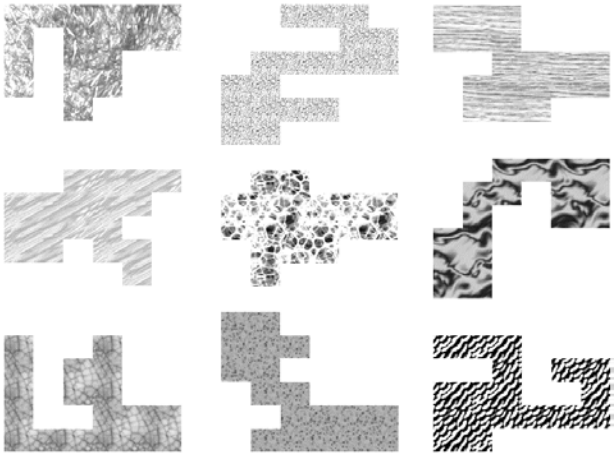


Figure 1: The 9 shapes and 9 textures.

Materials

The linguistic materials consisted of the 18 artificial words (9 nouns, referring to shapes, and 9 adjectives referring to textures) shown in Table 1. The auditory stimuli were produced by a male native speaker of English in a sentence context (e.g., "Click on the /bupe tedu/."). The stimuli were recorded using a Kay Lab CSL 4000 with 16 bit resolution and a sampling rate of 22.025 kHz. The mean duration of the "Click on the…" portion of the instruction was 475 ms for adjective instructions, and 402 ms for noun instructions. For adjective instructions, mean adjective duration was 487 ms, and mean noun duration was 682 ms. For noun instructions, mean noun duration was 558 ms.

We examined the neighborhoods our artificial words would fall into were they real words of English; none would be in a dense English neighborhood (9 had 0 neighbors, and 7 had 1 neighbor). (See Magnuson [2001] for evidence that artificial and native lexicons do not interact, even when artificial items are constructed to be maximally similar to real words.) The visual materials consisted of unfamiliar shapes generated by randomly filling 18 contiguous cells in a 6x6 grid. We selected a set of 9 subjectively dissimilar shapes. These shapes provided referents for the nouns. In addition, 9 textures were selected from among the set distributed with Microsoft PhotoDraw. Figure 1 shows each of the 9 shapes, with a different one of the 9 textures applied to each (note that picture quality was substantially higher on the computer display). Names were randomly mapped to shapes and textures for each participant.

Eye tracking

During the tests (see Procedure), eye movements were monitored using a SensoriMotoric Instruments (SMI) EyeLink eye tracker, which provided a record of point-of-gaze in screen coordinates at a sampling rate of 250 hz. Saccades and fixations were coded from the point-of-gaze data using SMI's software.

Eye movements were used because they are closely time-locked to speech in a properly constrained task. Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy (1995) found time locked fixations when subjects followed spoken instructions to perform a visually-guided task (e.g., "pick up the candle"). Because the subject must foveate the target item in order to efficiently follow the instruction, there is a functional link between the speech stimulus and dependent measure. This link avoids the pitfalls of interpreting eye movements described by Viviani (1990).

Allopenna, Magnuson & Tanenhaus (1998) extended this work to a time-course issue in spoken word recognition. Whereas studies using more conventional tasks had failed to find evidence for the activation of rhymes during lexical competition, eye tracking proved sensitive enough to detect the robust (if relatively weak) rhyme activation predicted by various models (e.g., Luce & Pisoni, 1998; McClelland & Elman, 1986). Dahan, Magnuson & Tanenhaus (2001) applied the approach to a debate regarding frequency effects in spoken word recognition. Competing theories made conflicting predictions at the level of time course; for example, some argued it kicked in as a late bias (Connine, Blasko & Titone, 1993). Dahan et al.'s eye tracking measures demonstrated that frequency has a continuous but gradual influence from the earliest moments of processing, leading to the appearance of a late locus in less sensitive paradigms.

The eye tracking paradigm imposes different constraints than more conventional paradigms, such as lexical decision. In a conventional task, the stimuli are typically decontextualized; there is nothing about the task that predicts what word one might hear next. In the eye tracking paradigm, the stimuli are presented in the context of a display of items. While this allows more naturalistic tasks, it might also allow strategic processing. For example, participants might activate lexical representations in response to the visual display prior to any bottom up information, or the displayed set of items might provide a verification set to guide recognition. There is no evidence for lexical activation prior to the bottom-up signal; fixation proportions map precisely onto emerging phonetic similarity over time. We have also found that recognition in this paradigm is not based on lexical activations constrained to the displayed items: artificial lexical items (Magnuson, Tanenhaus, Aslin & Dahan, 1999, in preparation) and real words (Magnuson, 2001) in dense neighborhoods (i.e., with many or very frequent neighbors) are recognized more slowly than words from sparser neighborhoods, even when the neighbors are not displayed. This suggests the representations of the neighbors were activated and competed for recognition.

In summary, eye movements provide an extremely sensitive time course measure of lexical activation and competition. We need just such a measure to resolve the time course debate we are concerned with here: when

are top-down and bottom-up information integrated during spoken language understanding?

Procedure

Participants were trained and tested in sessions on two consecutive days. Each session lasted between 90 and 120 minutes. On day 1, participants were trained first on the nouns in a two-alternative forced choice (2AFC) task. On each trial, two shapes appeared (both with solid black texture) and the participant heard an instruction to click on one (e.g., "click on the bupo"). The auditory stimuli were presented binaurally through headphones (Sennheiser HD-570) using standard Macintosh Power PC digital-to-analog devices.

When the subject clicked on an item, one item disappeared, leaving the correct one, and its name was repeated. There were 14 repetitions of each item, split into 3 blocks of 48 trials. Items were not repeated on consecutive trials, and were ordered such that every item was repeated 7 times every 72 trials. Following the 2AFC blocks, noun training continued with 3 blocks of 4AFC, with identical ordering constraints and numbers of trials. Each shape appeared equally often as a distractor.

Adjective training then began. First, participants saw two exemplars of one shape, with different textures. They heard an instruction, such as "click on the bupe pibo". Since they already knew that, e.g., "pibo" referred to one of the shapes, participants found it transparent that "bupe" referred to one of the textures. As in the noun training, after they clicked on one item, the incorrect one disappeared and the full name was repeated. Each adjective and each noun was a target in 8 trials in each block; each adjective was randomly paired with 8 different nouns in each block. After three 48-trial 2AFC blocks, there were three 4AFC blocks, with four exemplars of the same shape with four different textures. These were followed by three more blocks of 4AFC, but with two exemplars each of two shapes, each with a different texture (requiring participants to recognize both the adjective and noun).

After this, a more complex training regime began. On some trials, four different shapes appeared. On others, two pairs of shapes appeared. On every trial, each shape had a different texture. On trials with two pairs of shapes, an adjective was required to make unambiguous reference, and the full referent was specified on such trials (e.g., "click on the bupe pibo"). On trials with four different shapes, the adjective was not required – each item could be identified unambiguously by the name of the shape, and so only the noun was specified in the instruction (e.g., "click on the pibo"). Using the adjective would be infelicitous, on Grice's (1975) maxim of quantity (one should not over-specify, which is the observed tendency in natural conversation). Each adjective was repeated 8 times in every block of 144 trials, paired each time with a different, randomly selected noun. Each noun was repeated as the target item 8 times in the 4-noun trials. Trials were presented in blocks of 48. Participants completed 3 blocks of this mixed training on Day 1. On Day 2, they completed 12 more, which comprised the entire training phase on Day 2.

After each 48-trial block, the participant saw a summary of his or her accuracy in that block. To motivate participants, we told them that each training segment would continue until they reached 100% accuracy. Typically, we moved to each successive training phase after the number of blocks listed above for each segment, except in a few rare cases where participants were below 90% accuracy after the specified number of blocks, in which case training continued for another 1-2 blocks.

Each day ended with a 4AFC test with no feedback. We tracked participants' eye movements during the test. There were six basic conditions in the test. In the *noun baseline condition*, there were four different shapes, and no shape's or texture's name was a competitor of the target noun. In the *noun plus noun cohort condition*, there were four shapes, and one of them was a cohort to the target (e.g., the target might be /pibo/, and /pibe/ would also be displayed), but no shape had the target's adjective cohort texture applied (e.g., no shape would have the /pibʌ/ texture). In the *noun plus adjective cohort condition*, four different shapes were displayed. The noun cohort was not displayed, but the adjective cohort was (e.g., a distractor might be /pibʌ tedu/). In these conditions, the instruction would only refer to the noun (e.g., "click on the pibo").

In the other three conditions, two exemplars of two different shapes were displayed, requiring the adjective to be used in the instruction. In the *adjective baseline condition*, none of the distractor textures were cohorts of the target, and neither were any of the nouns. In the *adjective plus adjective cohort condition*, one of the non-target textures was a cohort to the target (e.g., the target might be /tedi dotu/, and one non-target might be /tedɛ bupo/), but no noun cohorts of the target would be displayed. In the *adjective plus noun cohort condition*, none of the distractors would have textures that were cohorts to the target texture, but a noun cohort would be displayed (e.g., given /tedi dotu/ as the target, /bupe tedu/ might be included).

The following scheme was used to ensure that each adjective and target appeared 9 times as targets in the test. Note that nouns and adjectives either had one competitor in each form class, or two in the opposite form class. Nouns with noun cohorts appeared in six *noun baseline* trials, two *noun plus noun cohort* trials, and once in the *noun plus adjective cohort condition*. Nouns with two adjective cohorts appeared in 7 *noun baseline* trials, 0 *noun cohort* trials, and two *noun plus adjective cohort* trials. The same pattern was used with adjective conditions, giving a total of 162 test trials.

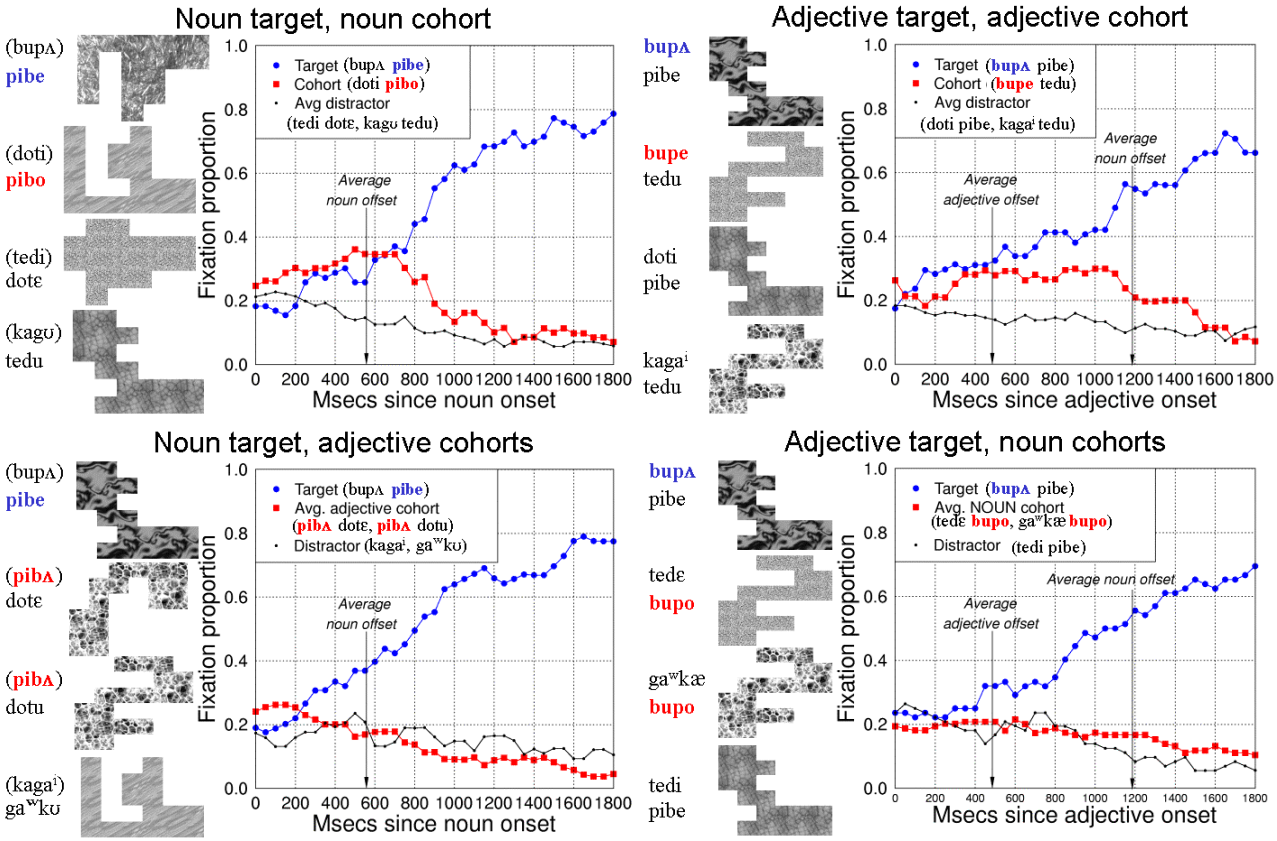


Figure 2: Results from the 4 critical conditions. The top panels show competition between within-class cohorts. The bottom panels show the failure to find competition for cohorts from different form classes.

Results

Participants attained high accuracy quickly (two failed to reach ceiling levels of accuracy, performing at less than 90% correct on the test on Day 2, and their data was excluded from the analyses). Mean accuracy on nouns and adjectives was 96% at the end of Day 1, and 98% at the end of Day 2. The results from the test on Day 2 are shown in Figure 2. Examples of possible stimulus items are shown to the left of each panel (these would be arranged around the central fixation cross in an actual experimental display). Note that in the cross-form class conditions (*noun with adjective cohorts* and *adjective with noun cohorts*) there were two cohorts in the display. This was necessary in the case of the *adjective plus noun cohort condition*; in order for the display to demand that an adjective be used, two exemplars of two different shapes had to be displayed. To make the *noun plus adjective cohort condition* comparable, two items were displayed with textures whose names were cohorts to the noun target.

The results show strong, immediate effects of the form-class constraints on lexical access. Compare the upper and lower panels of Figure 2. While strong cohort effects are apparent in the upper panels (the within-form class competitor conditions), there is no evidence for cohort effects in the lower panels (between-form class conditions). Analyses of variance on mean fixation proportion in the noun conditions over the window from 200 ms (where we first expect to see signal-driven fixations, since it takes 150 – 180 ms to plan and launch saccades in much simpler tasks) to 1400 ms (where the target proportions asymptote) confirm the trends. There was a reliably greater proportion of fixations to the cohort than to the distractors in the *noun plus noun cohort condition* (cohort=.25, mean distractor=.12; $F(1, 11)=10.16$, $p=.009$), but not in the *noun plus adjective cohort condition* (cohort=.15, mean distractor=.15). The same was true for the adjective conditions, over the window from 200 to 1800 (the window was extended because of the longer lag prior to disambiguation). There were reliably more fixations to the cohort in the *adjective plus adjective cohort condition* (cohort=.22, mean distractor=.15; $F(1,11)=7.2$, $p=.02$), but not in the *adjective plus noun cohort condition* (cohort=.16, mean distractor=.15, $p=.59$).

Discussion

The results demonstrate that higher-level linguistic constraints (in this case, syntactic expectations based on a visually-defined referential context) influence even the earliest moments of lexical access when the constraints are highly-predictive. Phonemically similar

items competed only when they were from the same form class. This suggests, contra strong modularity (e.g., Fodor, 1983), that lexical activation can be constrained given a highly informative context.

In future research, it will be important to establish the limits of such effects. It may be the case that form class constraints would be weaker were there more members of each form class (as predicted by the argument that a noun or verb expectation in English is an extremely weak constraint). We are currently exploring this possibility with an expanded lexicon.

It is possible that visual/pragmatic constraints swamp lexical activation, and turn the display into a verification set. To eliminate this possibility, we will use a neighborhood density manipulation. We should find faster increases in target fixations for items in sparse neighborhoods *in addition* to the form class/pragmatic effects observed here.

The timing of these sorts of effects will be informative about how different classes of constraints are integrated in real-time spoken word recognition. The current results provide a starting point for further explorations while demonstrating that the artificial lexicon paradigm can be adapted to a wide range of microstructural issues in spoken word recognition. Moreover, they suggest that the failure to find immediate effects in earlier studies does not reflect an architectural property of the word recognition system (i.e., encapsulation), but rather reflects the pattern predicted by constraint-based models when contextual constraints are only weakly predictive.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38, 419-439.

Connine, C.M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. Journal of Experimental Psychology: Learning, Memory & Cognition, 19, 81-94.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology, 42, 317-367.

Frazier, L. & Clifton, C. (1996). Construal. Cambridge, MA: MIT Press.

Fodor, J. A. (1983). The Modularity of Mind. Cambridge, MA: MIT Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), Syntax and Semantics, Vol. 3, Speech Acts (pp. 41-58). NY: Academic Press.

MacDonald, M. C, Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676-703.

Magnuson, J. S. (2001). The Microstructure of Spoken Word Recognition. Unpublished doctoral thesis, University of Rochester Department of Brain and Cognitive Sciences.

Magnuson, J. S., Dahan, D., Allopenna, P. D., Tanenhaus, M. K., and Aslin, R. N. (1998). Using an artificial lexicon and eye movements to examine the development and microstructure of lexical dynamics. In Gernsbacher, M.A., & Derry, S.J. (Eds.), Proc. of the Twentieth Annual Conference of the Cognitive Science Society, 651-656. Mahwah, NJ: Erlbaum.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (1999). Spoken word recognition in the visual world paradigm reflects the structure of the entire lexicon. In M. Hahn & S. Stoness (Eds.), Proceedings of the Twenty First Annual Conference of the Cognitive Science Society, pp. 331-336. Mahwah, NJ: Erlbaum.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (in preparation). The microstructure of spoken word recognition: Insights from investigations with artificial lexicons.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. Behavioural and Brain Sciences, 23, 299-370.

Saslow, M. G. (1967). Latency for saccadic eye movement. Journal of the Optical Society of America, 57 (8), 1030-1033.

Shillcock, R. C. and Bard, E. G. (1993). Modularity and the processing of closed-class words. In G. T. M. Altmann and R. Shillcock (Eds.), Cognitive Models of Speech Processing: The Second Sperlonga Meeting, pp. 163-185. Erlbaum.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. J. Verbal Learning & Verbal Behavior, 15, 545-569.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J. Verbal Learning & Verbal Behavior, 18, 427-441.

Tanenhaus, M. K., and Lucas, M. M. (1987). Context effects in lexical processing. Cognition, 25, 189-234.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. C. (1995). Integration of visual and linguistic information is spoken-language comprehension. Science, 268, 1632-1634.

Trueswell, J. C. & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In Clifton, C., Frazier, L. and Rayner, K. (Eds.) Perspectives in Sentence Processing. Erlbaum: Hillsdale, NJ.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), Eye Movements and Their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research V4. Amsterdam: Elsevier.

Three-year-old Children's Use of Category Labels and Motion in Drawing Inferences about Animal Kinds

Benise S.K. Mak (benise@hku.hk)
Department of Psychology, The University of Hong Kong,
Pokfulam Road, Hong Kong, PR China

Alonso H. Vera (avera@arc.nasa.gov)
Cognition Lab, NASA Ames Research Center,
Mailstop 262-4, Moffett Field, CA 94035-1000, USA

Lap Yan Lo (h9605361@hkusua.hk)
Department of Psychology, The University of Hong Kong,
Pokfulam Road, Hong Kong, PR China

Abstract

Linguistic labels, information about category memberships, have been found to be more important than perceptual information in guiding young children's inferences about animal kinds. However, perceptual information of static shape cues has often been stressed. A recent study has shown that young children tended to use dynamic perceptual cues, such as motion, more often than static shape cues to make categorical judgments. The overriding effects of linguistic labels over perceptual information in young children's inferences need to be re-examined. This paper was an attempt to examine how 3-year-old children use category labels and motion cues to draw inferences about animal kinds. Data showed that preschool children tended to use motion more often than labels when confronted with a choice between labels and motion. This provides support for our view that the role of category labels in young children's categorical judgments is not as important as what has been suggested in previous studies.

Introduction

The Importance of Category Labels

The importance of linguistic information in preschool children's inductive inferences about animal kinds has well been demonstrated in a series of studies by Gelman and her colleagues (e.g., Gelman & Markman, 1986, 1987; Gelman & Coley, 1990). In their studies, the extent to which children use category labels and perceptual appearance was examined. For instance, 2½ years old children were found to be more likely to assign a property, e.g., "lives in a nest", from a bluebird to an atypical bird, dodo, which looked different from the target but shared the same label "bird" than to a dinosaur which looked similar to the bluebird but carried a different label "dinosaur" (Gelman and Coley, 1990). However, when the stimuli were not labeled, they tended to draw inferences based more on perceptual similarity than on the similarity of verbal labels. The young children were able to go beyond perceptual appearance and use linguistic labels to draw inferences, suggesting that linguistic information is more powerful than perceptual information for preschoolers to draw accurate categorical judgments. As Gelman and Coley (1990) stated that for young children "language conveys important information beyond that which meets the eye" (p.804).

Despite of this, the role of category labels remains to be determined as perceptual characteristics have generally been taken as static perceptual cues in Gelman et al.'s studies. Perceptual cues should also include dynamic properties, such as motion.

The Importance of Motion Information

In a recent study by Mak and Vera (1999), 4 and 7-year-old children have been found to categorize animals based more on motion similarity than on static shape similarity. For instance, they were more likely to categorize, for example, a donkey with an antelope than with a horse when the antelope and the donkey were shown to jump in the same manner even thought the donkey looked more similar to the horse than to the antelope. The children tended to infer that the donkey shared the same property of "having poor vision" that was ascribed to the antelope rather than the property of "having good vision" ascribed to the horse. The role of motion information has been stressed.

As such, this paper tried to look further into the role of category labels and include motion information to study young children's inductive inferences about animal kinds. We will argue that the effectiveness of verbal labels is not as important as Gelman et al. have suggested.

Effectiveness of Category Labels?

Doubts have been raised about the significance of verbal labels in guiding children's categorical judgments. Although Gelman and Coley's (1990) studies found that young children were ready to draw inferences about animals based more on label similarity than on static shape similarity, their use of linguistic information seems to rely on the correspondence between the label and the perceptual information of the animal stimuli. If children find the labels are not congruent with the animal stimuli, they would not use the linguistic information. They would do so only when they believe that the labels and the stimuli match with one another. This has been illustrated in some of Gelman et al.'s studies.

For instance, Davidson and Gelman (1990) found that young children did not use familiar verbal labels to draw inferences about novel animal categories. In this study, 4- and 5-year-old children were tested with some imaginary animals; in one of the experimental conditions, familiar category labels (e.g., "cow" and "deer") were used. Results, which were different from those in Gelman and Coley's (1990) study, showed the preschool children tended not to draw inferences based more on label similarity than on perceptual similarity. This might be due to the fact that the children did not find the labels "cow" and "deer" congruent with the imaginary stimuli and therefore did not make judgments based on the labels.

Even when novel linguistic labels (e.g., "zav" and "traw") were used, the children did not use the novel labels and were more likely to use labels than perceptual appearance to draw inferences. The children, in this instance, might not to be able to relate the novel labels to the imaginary animals; confronted with a choice between perceptual and linguistic information, they opted for static perceptual cues.

Therefore, if children find that the familiar labels and the animal stimuli do not match with one another, they would not use category labels. Only when children believe that the labels go well with the stimuli do they begin to find the linguistic information useful. Here are some more examples.

In Gelman and Markman's (1987) study, familiar animal categories (e.g., "cat" and "skunk") were used. Three- and 4-year-old children were found to be more likely to assign a property, "can see in the dark", from a target cat to another cat (which was shown in different coloring and posture) than to a skunk (which looked similar to the target cat) when the animal stimuli were labeled accordingly. However, the children also succeeded in doing so even when verbal labels were not provided. A follow-up study showed that children were able to determine the categorical memberships of the stimuli perceptually when the stimuli were not labeled. The animal stimuli seemed to be too familiar to the young children that they could make use of the subtle perceptual cues to draw accurate inferences and did not need to rely on the linguistic information (Gelman & Markman, 1987). Gelman and Markman (1987) unexpectedly found that familiar linguistic labels could not help the 3- and 4-year-olds. Only when 2½-year-old children were tested did they begin to find that verbal labes could help young children to draw inferences about familiar animals (e.g., "bird" and "dinosaur"). This may be due to the fact that the 2½ year-olds were somewhat but not completely familiar with the animal stimuli. They were not able to use the subtle perceptual cues to determine the category memberships of the stimuli and needed to rely on the category labels provided to draw accurate inferences.

This may also be true in the 3- and 4-year-olds in Gelman and Markman's (1986) study who were found to be guided by the similarity and dissimilarity of linguistic labels in drawing inferences about familiar animals (e.g., "bat", "bird", "dolphin", "fish"). In the experiment, young children were shown, for example, a "bird" set which included two target animals, a flamingo and a bat (which looked very different from one another, and a test animals), a blackbird (which looked more like the bat than the flamingo). In the conflict condition where the flamingo and the blackbird were labeled a "bird" and the bat a "bat", the children were more likely to categorize the blackbird with the flamingo than with the bat even though the blackbird looked more similar to the bat than to the flamingo. However, in the no-conflict condition where the bat and the blackbird were labeled a "bat" and the flamingo a "bird", the children tended to categorize the blackbird and the bat together. In other words, the children tended to make categorical judgments based more on linguistic similarity than on perceptual similarity. They were more likely to categorize the blackbird and the flamingo together when they shared the same label "bird" than when the blackbird was labeled a "bat" instead of a "bird", suggesting that the young children were not so familiar with the animal categories, birds and bats, that they were willing to accept the names, "bird" and "bat", to be used to label the blackbird. If the stimuli were completely familiar to them, they would not have been guided by the similarity and dissimilarity of verbal labels, like the 4 and 5-year-olds in Gelman and Markman's (1987) later study.

To summarize, the effectiveness of familiar linguistic labels in guiding young children's inferences is rather limited, which relies on two major factors – whether children find the familiar linguistic labels and the perceptual information of the animal stimuli match with one another and their familiarity with the stimuli. Only when children believe that the linguistic labels are not in conflict with the stimuli and they are somewhat but not completely familiar with the stimuli does the linguistic information become useful. Otherwise, linguistic labels cannot help children to draw accurate

Table 1

Animation items, category labels and target properties used.

Pair	Test displays			Target displays			
	Animal	Motion	Label	Animal	Motion Same/Different	Label Same/Different	Property
1	Horse	Walk	"Horse"	Donkey	Walk/Jump	"Horse"/ "Donkey"	Good vision
2	Quail	Walk	"Quail"	Sparrow	Walk/Hop	"Quail"/ "Sparrow"	Good hearing

inferences. The effectiveness of linguistic labels in guiding children's inferences is not as significant as Gelman et al. have proposed.

The Present Study

In view of the above, it seems to be reasonable to believe that perceptual information, including both static and motion cues, may not be relatively unimportant in comparison to familiar verbal labels. It is evident that young children not only rely on linguistic labels but also use perceptual characteristics to draw inferences about animal kinds. In this paper, we would like to like to examine to what extent young children use familiar linguistic labels and motion cues to draw inferences.

It is postulated that young children, such as 3-year-old children (who have been found to be ready to draw inferences based on the similarity and dissimilarity of familiar linguistic labels), would use both verbal labels and motion cues to draw inferences about animals. However, when confronted with a choice between labels and motion, they would tend to use motion information more often than linguistic information.

Method

Design

To examine the extent to which 3-year-olds use familiar linguistic labels and motion cues to draw inferences about animal kinds, an inductive methodology used in previous studies (Gelman & Markman, 1986, 1987; Gelman & Coley, 1990; Mak & Vera, 1999) was adopted. Here, young children were required to consider a pair of animals, a target and a test. They were taught a new property about the target animal and were asked whether the target property was true for the test animal.

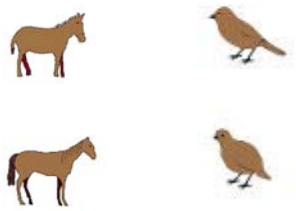
The independent variable was three *motion/label* conditions: *label and motion*, *label only* and *motion only*. The dependent measure was children's responses on the inference questions: whether they would generalize the target properties to the test animals.

Participants

Two hundred and twenty 3-year-old children from three kindergartens in Hong Kong participated: 110 girls and 110 boys, ranging in age from 3 years to 3 years 11 months with a mean age of 3 years 7 months.

Stimuli

There were two pairs of animal stimuli: (1) a donkey and a horse, and (2) a sparrow and a quail. Each pair consisted of animals with similar appearance and two series of animations: same and different movement. The movement of the animals in the animations was slightly slower than the corresponding real motion, but they all moved in the same speed. Also, all the stimuli were drawn in the same brown color with black outline. The drawings of the animal stimuli used for the animations are shown in Fig. 1, and the details of the animation items are shown in Table 1.



Pair 1 Pair 2

Figure 1
The drawings of animal stimuli used for the animation items.

Control Studies

Property Controls

Some control studies were run to make certain if 3-year-old children have any biases in assigning the target properties to the test animals which were shown to be static or with certain movement and linguistic labels. In these controls, children were shown the test animals alone (i.e., the horse and the quail) one at a time and were asked if the target properties (i.e., "good vision" and "good hearing") were true for the horse and the quail respectively. Results are summarized in Table 2.

Table 2
Percentages of "yes" responses for 3-year-old children in the Property Controls.

Control Condition	Percentage
Static	52.5
Label only	57.5
Motion only	52.5
Label and Motion	55

n=20 (half boys and half girls)

Data showed that 3-year-olds performed at about chance level in saying that the target properties were true for the test animals, showing no biases in giving "yes" responses to the test animals.

Static Controls

Two animals of similar appearance were adopted. To confirm if young children were ready to draw inferences based on the static perceptual similarity of the animal stimuli, a static control study was also conducted. In this control, 3-year-old children (10 boys and 10 girls) were presented with two sets of animals but one set at a time without motion. Children were tested to see if they would generalize the target properties to the test animals based on static shape information alone. Data showed that the young children were willing to do so 85% of the time, which was significantly above chance level, $p<.0005$ (1-tailed), suggesting that 3-year-olds were ready to draw inferences based on the static shape similarity of the animal stimuli.

Procedure

This was a between-subject design, so that a child (e.g., being assigned to the *Same-Label and Different-Movement* condition) was tested with two pairs of animals. One was the *donkey/horse* pair; both were labeled a "horse" but were shown to move in different manners (i.e., the donkey jumped and the horse walked). The other one was the *sparrow/quail* pair; both shared the same name "quail" but were shown to move differently (i.e., the sparrow hoped and the quail walked). The presentation order of the two pairs was counterbalanced across participants. Each child was tested individually. Throughout the experiment, instructions were given in Cantonese, the major Chinese dialect used in Hong Kong. Each child was tested individually

Children were shown two pairs of animals, one pair at a time. They were first taught a new property about the target animal in each pair and were then asked to infer whether the target property applied to the test animal. The animal stimuli were labeled according to the label conditions. Taking the *donkey/horse* pair in the *Same (Different) Label* conditions as an example, the experimenters first pointed to the donkey and said, "See this *horse* (*donkey*). This *horse* (*donkey*) has *good vision*; it can see things clearly at a great distance." The experimenter then pointed to the horse and said, "See this *horse* (*horse*). Does this *horse* (*horse*) have *good vision*, like this *horse* (*donkey*) (referring to the donkey) that can see things clearly at a great distance? Or, this *horse* (*horse*), unlike this *horse* (*donkey*), does not have *good vision*?"

At the end of the experiment, children in the *Same Label* groups, who were given somewhat misleading labels for the target animals, were shown the stimuli again and were told that the experimenters had made a mistake in saying that the donkey (sparrow) was a *horse* (*quail*). The experimenters further explained to the children that its proper name should be *donkey* (*sparrow*) instead of *horse* (*quail*).

Results

Children's responses on the inference questions were coded as 1 when they said "yes" (i.e., generalizing the target properties to the test animals) and 0 when they said "no" (i.e., not generalizing the target properties to the test animals). These scores were summed within participants, and the score for each participant ranged form 0 to 2. For each condition, a one-sample *t*-test was conducted to examine if children performed significantly above or below 50% chance level. Results are summarized in Table 3.

Motion only condition

Data in these conditions clearly show that 3-year-old children were ready to draw inferences about the animal stimuli based on the similarity and dissimilarity of the motion information. When two animals with similar appearance were shown to move in the same manner, the young children tended to infer that the test animals shared the same property ascribed to the target animals (87.5% of the time, significantly above chance level); when the two animals moved differently, the children tended not to do so (20% of the time, significantly below chance).

Table 3
Percentages of "yes" responses for 3-year-old children in *Label only*, *Motion only* and *Label and Motion* conditions.

Condition		Percentage	
Motion only	Same	87.5	***
	Different	20	++
Label only	Same	90	***
	Different	15	+++
Label and Motion	Same Label/Different Motion	10	+++
	Different Label/Same Motion	82.5	***

n=20 (half boys and half girls)
*** above chance, $P<.0005$, one-tailed
+++ below chance, $P<.0005$, one-tailed

++ below chance, $P<.005$, one-tailed

Label only condition

Three-year-old children were also found to draw inferences based on the similarity and dissimilarity of the linguistic labels provided in the experiment. When two animals with similar appearance shared the same name, the 3-year-olds tended to agree that the target properties were also true for the test animals (90% of the time, significantly above chance). The children did not do so (15% of the time, significantly below chance) when the two animals were labeled with different names.

Label and Motion condition

In these conditions, label similarity and motion similarity were in contrast with one another. Results indicate that the 3-year-olds were more likely to draw inferences about the animal stimuli based on motion information than on linguistic information. When two animals moved in the same manner, the children tended to draw inferences 82.5% of the time (significantly above 50% chance level) even though the stimuli were labeled with different names; when two animals moved in different manners, they tended to draw inferences 20% of the time (significantly below chance) although the animal stimuli shared the same name.

In order to have a clearer picture of how children changed their choices between linguistic and motion information, 3-year-old children's responses in *Label only* and *Label and Motion* conditions are shown in Figure 2.

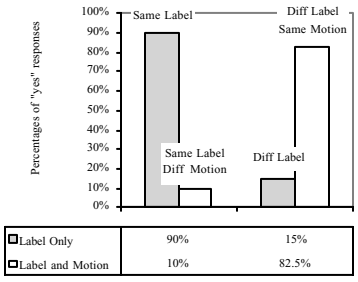


Figure 2
Percentages of "yes" responses for 3-year-old children in *Label only* and *Label and Motion* conditions.

This figure shows that 3-year-olds tended to shift their choices from linguistic labels to motion when contrasting motion information was introduced in addition to the linguistic information. Children drew inferences 90% of the time from a target to a test animal when two animal stimuli shared the same name in the *Label only* condition, whereas they made inferences only 10% of the time when additional different movement patterns were introduced to the stimuli in the *Label and Motion* condition. Moreover, children drew inferences 15% of the time when two stimuli were labeled with different names in the *Label only* condition, while they gave "yes" responses 82.5% of the time when the stimuli were shown to move in the same manner even though they were labeled with different names in the *Label and Motion* condition.

However, 3-year-olds did not change much about their choice of motion information when additional contrasting label information was introduced. Children's responses in *Motion only* and *Label and Motion* conditions are shown in Figure 3.

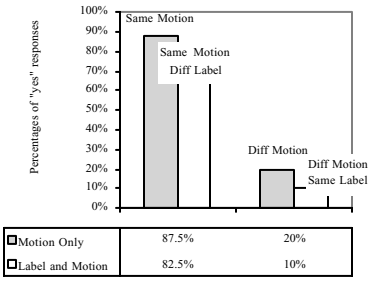


Figure 3
Percentages of "yes" responses for 3-year-old children in *Motion only* and *Label and Motion* conditions

This figure shows that children drew inferences 87.5% of the time when two animals moved in the same manner in the *Motion only* condition and did likewise 82.5% of the time even when different labels were ascribed to the stimuli in the *Label and Motion* condition. Moreover, when two animals moved in different manners in the *Motion only* condition, children made inferences only 20% of the time, and they drew inference 10% of the time even though the same label was ascribed to the stimuli in the *Label and Motion* condition.

Discussion

The current findings provide support for our hypothesis that linguistic information may not be relatively more important than perceptual information, including both static and dynamic characteristics, in guiding young children's inferences about animal kinds. Data show that 3-year-old children were ready to use both category labels and motion information to draw inferences. However, having to make a choice between labels and motion, children tended to use motion more often than familiar category labels.

Also, this seems to provide support our view that the effectiveness of category labels in guiding young children's inferences about animals may be rather limited. The importance of linguistic labels that has been suggested by previous studies need to be re-thought.

Although to what extent children use category labels and motion was compared in this study, we are not suggesting that linguistic and perceptual information play distinct roles in young children's inferences (for this argument, see also Gelman & Medin, 1993; Jones & Smith, 1993). Evidence has shown that children use both linguistic and perceptual (including static and dynamic) information to draw inferences. However, how static shape cues, motion and linguistic labels interact to guide children's inferences remains to be determined in future studies.

Acknowledgments

This research was supported by a CRCG Grant (#10203543) from the Hong Kong University Research Committee.

References

Davidson, N.S., & Gelman, S.A. (1990). Inductions from novel categories: The role of language and conceptual structure. *Cognitive Development, 5*, 151-176.

Gelman, S.A., & Coley, J.D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26 (5)*. 796-804.

Gelman, S.A., & Markman, E.M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Gelman, S.A., & Markman, E.M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58*, 1532-1541.

Gelman, S.A., & Medin, D.L. (1993). Commentary: what's so essential about essentialism? A different perspective on the interaction of perception, language and conceptual knowledge. *Cognitive Development, 8*, 157-167.

Jones, S.S., & Smith, L.G. (1993). The place of perception in children's concepts. *Cognitive Development, 8*, 113-139.

Mak, B.S.K., & Vera, A.H. (1999). The role of motion in children's categorization of objects. *Cognition, 71*, B11-B21.

Incorporating Cognitive Styles into Adaptive Multimodal Interfaces

Halima Habieb Mammar (habieb@ictthp.insa-lyon.fr)
Laboratoire d'Interaction Collaborative Télé-enseignement Téléactivités,
Edifice Léonard de VINCI 21 Avenue Jean Capelle INSA de Lyon 69211 Villeurbanne Cedex - FRANCE

Franck Tarpin Bernard (tarpin@gprhp.insa-lyon.fr)
Laboratoire d'Interaction Collaborative Télé-enseignement Téléactivités,
Edifice Léonard de VINCI 21 Avenue Jean Capelle INSA de Lyon 69211 Villeurbanne Cedex - FRANCE

Abstract

Many applications accessible through the web suffer from a noticeable lack of support in adapting the information presentation to users. The way users learn differs from an individual to an other, if not for the same individual from an application to an other one. These individual differences affect the learning style of users. They are classified into 3 categories which are: *affective*, *cognitive* and *physiologic styles*. There is little research to examine how to design adaptive systems based on user's cognitive styles. In this paper, we are focusing on user cognitive styles definition and suggest a technique in the design of an adaptive hypermedia system. We investigate the selection of the output modality that best tailor the user profile.

In section 1 we introduce the problematic of learning from net-structured knowledge then we define the cognitive styles. In section 2, we present the main cognitive styles which are the most mentioned in literature. The taxonomy of these cognitive styles and techniques to assess them are detailed in section3. In the last section, we present the structure of our site and model. We investigate the relationship between the cognitive style and the filtering process of the outcome modalities. For the development of the system, we have chosen the two technologies : XML and ASP.

Introduction

In e-learning systems, the user is confronted to lessons, exercises, games... which, on the one hand are relevant to his/her needs and preferences i.e. educational level, domain knowledge (expert or novice) but which, on the other hand, do not take into account his/her abilities for assimilation, memorization, etc. which are parts of the cognitive abilities (Lemaire, 1999). In some cases, the learner gives up the game or the exercise because of the frequent situations of defeats. In other cases, he/she tries hard to make his/her best in order to avoid these situations which overload his/her abilities. The aim of our project consists in developing an adaptive multimodal interfaces where individual cognitive styles are considered. Indeed, the adaptation process deals with the estimation of the document combination (in other word, the multimodal interface) which is the most compatible with the cognitive profile.

In the following sections, we present our site then we define the term "cognitive styles", assess the main styles encountered in literature then discuss the taxonomy of ways employed to measure the cognitive style. Later in section 4, we describe the functionalities of our site and the model we adopted to establish the relationship between the cognitive style and the filtering process of the outcome modalities.

Our Site

In collaboration with the society SBT, we work on an interactive web site for a supervised cognitive training (www.happyneuron.com). During each training session (i.e. each connection), the user executes a set of exercises that the system suggests. Presented into a playful and cultural dimension, the exercises vary in difficulty's level, speed… in order to entertain the user (Habieb-Mammar et al, 2001).

A database stores normalized data (means and standard deviations) for each variant of exercise and family of population distinguishing gender, level of education and age (Tarpin-Bernard et al, 2001). The current statistics show that since the web site has been opened to the public, the number of performed exercises exceeds 400.000. Comparing the trainee's results and the normalized data we progressively build his/her cognitive profile. Thanks to it, the system advises the elderly user in the choice of exercises.

In this context, we built an evaluation module composed of ten precise exercises that allows to quickly build a cognitive profile. Then, this profile, which is quite stable, can be used in very different context. Our first purpose is to elaborate an adaptive multimedia course on the brain. Depending on one's profile, the lesson will be presented using the most adapted medias.

Before describing the relationship between cognitive styles and interactive styles let start with the definition of these cognitive styles and the methods of their assessment.

What are Cognitive Styles?

Cognitive styles refer to a person's habitual, prevalent, or preferred mode of perceiving, memorizing, learning, judging, decision-making, problem-solving (Dufresne, 1997).

Individual differences about how people carry out tasks involving these functions may constitute a style if they appear to be: pervasive, which means that they emerge consistently in different contexts, independently of the particular features of situation; or stable, that is, they are always the same at different times.

They are one of the most stable user characteristics overtime (Dufresne, 1997). they are consistent across a variety of situations, as opposed to user knowledge or experience that are more specific and evolving . Many research have shown the importance of cognitive styles in the area of HCI and their implication in the interface design (Muylwijk et al, 1983).

Cognitive styles induce persons to adopt similar attitudes and behaviors in a variety of domains they concern (Daniels, 1996). Cognitive styles are important in determining the most effective interface for a particular category of user, especially in the formative stages of an interaction (Fowler et al, 1985).

They can be conceptualized as a cross-road of thinking, personality, and motivation. In fact they concern the kind of strategies which an individual tends to apply when he/she faces a situation or the preferred way of processing information.

The Main Cognitive Styles

Field Dependence

The first style we introduce is: field-independent style. People tend to have good analytical and cognitive restructuring skills. They will actively reorganize information according to contextual demands and impose structure when necessary according to their experience. They are likely to form a mental model of the situation before proceeding with their task. Field-independent people seem to follow more easily a restructuring approach and use internal referents in other situations (Antonietti et al, 2000).

Field-dependent people tend to adopt a passive approach in learning and problem solving. They prefer to be guided and to rely on external referents. Perception is dominated by the prevailing field.

When internal referents are less available, field dependent people are more likely to respond to the dominant properties of the field as given.

Lesser use of restructuring may handicap field dependent people in unstructured situations. field dependent people may need more explicit instructions in problem solving strategies or more exact definitions of performance outcome than field independent, who may even perform better when allowed to develop their own strategies (Witkin et al, 1981).

However, the restructuring process occurs only when the field lacks organization. When the material to be learned is presented in an already organized form, so that structuring is not particularly called for, field dependent and field independent people are not likely to differ in their behavior and learning (Antonietti et al, 2000).

In general field independent subjects :

- Perceive objects as separate from the field;
- Can disembed relevant items from non-relevant items within the field;
- Provide structure when it is not inherent in the presented information;
- Reorganize information to provide a context for prior knowledge;
- Tend to be more efficient at retrieving items from memory.

Conversely, field dependent subjects:

- Rely on the surrounding perceptual field;
- Have difficulty attending to, extracting, and using non salient cues;
- Have difficulty providing structure to ambiguous information;
- Have difficulty restructuring new information and forging links with prior knowledge;
- Have difficulty retrieving information from long-term memory.

The test of field dependent-independent subjects is done through several exercises where individual are asked to remember shapes or other types of information whether they where presented in significant context (Fig. 1) or not (Fig. 2) (Tarpin-Bernard et al, 2001).

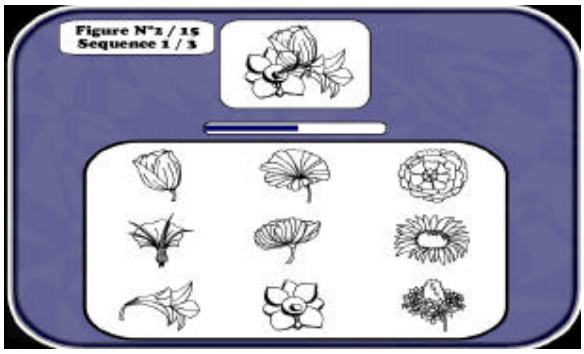


Figure 1: Significant context.

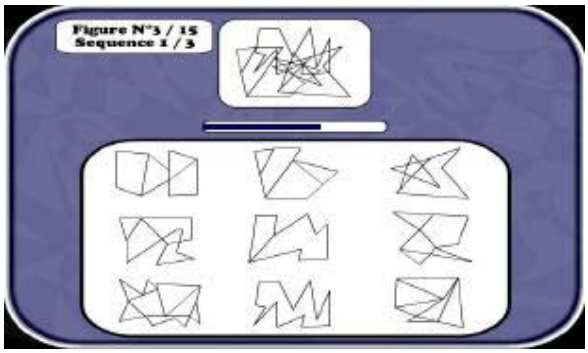


Figure 2: Non-significant context.

Impulsive reflective style

The impulsive subject tends to put forward the first idea that comes to him/her, whereas the reflective subject considers alternatives. This style is generally assessed by measuring differences in decision-making under conditions of uncertainty. Tasks used present several plausible choices, only one of which is correct:

- who responds quickly often errs;
- who pauses to reflect is more often correct.

Fig. 3 is an example of exercise where it is possible to identify one of the following categories:

- *fast-responding/high-error* (FH),
- *fast responding/low-error* (FL);
- *slow-responding/high-error* (SH);
- *slow-responding/low-error* (SL).

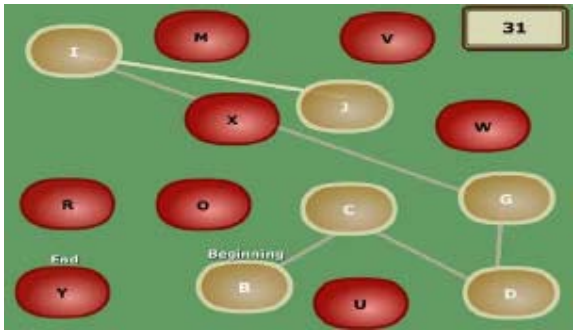


Figure 3 Exercise type for impulsive reflective identification.

Categorization style

Other individual differences consist in giving a number of objects and asking subjects to sort them into categories. Some individuals place objects into a wide number of small categories, so that each category contains only objects sharing a high number of similar features; other individuals place objects into a small number of wide categories which include items with few common features. Other individuals may group objects into different categorization where the criteria are not only the width:

- *analytic-descriptive style* induces to include in the same category items showing surface physical-perceptual similarities;
- *conceptual-inferential style* induces to define categories on the basis of similarities in objects' functions;
- *thematic-relational style* induces to include in the same category disparate objects which have in common only the fact that they occur in the same action or situation.

Figure (Fig.4) shows an exercise where subjects are invited to sort objects into categories suggested by the supervisor. Firstly, the users select the category then they sort object into this category.

Figure 4 Exercise type for the category identification

Analytic-global style

The last style we consider is the analytic-global style which refer to either considering the details of a situation or the whole picture (Euzeby, 1999). Analytic individuals have a focused attention, an interest in operations and procedures or the 'proper' ways of doing things and prefer step-by-step schemes; their thinking is controlled and consciously directed. Global persons tend toward scanning, leading to form overall impressions, including entry of feelings into decisions; their organizational schemes involve random or multiple accessibility of components and varied associations between them.

Some tests in cognitive styles analysis (Riding and Rayner, 1998) allow to measure the analytic dimension by presenting items each comprising a simple geometrical shape and a complex figure and by asking to indicate whether or not the simple shape is contained in the complex figure.

How do we Assess Cognitive styles?

Three main kinds of data can be employed to measure cognitive styles: behavioral, self-report, and physiological (Antonietti & Giorgetti, 1998)

Behavioral data can be obtained by recording the final result of a given task or the procedure followed in performing the task. The task may consist in filling out a paper-and-pencil test or a sorting test, in carrying out trials by means of an experimental apparatus, or in interacting with the computer like during exercise running (Tarpin-Bernard et al., 2001). For an example, to assess whether a person is a visualizer or a verbalizer, it is possible to present him/her with tasks which can be performed through both visual and verbal strategies and to record the extent to which each of the two kinds of procedures has been followed.

Self-reports require that people evaluate themselves by describing by introspective manner the way in which they performed tasks, by checking personal habits or preferences, or by endorsing statements about what they think of themselves. This may be done, for example, by asking subjects to keep a diary of what occurred to them during a period of their life, by interviewing them, or by adopting questionnaires.

The following example is given by Antonietti (Antonietti & Giorgetti, 1998):

In order to understand how much an individual tends to visualize, he/she can be requested to keep a record of the times in which he/she has experienced imagery during the day. Information of this kind may be derived also through questionnaires in which people are asked to rate how frequently they create and process various kinds of mental images. These instruments incite subjects to consider their habitual modes of thinking as they emerge in the complete range of mental activities and to assess the occurrence of visual images in different tasks, domains, contexts, and so on.

Finally, some physiological measures can be interpreted as indices of particular cognitive preferences in processing stimuli. Indeed, Physiological measures observations have indicated that when someone is asked a question requiring a little thought the eyes make an initial movement to the left or right. Since it was argued that the right cerebral hemisphere is associated with the processing of visual information and that the spontaneous lateral eye movements are under the control of the counter-lateral hemisphere, it was claimed that the presentation of a visual-spatial question produces the activation of the right hemisphere and, consequently, left lateral eye movements. However, verbalizers should turn their eyes consistently to the right and visualizers to the left, whatever the kind of question. Thus, it has been suggested to use lateral eye movements as a criterion to assess the preference for either a visual or a verbal processing.

Relationship Between Cognitive Styles and Interaction Styles

The most important components of HappyNeuron's technology are structured as follows (Fig.)5 :

First stage: User's profile generation process (Tarpin-Bernard et al, 2001)

- Questionnaire;
- Exercises;
- Supervision process;
- Cognitive styles
- User profile;

Second stage: Adaptation process

- Compatibility matrix;
- User profile;
- XML (eXtensible Markup Language) documents (multimodal documents);
  - Text;
  - Image or graphics;
  - Sound;
  - Video.
- Stylesheet;

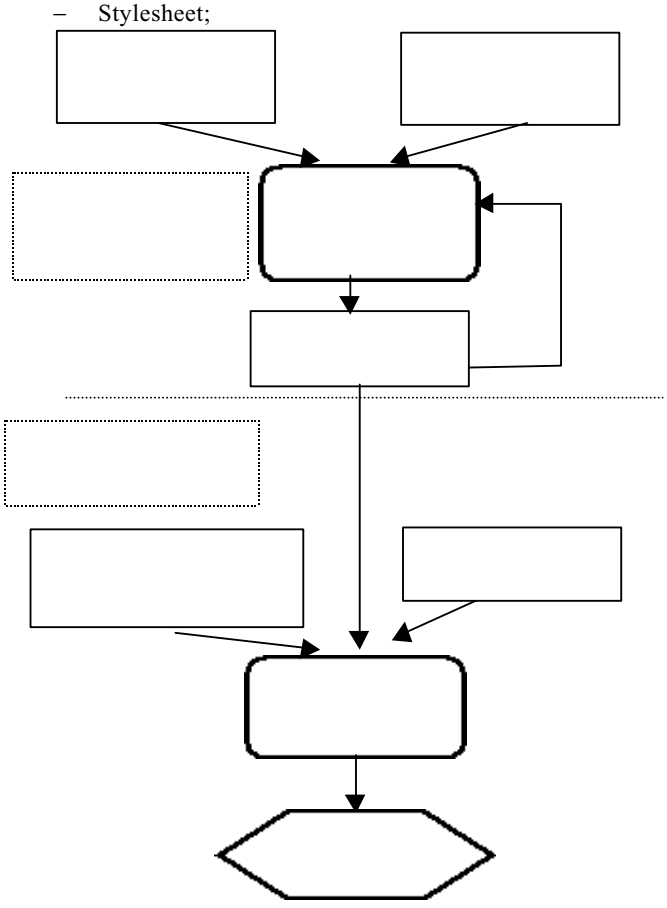


Figure 5: User profile generation and adaptation process

User's profile generation process

The user's profile generation process consists on the one hand in executing interactive exercises then constructing the user cognitive profile. On the other hand, it deals with other user's behavior such as the time to run an exercise, performance variations for the same exercise type, etc. This indicators are adjusted with each other and constitute the final user profile (Fig. 6).

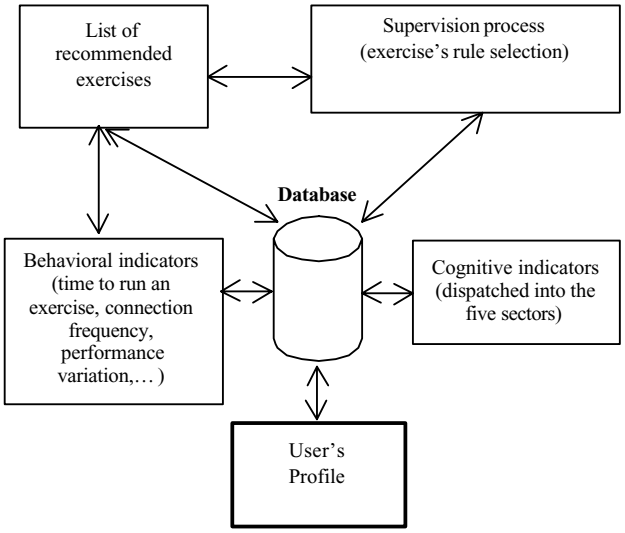


Figure 6:The user's profile generation process

During each training session, the supervisor suggests a set of three or four exercises. The user can also select other exercises in the complete list. At any moment, the system gives users feedback about their progression and enables them to:

- check their performance by consulting *the profile performance* page,
- have a summary of the exercises they have already done,
- browse some documents (news, forums,…).

Thus, the main components of the output profile are:

1- Cognitive indicators dispatched into 5 sectors: *memory, attention, executive functions, language and visual and spatial capacities* [Tarpin-Bernard et al, 2001]. In total, 25 indicators have been determined, we can mention several of them as an illustration: cultural memory, old personal memory, recent memory (verbal, visual or musical), working memory and short term memory with the tree modalities, lexical spelling, categorization, comprehension, arithmetic, planning, reasoning, mental imagery, form recognition, etc ;

2- Behavioral indicators (time during exercise running, connection frequency, etc.);

3- Indicators revealing some characteristics of styles such as field dependent or independent.

These indicators are affected with some weightings and contribute in the adaptation process. For example to determine whether the subject is field dependent or independent, we use the indicator which measures the difference in performance after running the same exercise with two different images (the first with significant context and the second with non-significant context) then we adjust it with other cognitive indicators such as recent memory (verbal, visual or musical) and comprehension.

Adaptive Process

As described in figure 5, the training process yields a user profile which constitutes the input data for the adaptation process. Indeed, this profile enables the selection of the outcomes style sheets. The multimodal document is defined into an XML document.

Prior to any process, each style sheet contains the layout of a complete page to be presented. For a specific subject, the final layout of this page is brought through the adaptive process. This page is the most compatible one to the user profile.

To illustrate the adaptive process, we give hereafter an example of a page to be presented dealing with the following subject : "The main parts of a language : the vocabulary and the syntax". In the XML document, the page falls under 2 elements. Each one could be presented according to different modalities. (Fig.7).

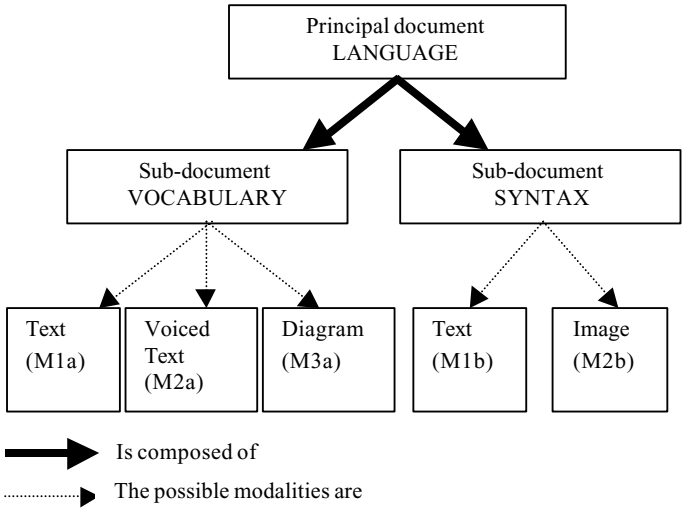


Figure 7: XML document structure

The problem is to find the "best" combination of modalities according to the willing of the designer and the abilities of the reader. According to the XML structure, the possible combinations are: (M1a, M1b), (M1a, M2b), (M2a, M3a, M1b), etc. Then, we can build

Metacat: A Self-Watching Cognitive Architecture for Analogy-Making

James B. Marshall (marshall@cs.pomona.edu)
Computer Science Program
Pomona College
610 N. College Ave.
Claremont, CA 91711 USA

Abstract

This paper describes Metacat, an extension of the Copycat analogy-making program. Metacat is able to monitor its own processing, allowing it to recognize, remember, and recall patterns that occur in its "train of thought" as it makes analogies. This gives the program a high degree of flexibility and self-control. The architecture of the program is described, along with a sample run illustrating the program's behavior.

Introduction

This paper describes Metacat, an extension of the Copycat analogy-making program originally developed by Hofstadter and Mitchell (Hofstadter, 1984; Mitchell, 1993). Copycat was developed as a model of the complex interplay between bottom-up and top-down perceptual processes in the mind, which together enable humans to perceive analogies between different situations in remarkably flexible ways. The program operates in an idealized microworld of analogy problems involving short strings of letters. Although the program understands only a limited set of concepts pertaining to its letter-string world, its "fluid" processing mechanisms give it considerable flexibility in recognizing and applying these concepts in many diverse situations.

The long-term goal of the Copycat line of research is to computationally model how high-level cognitive phenomena such as creativity, analogical perception, understanding, and self-awareness can arise out of a subcognitive substrate composed of a huge number of tiny, nondeterministic processes, each of which is far too small by itself to support such phenomena. Few people would suggest that individual neurons in the brain (or individual molecules) are "conscious" in anything like the normal sense in which humans experience consciousness. One is forced to accept the fact that self-awareness arises, somehow, out of nothing but billions of molecular chemical reactions and neuronal firings. How can individually meaningless physical events in a brain—even a huge number of them—ultimately give rise to meaningful awareness? Hofstadter has argued that two key ideas are of paramount importance (Hofstadter and FARG, 1995):

> What seems to make brains conscious is *the special way they are organized*—in particular, the higher-level structures and mechanisms that come into being. I see two dimensions as being critical: (1) the fact that brains possess *concepts*, allowing complex representational structures to be built that automatically come with associative links to all sorts of prior experiences, and (2) the fact that brains can *self-monitor*, allowing a complex internal self-model to arise, allowing the system an enormous degree of self-control and open-endedness.

Work on Copycat explored the first idea through the development of a computer model of analogy-making in which the program's representation of concepts is intimately intertwined with its mechanisms for perceiving similarity between different idealized situations. Recent work has focused on the second idea by incorporating *self-watching* into the model—namely, the ability of a system to perceive and to explicitly characterize its own perceptual processes. The objective of this work has been to develop mechanisms that allow the program to monitor its own actions and to *make explicit* the ideas that come into play during the course of solving analogy problems. This can be thought of as adding a higher "cognitive" layer on top of the program's "subcognitive" layer, enabling the program to watch and remember what happens at its subcognitive level as perceptual structures are built, reconfigured, and destroyed. Humans are capable of paying attention to patterns in their own thinking in a similar fashion (see, for example, Chi *et al.*, 1994).

Self-watching in Copycat and Metacat

The Copycat architecture has been discussed at length elsewhere (Mitchell, 1993; Hofstadter and FARG, 1995), so details will be omitted here. Briefly, the program consists of a long-term memory of concepts about the letter-string world, called the *Slipnet*, together with a short-term memory for perceptual structures, called the *Workspace*. In the Workspace, small nondeterministic agents called *codelets* examine the letters of an analogy problem (*"**abc** ⇒ **abd**; **mrrjjj** ⇒ **?**"*, for example), and build

up structures around the letters representing a particular interpretation of the problem. The program's high-level behavior emerges in a bottom-up manner from the collective actions of many codelets working in parallel, in much the same way that an ant colony's high-level behavior emerges from the individual behaviors of the underlying ants, without any central executive directing the course of events.

Guiding the search for a mutually-consistent set of structures are concepts in the Slipnet, which become activated to different degrees depending on the activity in the Workspace. This activation may spread to neighboring concepts, and strongly influences codelet decisions, resulting in top-down pressure that guides the program in its search for a good interpretation of a problem.

The overall degree of Workspace organization is measured by a number called the *temperature*. Temperature not only reflects the state of the Workspace, it also continuously regulates the amount of randomness used by codelets in making decisions. At high temperatures, few Workspace structures exist, so decisions are made in a highly random manner, since not much is yet known about the problem. However, as relationships among the letters are noticed and structures are built, the temperature falls, and Copycat begins to gain "confidence" in its understanding of the situation. At lower temperatures, decisions are still probabilistic, but are much less random, being strongly biased by the estimated promise of newly emerging structures, all of which compete for attention by codelets. At very low temperatures, codelets pay attention to only the most promising structures, and decisions become largely deterministic. Thus the type of strategy used by the program to explore its search space ranges along a broad continuum, from being very diffuse and highly parallel at high temperatures to being very serial and focused at low temperatures.

To summarize, Copycat's search proceeds via a large number of fine-grained stochastic decisions, which depend on the temperature. These decisions may cause new structures to be built or existing structures to be destroyed, which changes the temperature and subsequently alters the course of structure building, forming a kind of feedback loop. Temperature thus serves as a very crude mechanism for self-watching in Copycat, since it allows the program to regulate its own behavior to a limited extent. That is, by tying the stochastic activity of codelets to the temperature, the program becomes sensitive to the consequences of its own actions, since the temperature reflects the result of these actions, albeit in a very coarse way (*i.e.*, in the form of a single number).

This type of rudimentary self-watching, however, is quite primitive. Copycat can characterize patterns within its perceptual input (the letter strings), but is completely oblivious to patterns that arise in its *processing* of that input. For example, when solving the problem *"**abc**⇒ **abd**; **xyz**⇒ **?**"*, Copycat usually attempts to take the successor of ***z***, which is impossible in the program's microworld. It "hits a snag", and is forced to try something else. However, it often just tries the same thing again, over and over, sometimes as many as thirty or forty times before stumbling by chance on an alternative approach (such as the answer ***xyd***). Unlike humans, the program is unable to recognize when it has fallen into a repetitive pattern of behavior. It has no memory of its actions over time, and thus cannot recognize when it has encountered the same situation in the past. As a result, Copycat lacks insight into how it arrives at its answers, and consequently cannot explain what makes one answer better or worse than another.

In contrast, Metacat is able to create much richer representations of the analogies it makes, enabling it to compare and contrast answers in an insightful way. This has involved incorporating an episodic memory into the original Copycat architecture, along with new mechanisms that allow the program to monitor itself, so that it can recognize, remember, and recall patterns that occur in its "train of thought" as it makes analogies.

To do this, Metacat creates an explicit sequential record of the most important processing events that occur during a run. This temporal record is examined by codelets for patterns—in the same way that Copycat's codelets examine letter-strings for patterns—and serves as the basis for constructing an abstract description of an answer in terms of the key concepts and events that led to its discovery. Furthermore, by monitoring its own processing in this way, Metacat can recognize when it has become "stuck in a rut", enabling the program to break out of the rut by focusing on ideas other than the ones that seem to be leading it nowhere.

The Architecture of Metacat

Metacat's architecture includes all of Copycat's architectural components, such as the Workspace and the Slipnet, as well as three new components: the *Episodic Memory*, the *Themespace*, and the *Temporal Trace*. When the program discovers a new answer, it pauses to display the answer along with the Workspace structures that gave rise to it. These structures represent a way of interpreting the problem that yields the answer just found. All of this information is then packaged together into an *answer description* and stored in the Episodic Memory, after which the program continues searching for alternative answers to the problem, instead of simply quitting. Gradually, over time, a series of answer descriptions accumulates in memory, each one containing much more information than just the answer string itself.

The most important structures stored in answer

descriptions are called *themes*, which represent the essential ideas underlying an answer. The collection of themes associated with an answer serves as the basis for comparing it to other answers stored in memory. Furthermore, Metacat may be reminded of other answers it has encountered in the past if the themes associated with a newly discovered answer, acting as a memory retrieval cue, are similar enough to those of a previously stored answer description. Thus an answer's themes act as an index under which it is stored and retrieved from memory.

Themes get created in Metacat's Themespace as codelets build structures around the letter-strings, and are composed of Slipnet concepts. For example, in the problem *"**abc**⇒**abd**; **xyz**⇒**?**"*, an *Alphabetic-Position: opposite* theme representing the idea of alphabetic-position symmetry between the letters ***a*** and ***z*** might get built if the program perceives these letters as playing analogous roles in their respective strings (an interpretation that may lead to the "mirror image" answer ***wyz***).

In some ways, themes are like ordinary Workspace structures. They are not initially present in the Themespace; rather, they arise during the course of a run as the result of codelet activity occurring in the Workspace. In other ways, however, themes behave like Slipnet concepts. They can take on different levels of activation, reflecting the extent to which the ideas they represent are supported by structures in the Workspace. A theme's activation level decays over time, and is influenced by the activation levels of other themes. Like Slipnet concepts, themes can, under certain conditions, exert strong top-down pressure on perceptual activity in the Workspace. In fact, themes can assume both positive and negative levels of activation, ranging from −100 to +100. A positively-activated theme exerts "positive thematic pressure", encouraging the creation of Workspace structures that support the idea represented by the theme. A negatively-activated theme, on the other hand, exerts "negative thematic pressure", which *discourages* the creation of structures related to the theme, promoting instead the creation of alternative structures.

The Temporal Trace serves as the focal point for self-watching in Metacat. Like the Themespace, the Trace accumulates information over the course of a single run, and can be viewed as an extension of the Workspace. The Trace stores an explicit temporal record of the most important processing events that occur while the program works on an analogy problem. Examples of such events include recognizing some key idea pertaining to the problem (by noticing the strong activation of a theme or concept, for instance), hitting a snag, or discovering a new answer. Once processing events have been explicitly represented in the Trace as "reified" structures in their own right, they are subject to examination by codelets as well. Metacat thus uses a single set of mechanisms for perceiving patterns in its perceptual input and in its own processing of that input. When a new answer is found, an answer description can be formed by examining the temporal record in the Trace to see which events contributed to the answer's discovery.

This approach is similar in flavor to work on derivational analogy, in which the trace of a problem-solving session is stored in memory for future reference, together with a series of annotations describing the conditions under which each step in the solution was taken (Carbonell, 1986; Veloso, 1993). In Metacat's case, however, the information in the Trace is used as the basis for constructing an abstract description of the answer found, rather than being permanently stored itself.

One way to appreciate the abstract, chunked nature of the information in the Trace is to consider the number of "steps" that occur during a typical run of Metacat. At a very fine-grained level of description, where each step corresponds to an action performed by a single codelet, a run consists of many hundreds or thousands of steps. At this level of description, no two runs are ever *exactly* the same, even if they involve the same letter-strings (unless, of course, both runs start with the same random number seed). On the other hand, at the level of description of the Trace, a typical run consists of a few *dozen* steps. At this level of granularity, each step corresponds to a single event represented in the Trace—each of which arises from the actions of many codelets.

For example, Figure 1 shows the contents of the Trace after a run on the problem *"**abc**⇒**abd**; **xyz**⇒**?**"*, in which the program, after trying unsuccessfully a couple of times to take the successor of ***z***, answers ***xyd***. The events that occur during the run appear left to right in chronological order. Although this run involves a total of 1,558 codelets, the high-level picture shown in the Trace consists of just twelve events, which represent the "major milestones" encountered along the way in the program's search for an answer. Such events include the activation of concepts in the Slipnet, perceiving entire strings as single, chunked wholes, creating new rules for describing string changes, hitting a snag, and discovering a new answer.

For instance, as can be seen in the figure, the Slipnet concept *identity* gets activated early on in this particular run (due to the program perceiving the ***a***'s and ***b***'s in ***abc*** and ***abd*** as corresponding). This is followed by the perception of ***abc*** and ***xyz*** as *predecessor* groups going in the same direction (to the left). The next event records the creation of the rule *Change letter-category of rightmost letter to successor* for describing ***abc*** ⇒ ***abd***, which leads inevitably to a snag. In the aftermath of the snag, another rule is created (*Change letter-category of rightmost letter to 'd'*), and ***abc*** and ***xyz*** are reperceived as *successor* groups (again going in the same direction—only

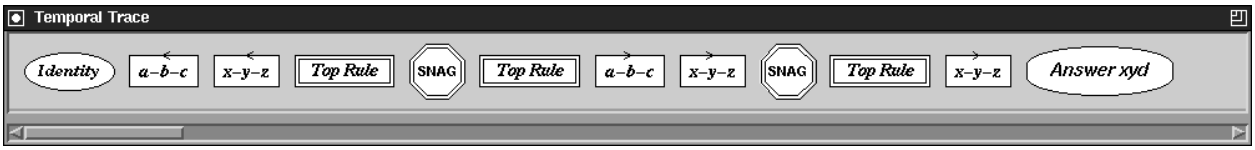


Figure 1: The temporal record of a run on the problem *"**abc*⇒ *abd; xyz*⇒ *?"*.

this time to the right). However, the program again attempts to use the first rule, resulting in another snag. Finally, after creating yet another rule and again perceiving ***xyz*** as a successor group, the program finds the answer ***xyd***.

Pattern-clamping and Self-control

The Trace allows Metacat to monitor the subcognitive processing activity in the Workspace at a very abstract and highly-chunked level of description, enabling the program to "see" what it is doing during a run. Equally important, however, is the program's ability to *respond* to what it sees by clamping particular themes and concepts at high activation, resulting in strong top-down pressure on processing. Various types of *patterns* consisting of sets of themes, concepts, or codelet types can be clamped by the program in response to different situations that arise. Clamping a pattern alters the probabilities that certain types of codelets will run, or that certain types of Workspace structures will get built, effectively steering the behavior of the program in particular directions. This may lead the program to revise its interpretation of a problem, by catalyzing the reorganization of structures in the Workspace in accordance with the ideas represented by the pattern.

Metacat's ability to revise its perception of a situation in response to events in the Trace affords the program a very powerful degree of self-control. Patterns—especially patterns of themes—act as a "medium" through which the program is able to wield control over its own behavior. For example, in the problem *"**abc*⇒ *abd; xyz*⇒ *?"*, the program usually perceives ***abc*** and ***xyz*** as going in the same direction at first, which leads to a snag (as in the run shown in Figure 1). This interpretation of the problem, based on the idea that letters having identical positions in their respective strings correspond to one another (***a*** to ***x***, ***b*** to ***y***, ***c*** to ***z***), is characterized by a *String-Position: identity* theme. When an event is recorded in the Trace, the themes most active at the time of the event are also noted along with it. These themes serve as the event's *thematic characterization*. In the case of a snag event, the thematic characterization represents an interpretation of the problem that has just led to failure. If Metacat continues to hit the same snag several times in succession, a series of snag events will accumulate in the Trace, all with very similar thematic characterizations. This similarity may be noticed by codelets (the probability becoming higher as more snags accumulate), causing them to take action by clamping the "offending" themes (such as *String-Position: identity*) with strong negative activation. This encourages the program to explore alternative interpretations of the problem by steering it away from the ideas causing the snag, which may subsequently lead it to the discovery of other answers, such as ***wyz***. In this way, Metacat can recognize its own repetitive behavior and respond accordingly.

Two types of codelets are responsible for examining and responding to events unfolding in the Trace. The first type, called a *Progress-watcher*, is responsible for deciding whether or not to unclamp a clamped pattern. If a *Progress-watcher* codelet runs while a pattern is clamped, it examines the most recent event in the Trace to determine how much time has elapsed since the event occurred. Generally speaking, the purpose of clamping a pattern is to precipitate a series of events that reorganize the perceptual configuration of the Workspace in some way. It is therefore better to wait until the structure-building activity occurring in the wake of a clamp has settled down before concluding that the clamp has "run its course". Accordingly, if the amount of time since the most recent event in the Trace is less than some minimal settling period, then the codelet simply fizzles, leaving the clamped pattern still in effect. On the other hand, if enough time has passed without any new important events having transpired, the codelet unclamps the pattern and then determines the amount of progress that was made since the clamp occurred. Depending on the amount of progress achieved, the codelet may decide to post a follow-up codelet in order to see whether a new answer can be made based on the newly-created structures.

The criteria for judging the success of a clamp can vary. Sometimes, the purpose of clamping a pattern is to promote the creation of *specific* types of Workspace structures. Other times, the purpose is to encourage the creation of structures of *any* type, so long as they are compatible with the clamped pattern. The progress achieved by a clamp can be measured by observing the number of structures that get built in the immediate aftermath of the clamp, and the extent to which they are compatible with the pattern.

If no patterns are clamped when a *Progress-watcher* codelet runs, then instead of checking on the progression of events in the Trace, the codelet

checks on the current rate of structure-building activity in the Workspace. This activity is measured by a single number that serves as a quick estimate of the "freshness" of the current Workspace structure configuration. More precisely, it is an inverse function of the average age of the most recently created structures. Thus the activity level tends to remain high as long as new structures are being built, but eventually drops to zero in the absence of new structures.

If the activity level is zero, indicating that nothing much is happening in the Workspace, then Metacat may have arrived at an impasse in its search for answers to the current problem. This is not quite as bad as hitting a snag, but it still ought to prod the program into trying something different. However, in the case of an impasse, there is usually no clear set of "offending" structures or themes on which to pin the blame, unlike in the case of a snag. Indeed, the impasse may well arise from a *lack* of appropriate structures, rather than from the existence of the "wrong" structures.

Therefore, in the absence of Workspace activity, *Progress-watcher* codelets check to see whether particular types of new structures are needed. For example, a codelet may examine the quality of the rules that have been built so far. If no good rules yet exist, the codelet may try to encourage the creation of better rules by clamping a pattern that strongly increases the probability that rule-seeking codelets will run, while simultaneously inhibiting other types of codelets. Eventually, other *Progress-watcher* codelets will turn off the clamp once enough time has passed without any more events being added to the Trace. Since this type of clamp is only concerned with the creation of new rules, the amount of progress achieved is judged solely on the basis of the quality of the rules that get created in the clamp's wake.

The second type of codelet that "watches the action" from the high-level vantage point of the Trace is called a *Jootser* (short for "*j*umping *o*ut *o*f *t*he *s*ystem"). These codelets are responsible for noticing repetitive behavior that the program has fallen into. An example of such behavior arising from a snag was sketched above. However, *Jootser* codelets are sensitive to other kinds of situations as well. For example, it is possible for Metacat to become "fixated" on some idea, such that it ends up clamping the same pattern over and over again, without making any significant progress. In this case, too, *Jootser* codelets may notice the series of recurring events in the Trace and take action.

For instance, if an analogy problem happens to involve a string that changes in some difficult-to-describe way, the program may end up repeatedly clamping patterns in an attempt to spur the creation of better rules for describing the change. Repetitive clamping behavior can even arise from unsuccessful attempts to break out of a cycle of snags. That is, clamping a pattern in response to a recurring snag may prove to be ineffective, leading only to further snags and more pattern-clamping, rather than to a new interpretation of the problem. Faced with several similar clamp events in the Trace, a *Jootser* codelet decides probabilistically whether to "joots" based on the number of clamps and the average amount of progress achieved by each. The more clamp events there are, the more likely jootsing is to occur, especially if the amount of progress is low, unless recent clamps appear to be making more headway than earlier ones. Unlike jootsing from snags, however, jootsing from a series of recurring clamp events does not involve the clamping of any new patterns in response. Instead, Metacat simply "gives up" in a graceful manner and stops.

In a sense, Metacat's ability to respond to a recurring snag by focusing on alternative ideas can be thought of as "first-order" jootsing. In contrast, the program's ability to eventually give up when it recognizes that its repeated attempts to circumvent a snag are leading nowhere can be thought of as "higher-order" or "meta-level" jootsing (*i.e.*, jootsing from repeated unsuccessful jootsing). The important point is that the same general mechanisms are responsible for first-order and meta-level jootsing in Metacat—namely, *Jootser* codelets and the explicit representation of processing events in the Trace.

An Example of Jootsing

The following example illustrates the idea of jootsing. In this run, Metacat is given the problem *"**eqe** ⇒ **qeq**; **abbbc** ⇒ **?**"*. The program builds up an interpretation of the string ***abbbc*** as a successor group composed of the letter ***a***, the group ***bbb***, and the letter ***c***. The two rules shown below are also created to describe the ***eqe*** ⇒ ***qeq*** change:

- *Swap letter-categories of all objects in string*

- *Change letter-category of leftmost letter to 'q'*
  Change letter-category of middle letter to 'e'
  Change letter-category of rightmost letter to 'q'

Around time step 1100, the program attempts to apply the first rule to ***abbbc***, which results in a snag, since the idea of a three-way swap involving the letters ***a***, ***b***, and ***c*** makes no sense (see Figure 2). Of course, if it had chosen to use the second rule instead of the first, then it would have found the answer ***qeeeq***, but it prefers the first rule, since this rule is more abstract.

Over the next 3000 time steps, the program tries again and again to swap the letters of ***abbbc***, often breaking various structures in the process, but always rebuilding them in the same way as before. Eventually, at time step 4280, a *Jootser*

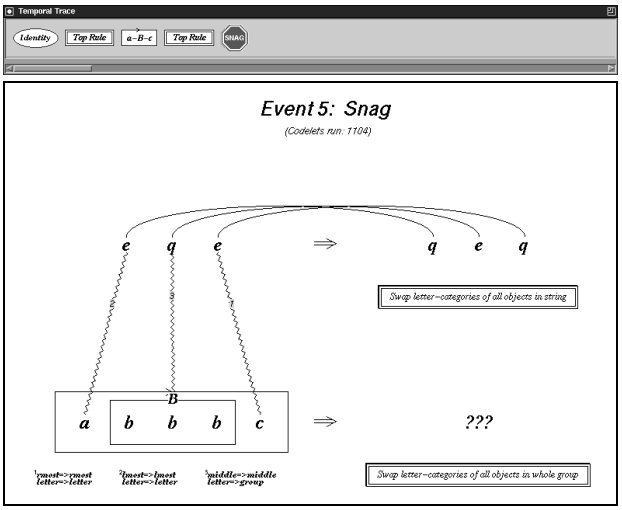


Figure 2: Attempting to swap the letters of ***abbbc***

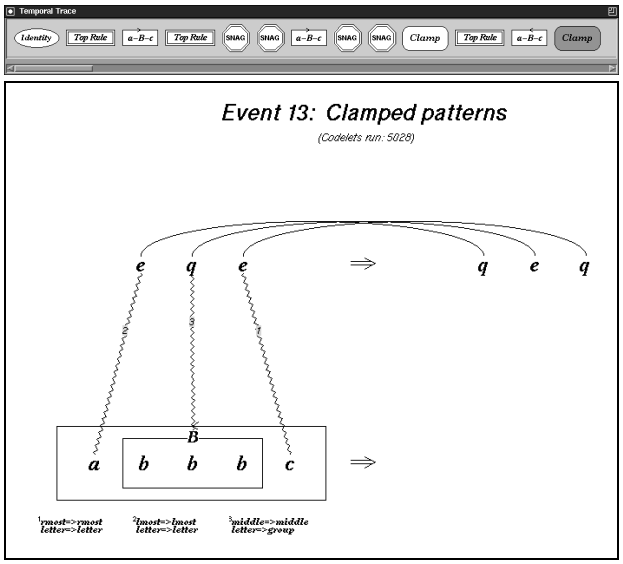


Figure 3: Clamping patterns in response to snags

codelet notices the pattern of recurring snag events in the Trace, all of which involve the themes *String-Position: identity*, *Object-Type: identity*, and *Object-Type: different*. These themes arise from the program's interpretation of the letters ***e***, ***q***, and ***e*** in ***eqe*** as corresponding, respectively, to the letter ***a***, the group ***bbb***, and the letter ***c*** in ***abbbc***. The *Object-Type: identity* theme is based on the ***e***−***a*** and ***e***−***c*** correspondences, while the *Object-Type: different* theme results from the correspondence between ***q*** and ***bbb***, since one is a letter and the other a group.

In hopes of finding a way around the recurring snag, the codelet decides to negatively clamp the *Object-Type: identity* theme. In the wake of the clamp, ***abbbc*** is reinterpreted as a predecessor group going to the left, and a new rule is created to describe ***eqe*** ⇒ ***qeq***, but these new structures do not really change the basic situation. Soon afterwards, another *Jootser* codelet tries again, this time clamping *both* *Object-Type* themes, which essentially "paralyzes" the program for the duration of the clamp, since no structures can be built that are compatible with both of these themes simultaneously. Figure 3 shows the state of the Workspace and Trace at the time of the second clamp.

A few hundred codelets later, the program hits the snag again. This is followed shortly thereafter by another clamp. This clamp, like the one before it, achieves no new progress. After hitting the snag yet again, the program finally decides to give up. More precisely, at time step 5933, a *Jootser* codelet notices the three clamp events in the Trace, all of which involve overlapping thematic characterizations. Moreover, neither of the two most recent clamps have resulted in any discernible progress, which further increases the probability of jootsing. Consequently, the program prints the message "this is getting boring, I can't think of anything else to try" and then ends the run.

As this example shows, Metacat is able to realize when it is "stumped", instead of just cycling endlessly. The program's ability to monitor its own processing at an abstract level of description affords it a great deal of flexibility and self-control, and, it is to be hoped, represents a step toward the goal of understanding the cognitive mechanisms underlying human self-awareness.

References

Carbonell, J. (1986). Derivational analogy: a theory of reconstructive problem solving and expertise acquisition. In R. Michalski, J. Carbonell, & T. Mitchell (Eds.), *Machine learning, volume 2*. San Francisco: Morgan Kaufmann.

Chi, M., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18:439–477.

Hofstadter, D. R. (1984). *The Copycat project: an experiment in nondeterminism and creative analogies*. AI Memo 755, MIT Artificial Intelligence Laboratory.

Hofstadter, D. R. & the Fluid Analogies Research Group (1995). *Fluid concepts and creative analogies*. New York: Basic Books.

Marshall, J. (1999). *Metacat: a self-watching cognitive architecture for analogy-making and high-level perception*. Doctoral dissertation, Department of Computer Science, Indiana University, Bloomington. http://www.cs.pomona.edu/marshall/metacat.pdf

Mitchell, M. (1993). *Analogy-making as perception*. Cambridge, MA: MIT Press/Bradford Books.

Veloso, M. (1994). *Planning and learning by analogical reasoning*. Berlin: Springer-Verlag.

Where do syllables come from?

Evelyn Martens and Walter Daelemans and Steven Gillis and Helena Taelman
Universitaire Instelling Antwerpen
Universiteitsplein 1
2610 Wilrijk
Belgium

Abstract

Young children are able to segment words into syllables, even though there are no perceptual or acoustic cues that indicate syllable boundaries in the primary linguistic data. We show that information about word boundaries can be used to predict syllable boundaries by replicating the results of experiments done by Gillis and De Schutter (1996) with children who syllabified Dutch disyllabic monomorphemes with a single intervocalic consonant. Word boundary probabilities were statistically computed in child language corpora and used to predict syllable boundaries with a simple statistical model. The children's syllabification behavior could be simulated using word-boundary probabilities estimated from child language corpora. Similar results were obtained for three different corpora. In our simulations, we also investigate the question whether children acquire their knowledge of word boundaries from words from the input, from the intake, or from their own output.

Introduction

The syllable is an important construct in phonological descriptions of languages (Van der Hulst & Ritter, 1999) as well as in models of language acquisition (Jusczyk, 1997) and language processing (Levelt, 1989). In most contemporary phonological theories the syllable plays an important role at the segmental level (e.g., in consonant harmony) as well as at the supra-segmental level (e.g., in stress assignment). Across languages syllables adhere to a number of universal principles (Venneman, 1988) and Clements (1990) proposes a universally valid algorithm for syllabifying words. One of its operating principles is 'sonority sequencing': a syllable has rising sonority from the left edge to the vocalic nucleus and falling sonority from the vowel to the right edge. Irrespective of the theoretical framework in which the universals of syllabification are cast, it is accepted that the language universals, such as those incorporated in Clements' algorithm, can be overruled by language-specific constraints. For instance, at the end of a syllable long vowels are universally accepted, but languages differ as to whether there can be a short vowel at the end of a syllable (Kager, 1989).

In sharp contrast to the relatively clear phonological picture stands the phonetic reality: what are the acoustic correlates of the syllable in the speech stream? For instance, acoustic correlates of the 'sonority sequencing principle' are very difficult to determine, which led phoneticians to define the syllable from a phonetic point of view as that entity of which the word *syllable* has three. The syllabic nucleus (the vowel) is fairly easy to detect, but the syllable boundaries are not straightforward. For instance, the /I/ in *bitter* is the nucleus of the first syllable, but where is the boundary of that syllable: immediately after the vowel /bI.t@r/ or after the first consonant /bIt.@r/ or in the middle of the first consonant /bIt.t@r/, a case of ambisyllabicity? This brings us to the core issue addressed in the present paper: if from a structural perspective syllables are easy to describe, but if it is very difficult to depict the acoustic correlates of the syllable and its boundaries, it is an outstanding question how children arrive at detecting syllables and their boundaries.

Nevertheless, in early speech perception (Jusczyk, 1997) as well as in speech production (Wijnen, 1988) children appear to use syllables as organizing entities. The question is: how does a child acquire the knowledge of the structure of syllables?

In the acquisition literature there are basically two approaches: in a nativist approach, the universals of syllable structure are thought to be innately given: they are described as inborn parameters (Fikkert, 1998), or as inborn constraints (Kager, 1999; Levelt, Schiller, & Levelt, 2000). Acquiring the structure of syllables requires a child to figure out the language-specific parameter setting or the language-specific constraint ranking. Thus, the broad outlines are genetically given, so that only on the basis of the ambient language the child has to determine where precisely her mother tongue fits into these outlines. Appealing as this may sound, it is unclear on what basis parameters are set or constraints are ranked. The cues for parameter setting or constraint ranking can only be found in the input. However, the acoustic correlates of the syllable are not clear in the input (see second paragraph).

The alternative approach is that children do not start from a preset body of knowledge, but instead

use the information available in the input to arrive at linguistically relevant knowledge. For instance, Brent and Cartwright (1996) found that word boundaries can be learned on the basis of utterance boundaries. In a similar vein we want to investigate if syllable boundaries can be learned on the basis of word boundaries. Word boundaries are clear and usable cues, because words often occur in isolation in child-directed speech (van de Weijer, 1998; Brent & Siskind, 2001). Thus, the hypothesis tested in this paper is that syllable boundaries are learned on the basis of word boundaries.

In this paper we will test this hypothesis in a simulation experiment. The results of the simulations will be evaluated in the light of children's actual syllabification behavior. Gillis and De Schutter (1996) tested 5- and 6-year-old native Dutch-speaking children in a syllabification task: they syllabified disyllabic Dutch monomorphemes with a single intervocalic consonant, such as /Ap@l/ 'apple'. The children segmented the test words orally and of the possible syllabifications (V.CV, e.g. /A.p@l/, VC.V, e.g. /Ap.@l/, and VC1.C1V, e.g. /Ap.p@l/) the preferred syllabification pattern was V.CV (81.6%), i.e. before the intervocalic consonant. The next most frequent syllabification was the ambisyllabic pattern VC1.C1V (17.8%), and the children almost never (0.4%) put a syllable boundary after the intervocalic consonant (VC.V). Furthermore, children's syllabification of the intervocalic appeared to depend on the length of the preceding vowel, the stress pattern of the word, and the quality of the intervocalic consonant. These results will be taken as the background against which the results of the simulations will be evaluated.

Naive Bayesian learning of syllabification

Whether it is possible to learn syllable boundaries from information about word boundaries will be investigated with a naive Bayesian learning technique. A simple statistical model uses estimated word boundary probabilities of segments to predict syllable boundaries. This model takes into account the probability that a phoneme occurs at the end of a word and that the following phoneme occurs at the start of a word, and combines both features in a multiplicative way. Such a model does not take into account interactions between the features, hence Naive Bayesian learning, a well-known supervised learning approach (Mitchell, 1997). However, we don't use a normal supervised learning set-up in which training and testing is on the same data. In our case, training is on word boundary information and extrapolation is to syllable boundary decisions.

In the training data, the word-initial boundary probability and the word-final boundary probability of every phoneme are computed. This is done by counting the number of times the phoneme is at the end of a word (e), the number of times the phoneme is at the beginning of a word (b), and the total number of times that phoneme occurs (t). For every phoneme the word-initial boundary probability and the word-final boundary probability are then computed in the following way:

p(beginning(phoneme)) = b/t
p(end(phoneme)) = e/t

The model's task is to predict the syllable boundary in disyllabic monomorphemic Dutch words with one intervocalic consonant. To compute the probability of a syllable boundary between two phonemes of a test word, the word-final boundary probability of the first phoneme, and the word-initial boundary probability of the second phoneme are multiplied. This is done for the two possible syllabifications of the test word, considering that every syllable must contain a vowel.

E.g. *appel*, /Ap@l/, 'apple'
p(end(A))*p(beginning(p)) = p(V.CV)
p(end(p))*p(beginning(@)) = p(VC.V)

For all the test words the probabilities of V.CV and of VC.V are computed. I.e. the probability that the syllable boundary falls either before or after the intervocalic consonant. Ambisyllabicity occurs if the difference between those two numbers does not exceed a maximum limit. If it does, the pattern with the highest probability is chosen. This method forces the model to syllabify and to choose one of three syllabification patterns. No syllabification occurs, though, if the probability for both V.CV and VC.V is zero. This way, a fourth category of "no syllabification" is created, to make sure these cases are not counted as ambisyllabicity.

For n = threshold:

if	p(V.CV) = 0 and p(VC.V) = 0
	→ no syllabification
else if	|p(V.CV) - p(VC.V)| < n
	→ VC1.C1V (ambisyllabic)
else	max(p(V.CV), p(VC.V))

As the probabilistic model is trained on a two-way classification problem (either there is a word boundary or not), and the target classification problem is four-way (ambisyllabic, before or after the intervocalic consonant, no syllabification), we fixed the model on the proportion of ambisyllabicity found in the empirical data by setting the n threshold. This threshold value is determined by the amount of ambisyllabicity. The percentage of ambisyllabic syllabification is put as close as possible to 17.8%, which is the percentage of ambisyllabicity found in the experiments by Gillis and De Schutter (1996).

The fixing of a threshold parameter on the test data to be explained is an unfortunate consequence of the fact that the training data (word segmentation information) does not contain a similar concept to

ambisyllabicity at the syllable level. Nevertheless, the threshold value seems to be rather robust over different training data sets, and could be learned with simple hill-climbing type of algorithms (there is a smooth gradient).

Research questions

Considering the different factors that might play a role in syllabification, a number of research questions were formulated.

1. What is the nature of the child's primary linguistic data? To acquire knowledge of language, children may analyze all the language that they hear or that is addressed to them (i.e., child-directed speech). Alternatively, it may well be that it is not the *input*, but the *intake* (i.e., what the child picks

   up from the input) (Wijnen, 2000) that is crucial for analysis. Alternatively, proponents of the output-as-input hypothesis (Elbers, 2000) argue that the input for children's linguistic analysis is primarily their own production, their own *output*.

2. What type of words is children's language analysis based on? Judging from the absence of function words in children's early productive vocabulary, it may well be that only content words are vital. And since syllables play a role in children's earliest word productions (Fikkert, 1998), it is important to investigate if syllabification can be acquired solely on the basis of content words as opposed to function words.

   Judging from the predominance of monosyllabic words in children's early production (or even the fact that all children initially exclusively produce monosyllables (Fikkert, 1998)) also the opposition between monosyllables and polysyllables will be investigated.

3. What is the influence of frequency on the acquisition of syllabification? Frequent words in the input are more salient for children (Jusczyk, 1997). However, Schreuder and Baayen (1997) found that the word frequency effect is composite in nature in the sense that it has both a token and a type component.

4. What is the optimal representation? Are words best represented as phonemes, or as phoneme categories? And is stress part of the representation?

   Phoneme categories express distinctive articulatory and acoustic features of phonemes, which is the reason why they differ in their scale of sonority. Sonority is regarded as important in syllabification, e.g. the universal Sonority Sequencing Principle describes syllables in terms of rising and falling sonority (Selkirk, 1984; Clements, 1990).

   Stress as well has been suggested as a determining factor in syllabification. There is a significant interaction between stress and length of the first vowel (Gillis & De Schutter, 1996), and there is less syllabification after the vowel if the first syllable is stressed than if it is unstressed (Wijnen, 1988).

In the following sections, we will report on experiments in which these dimensions are systematically encoded in the training data. The degree to which the resulting syllabification behavior of our statistical model matches the empirical data may have heuristic value to answer the question which dimensions of language data and representation are relevant in explaining this aspect of language acquisition.

Experiments

The input for the learner consisted of data taken from three Dutch child language corpora, all available through CHILDES (MacWhinney, 2000). The research questions were translated into different selections of input material and different types of input representations that were systematically varied in order to figure out their influence on the learnability of the task. Experiments were performed

1. using as training material the input to the child, the child's intake, and the child's output (the concept of intake was operationalized by using the actual adult model form of a child production, which makes intake a subset of the input);

2. using as training material different types of words: all words vs. content words, monosyllabic vs. polysyllabic words;

3. with information about word frequencies: word types vs. word tokens, as calculated from the corpora;

4. in which the representation of the input was varied: raw segmental material (phonemes) vs. segment categories (stops, fricatives, nasals, liquids, glides, and vowels) both with and without primary stress marking.

Combining all these factors in three child language corpora leads to a total of 136 experiments. In each case, the test material consisted of the words that were used in the experiment with children (Gillis & De Schutter, 1996) (see introduction). The artificial learner is set to the same task as the children: predicting the syllable boundary in Dutch disyllabic monomorphemes with a single intervocalic consonant. Hence, the learner has to decide whether for a given word (e.g. *appel*, /Ap@l/, 'apple') the string VCV should be syllabified as V.CV (/A.p@l/), VC1.C1V (/Ap.p@l/) or VC.V (/Ap.@l/).

For the different datasets, word boundary probabilities are computed with a naive Bayesian learning technique as described above. The amount of ambisyllabicity will be more or less the same for all the experiments (as close as possible to 17.8%), because the threshold (n), which is needed to get this percentage of ambisyllabicity, is dataset-specific. It is the percentages of syllabification after the vowel and after the intervocalic consonant, and the amount of "no syllabification", which are of interest. The results will be evaluated by comparing the proportions of the chosen syllabification patterns using word boundary probabilities to those of the children in the experiment by Gillis and De Schutter (1996). This means very little syllabification after the intervocalic consonant (0.4%) and most syllabification after the vowel (81.6%) are best.

Results

In this paragraph we will systematically take up the research questions formulated above and discuss what answer is suggested by the results of the simulation experiments. We will then propose the characteristics of the 'optimal' simulation, i.e., the one that most closely matches the results of the experiment with children.

Overall effects

1. What is the nature of the primary linguistic data?

   It is not clear from the simulation experiments' results whether language input, intake or production is the source of linguistic knowledge.

   Overall, there is less syllabification after the intervocalic consonant and less after the vowel in experiments using input or intake than in experiments using language output (Table 1).

Table 1: Average results over all simulation experiments using input vs. intake vs. output.

	V.CV	VC.V
input	51.9%	18.3%
intake	50.8%	16.4%
output	57.9%	23%

2. What type of words is the language analysis based on?

   The results suggest that content words — both mono- and polysyllabic — are the words used in a syllabification task.

   On average, there is less syllabification after the intervocalic consonant and more after the vowel in experiments using content words than using all words. The results of experiments using both mono- and polysyllabic words are better than those using only monosyllabic words. There is less syllabification after the consonant and more after the vowel with monosyllabic content words (types) than with all monosyllables (types or tokens), but there is more syllabification after the consonant and less after the vowel with monosyllabic content words (tokens) than with all monosyllables (Table 2).

Table 2: Average results over all simulation experiments using all words vs. content words vs. monosyllables vs. monosyllabic content words.

	V.CV	VC.V
content words	60%	6%
monosyll. content words types	51.1%	7.7%
all words	59.8%	19.8%
monosyllables	44.6%	30.1%
monosyll. content words tokens	17.3%	39.2%

3. What is the influence of frequency on the acquisition of syllabification?

   The simulation experiments suggest that linguistic analysis is based on word types rather than on word tokens.

   If information of word tokens is taken from child language corpora as training material, syllabification occurs more often after the intervocalic consonant and less after the vowel than when word types are used (Table 3).

Table 3: Average results over all simulation experiments using word types vs. word tokens.

	V.CV	VC.V
types	55.9%	14.2%
tokens	48.2%	21.5%

4. What is the optimal representation?

   A representation in phoneme categories appears to be more appropriate than a representation in phonemes.

   Using phoneme categories instead of phonemes generally gives better results, because with phonemes "no syllabification" is often assigned. The amount of test words for which the probabilities for V.CV and for VC.V are both zero can reach up to 81.1%. With phoneme categories, on the contrary, there are no test words that do not get syllabified (Table 4).

   The effect of stress marking in polysyllabic words is not univocal (Table 5). Using phoneme categories, there is less syllabification after the vowel

and less after the consonant with stress marking; using phonemes, stress marking has the opposite effect. Thus, stress has a differential effect depending on the representation of the segments.

Table 4: Average results over all simulation experiments using a representation in phonemes vs. phoneme categories.

	V.CV	VC.V	no syll.
phoneme categories	68%	12.9%	0%
phonemes	36.1%	22.8%	22.5%

Table 5: Average results over the simulation experiments with polysyllabic words using a representation with vs. without stress marking.

	V.CV	VC.V
phoneme categories without stress marking	79.2%	4.3%
phoneme categories with stress marking	76.1%	4.1%
phonemes without stress marking	37.7%	21.6%
phonemes with stress marking	39.5%	22.2%

These tendencies concerning the composition and the representation of the input material are found over the total of all 136 experiments. Now we will discuss the individual experiments that most closely match the behavior of children.

Best results

Similar syllabification patterns (χ^2=1.16, p>0.05) to children's intuitive syllabifications in the experiments by Gillis and De Schutter (1996) are obtained when word boundary probabilities are computed in content words from the intake (types or tokens) or from the input (types) of a child language corpus, represented in terms of segment categories without stress assigned. These results are robust over the three language corpora, in the sense that we find the same results as displayed in figure 1 for the three corpora.

Not only the proportions of the syllabification patterns of the Naive Bayesian learner are similar to children's. Also the factors that influenced the children's syllabification patterns were replicated in the simulations. We will restrict the discussion to the factor of consonant quality.

Gillis and De Schutter (1996) found that children give significantly less ambisyllabic reponses if the intervocalic consonant is a stop (3.4%) than if it is a continuant (19.5%). Looking at the predicted syllable boundaries in the experiments, there are significantly less ambisyllabic responses as well if the intervocalic consonant is a stop (0%) than if it is a continuant (19.4%) (χ^2=9.05, p<0.01). This concerns the same training material, i.e. content words in phoneme categories from intake or input.

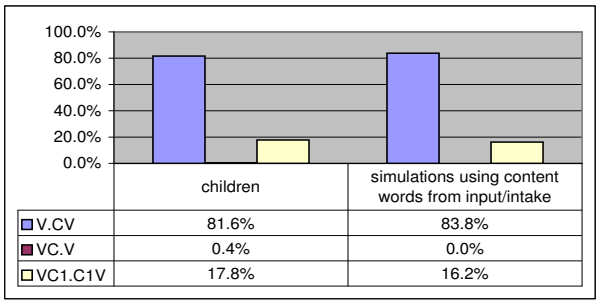


Figure 1: Comparison of syllabification patterns in five- and six-year-olds with results of simulation experiments using content words from input (types) or intake (types or tokens) in phoneme categories.

The observation that a simple statistical model trained on word boundary information (of content words in the input and using a representation in terms of phoneme categories) produces a tight fit with syllabification behavior in children, and the additional evidence that the model matches the children's behavior even at a detailed level of consonant quality is a strong existence proof of the possibility of data-oriented acquisition of the concept of syllables and of syllabification behavior.

Conclusion

Five- to six-year-old children that can't read nor write yet are able to syllabify disyllabic monomorphemic words according to universal rules of syllabification (Gillis & De Schutter, 1996). In this paper, we hypothesized that this intuitive knowledge of syllable boundaries is learned by attending to word boundaries.

To test this hypothesis, statistical word boundary probabilities of phoneme categories were used to predict syllable boundaries in disyllabic monomorphemes with one intervocalic consonant. To compute the probability of a syllable boundary between two phoneme categories, the word-final boundary probability of the first phoneme category and the word-initial boundary probability of the following phoneme category were multiplied. If the difference between the probabilities of the two syllabification possibilities (V.CV and VC.V) does not exceed a maximum limit, ambisyllabicity was assigned (VC1.C1V). Otherwise, the syllable boundary with the highest probability was chosen.

Using this naive Bayesian learning technique, similar syllabification patterns to children's intuitive syllabifications in the experiment by Gillis and De Schutter (1996) were obtained. Best results were achieved when the words used as material to compute word boundary probabilities were content words from the intake (types or tokens) or from the input (types) of a child language corpus, represented in phoneme categories. The same results were found with words from three different child language corpora. Moreover, the quality of the intervocalic consonant has a similar effect on children's intuitive syllabification and on the simulations using word boundary probabilities for syllabification. In both cases there is significantly less ambisyllabicity if the intervocalic consonant is a stop than if it is a continuant.

We have given an existence proof of the hypothesis that syllable boundaries can be learned from word boundaries. The fact that extrapolation from word boundaries to syllable boundaries can be modeled with such a simple statistical mechanism lends support to our initial hypothesis. Furthermore, varying the representations and input data used by this simple statistical learner, we were able to derive a number of interesting more detailed hypotheses about the type of representations and input children may use. More in particular, our results suggest syllable boundaries are most reliably learned from **content words**' boundaries. The semantic saliency of content words seems to be reflected in language production. Moreover, the best results are obtained using **phoneme categories**, rather than the phonemes themselves. This points at the role of sonority in the production of syllables. Phonological saliency is also shown to be an influencing factor, since disyllabic words with intervocalic stops are syllabified significantly differently from disyllables with intervocalic continuants. Finally, we found that the material that worked best to compute boundary probabilities are words from the **intake** or from the **input** of child language corpora. This suggests that children's productions — in this case intuitive syllabifications — could be based on their language input rather than on analysis of their own output. All these findings and predictions from the model have to be further investigated.

References

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–126.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Between the grammar and the physics of speech*. New York: Cambridge University Press.

Elbers, L. (2000). An output-as-input hypothesis in language acquisition. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: inductive and deductive approaches*. Oxford: Oxford University Press.

Fikkert, P. (1998). The acquisition of Dutch phonology. In S. Gillis & A. De Houwer (Eds.), *The acquisition of Dutch*. Amsterdam: John Benjamins.

Gillis, S. & De Schutter, G. (1996). Intuitive syllabification: universals and language specific constraints. *Journal of Child Language*, *23*, 487–514.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kager, R. (1989). *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris.

Kager, R. (1999). *Optimality theory*. Cambridge: Cambridge University Press.

Levelt, C. C., Schiller, N. O., & Levelt, W. J. (2000). The acquisition of syllable types. *Language Acquisition*, *8*, 237–264.

Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mitchell, T. M. (1997). *Machine learning*. Singapore: McGraw-Hill Companies.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118–139.

Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Oehrle (Eds.), *Language sound structure*. Cambridge, MA: MIT Press.

Van der Hulst, H., & Ritter, N. A. (1999). Theories of the syllable. In H. Van der Hulst & N. A. Ritter (Eds.), *The syllable: views and facts*. Berlin: Mouton de Gruyter.

Venneman, T. (1988). *Preference laws for syllable structure and the explanation of sound change*. Berlin: Mouton de Gruyter.

Weijer, J. van de (1998). *Language input for word discovery*. Doctoral dissertation, Max Planck Institute for Psycholinguistics, Nijmegen.

Wijnen, F. (1988). Spontaneous word fragmentations in children: evidence for the syllable as a unit in speech production. *Journal of Phonetics*, *16*, 187–202.

Wijnen, F. (2000). Input, intake and sequence in syntactic development. In M. Beers, B. van de Bogaerde, G. W. Bol, J. de Jong, & C. Rooijmans (Eds.) *From sound to sentence – Studies on first language acquisition*. Groningen: Centre for Language and Cognition.

Reasoning from Data: The Effect of Sample Size and Variability on Children's and Adults' Conclusions

Amy M. Masnick (masnick@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh PA 15213

Bradley J. Morris (bjmorris@pitt.edu)
Learning, Research, & Development Center, University of Pittsburgh
Pittsburgh PA 15260

Abstract

Interpretation of data is a critical part of scientific experimentation because it involves applying one's background theoretical knowledge to the characteristics of the data. Though many researchers have examined the impact of background knowledge, few have considered the impact of the characteristics of the data in making decisions. In this study, we presented 3rd graders, 6th graders, and college undergraduates with a series of datasets that varied in sample size, consistency in data pairs and variability relative to the mean. We found that at all ages, participants showed sensitivity to sample size and whether or not there were overlapping data points in comparative datasets, but that there were age differences in the justifications used and in conclusions drawn from the data.

Interpretation of data is a critical part of scientific experimentation. Expectations about features of the data have been suggested as an important component in assessing data (Kahneman & Tversky, 1973). These expectations are based both on theoretical knowledge about the domain under consideration and on features of the data itself. While a large body of research in scientific thinking examines the influence of domain theory on the evaluation of data (e.g., Klahr, 2000; Koslowski, 1996; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995), little is known about how the characteristics of data influence how children and adults interpret it.

An important component of science is distinguishing real effects from error, or effects caused by factors other than the ones being explored. In the science laboratory, statistics is a vital tool to help make these decisions. When there are differences that are highly unlikely to occur by chance, scientists can feel more confident about drawing conclusions from data.

In daily life, we regularly make decisions about evidence without the aid of formal statistics. In such cases, we resort to relying on theory and expectations. However, there are many situations in which we do not have strong background information, and thus only have evidence based in the data. Elementary school students seem likely to have an especially large handicap in evaluating data – they have a smaller knowledge base about the world and also have less formal knowledge about statistics and its applications.

Students in elementary school are beginning to learn about experimentation and data interpretation, and third through sixth grade is a time of important increases in understanding of basic science fundamentals, such as the control of variables strategy (e.g., Chen & Klahr, 1999). In addition, elementary school teachers routinely assign children to perform repeated trials of events, explaining that this is how science is done (Klahr, Chen & Toth, 2001). In evaluating data in and out of the classroom when children do not know formal statistical techniques, we expect them to rely on their informal knowledge of the area.

But what constitutes "informal" notions of statistical reasoning? We suggest two components: expectations about data distribution and expectations about the influence of sample size. Some research that has examined expectations for the distribution of data has looked at probability estimates. For example, when given data about a series of coin flips, participants expected that a coin would land on "heads" every other flip (Gilovich, 1991). This suggests that the participants had an implicit expectation of the distribution of data in a series of coin flips and that the judgment of "randomness" was (at least in part) based on a mapping between expectations and data patterns. More recently, some have argued that children as young as five or six have a functional understanding of probability (Schlottman, 2001).

Although there is related research in several areas, few studies focus explicitly on the characteristics of the data and the effects this focus has on conclusions. There is some evidence that children at different ages do recognize different properties of datasets, and that this recognition in turn affects the conclusions they draw. For example, Jacobs and Narloch (2001) found that children as young as seven could use sample size and variability information in inferring the likely frequency of a future event. The differences in variability were based on prior knowledge of base rates (i.e., how many elephants have two eyes, compared to how many birds are a specific color). The sample sizes used in this study varied dramatically, with either 1, 3, or 30 instances of an event before the

participant was asked to infer likelihood of other instances of the event, so it is unclear what sample size leads children to feel confident in their predictions.

However, there is also some evidence that children at ages 11 are still struggling to understand the value of repeated measurements within the context of a school science laboratory (Lubben & Millar, 1996). Some children at this age believe repeated measurements are important, but 18% thought that repeated measures are useful because they accommodate scatter in the data.

There is also some evidence that children can distinguish different kinds of variability. Masnick and Klahr (2001) examined second and fourth graders performing experiments in which two balls were simultaneously rolled down ramps and the distance each travels was measured. The children expected that on a new trial using the same experimental set-up, the relative positions of the two balls would remain the same, but that the precise location of each ball might be different. That is, they were able to make a distinction between small differences in individual data points and larger differences in sample means.

Students' expectations about the essential features of data and the features of a specific dataset may allow them to recognize data as consistent or inconsistent with their expectations about its distributions. These expectations may in turn guide decisions about the usefulness of the data and the extent to which the data are relevant to explanatory theories. Thus, the characteristics of the data partly determine the extent to which they are used to guide the formation or modification of explanatory theories.

One related body of research has examined how children use data (in this case covariation between events) to detect causal relationships between elements (Shaklee & Mims, 1981; Kuhn, 1989). These studies looked at how children evaluate evidence when events occur together all the time, some of the time, or none of the time.

In a series of studies by Shaklee and her colleagues, students in grades 2-8 and adults were presented with data about two events (e.g., plant growth-healthy/unhealthy and use of bug spray- yes/no) in a 2x2 contingency table. From these data, students were asked to determine the nature of the causal relationship between the events (i.e., presence and direction of relationship). A majority of participants at all ages did not use conditional probability rules to determine covariation, yet many children used a strategy for summing the diagonals in the contingency table (Shaklee & Paszek, 1985). However both children and adults could use a conditional probability rule if instructed (Shaklee, Holt, Elek, & Hall, 1988).

Kuhn and her colleagues (Kuhn, Amsel, & O'Loughlin, 1988; Kuhn et al., 1995) extended this line of research and examined the effect of detecting covariation on the participant's prior beliefs about an event. For example Kuhn et al. (1988) interviewed sixth and ninth grade children about the relationship between consuming different types of foods and catching colds. Information about each child's prior beliefs was used to provide each child with two sets of data: one that confirmed their prior beliefs and one that disconfirmed their prior beliefs. The researchers argued that children did not clearly distinguish theory and evidence because children often distorted the evidence to match their prior beliefs.

Although these studies suggest that data itself is important in detecting causal relationships and in evaluating hypotheses, there is little evidence about the point at which children (and adults) detect covariation in a particular dataset. In fact, in her review of scientific reasoning literature Zimmerman (2000) states that "it is not clear *how large* the difference must be in order to conclude that the two events are related" (p. 115).

If students do rely on evidence to extend or modify a theory, how do they go about such a task? Students' notions of data variability may help them determine how to weigh the potential importance of different types of information. For example, data with little variability may be considered more useful in drawing a conclusion than data with greater variability *a priori*. The relevance of the data to theory may be separately evaluated.

One approach to understanding how children use evidence to extend and modify theories is to look at category induction. In a series of studies, Gutheil and Gelman (1997) presented 8- and 9-year-old children and college adults with series of category exemplars. Participants were asked whether a given property would be expected to occur in a new exemplar. The diversity and sample size of the initial sets were varied. Results suggested that children used diversity and sample size information only in combination, but were unable to use just one successfully to infer category membership. Adults, in contrast, used each property independently, as well as jointly, in inferring category membership. In these studies, however, determining that a set was homogenous or diverse relied on domain knowledge acquired outside of the experiment.

Clearly, patterns of data play a key role in scientific inference, but what characteristics of the data guide inferences about their utility? We suggest that, in drawing conclusions about comparative data, three characteristics that indicate the amount of variation in the data are key: consistency within the patterns of data (i.e., the relative sizes of the data points), the magnitude of differences (i.e., the range of each set of data) and the presence of outliers. Data that show high consistency in the direction of effects, small differences in magnitude and few outliers suggests little variability. Data that shows low consistency in the direction of effects, large differences in magnitude, and many outliers suggests more variability. This information about variability can be assessed increasingly well with a larger number of data

points, increasing the degree to which the data itself can inform an interpretation.

As a preliminary exploration of this area, we presented children and college students with sets of comparative data, and asked them to draw conclusions about differences between the sets. The data were varied systematically in number of data points presented and consistency within the pattern of data.

Method

Participants Thirty nine third graders (mean age = 9.1), seventeen sixth graders (mean age = 11.8), and fifty college undergraduates (mean age = 20.2) participated.

Procedure All participants were interviewed individually. Participants were randomly assigned to one of two conditions. The conditions differed in the cover story for the data presented. In the first condition, each participant was read the following information:

> Some engineers are testing new sports equipment. Right now, they are looking at the quality of different sports balls, like tennis balls, golf balls and baseballs. For example, when they want to find out about golf balls, they use a special robot launcher to test two balls from the same factory. They use a robot launcher because they can program the robot to launch the ball with the same amount of force each time. Sometimes they test the balls more than once. After they run the tests, they look at the results to see what they can learn.

In the second condition, we used an isomorphic background story in which two athletes were trying out for one slot on a team in different sports. The coaches asked the participants to perform certain tasks (e.g., hit a golf ball as far as possible) to assess which athlete would be better for the team. This condition was designed to see if adding information about a highly likely potential source of variability (human error) would change participants' responses in any way.

After reading the cover story, the participants were shown a series of datasets, one at a time. For each example, there was data for either two different balls of the same type, which were not given any distinguishing characteristics (e.g., "Baseball A" and "Baseball B"), or for two athletes about which there was no information other than their names (e.g., "Alan" and "Bill"). In the athlete condition, different names were used for each story, to prevent any carry-over knowledge effect. For each dataset, there were 1, 2, 4, or 6 pairs of data. Each page contained two columns of data: one listing the distance the first ball traveled and one listing the distance the second ball traveled.

The datasets varied in (a) sample size, (b) whether the datasets overlapped or not, and (c) in whether the variability in the data was high or low relative to the means. Each participant received a total of 14 comparisons, with 8 trials including no overlap (sample size 1, 2, 4, and 6), and 6 trials including one or two overlapping data points (sample size 4 with one overlapping data point, and sample size 6 with one and two overlapping data points). Half of the trials had high variability, in which the standard deviation was 15-20% of the mean, and half had low variability, in which the standard deviation was less than 2% of the mean. Each of the fourteen trials tested a different type of sports ball. (See Table 1 for specific examples of the different data characteristics.)

Table 1: Examples of datasets

Example 1: Six data pairs, no overlapping data points, low variability within columns relative to the mean, robot condition

Golf Ball A	Golf Ball B
466 feet	447 feet
449 feet	429 feet
452 feet	430 feet
465 feet	446 feet
456 feet	437 feet
448 feet	433 feet

Example 2: Four data pairs, one overlapping pair (3 out of four times Carla throws farther), high variability within columns relative to the mean, athlete condition

Carla	Diana
51 feet	38 feet
63 feet	50 feet
43 feet	56 feet
57 feet	44 feet

For each dataset, participants were first asked what the engineer or coach could find out as a result of this information and to explain any reasons for their answer. Then they were asked how sure they were about these conclusions. To answer the questions about sureness, participants were offered a four-level scale from which to select their answer, choosing among "not so sure," "kind of sure," "pretty sure," and "totally sure."

Participants were next asked if the robot or athlete launched Ball A again, exactly how far they thought the ball would go, and how sure they were that Ball A would travel X feet. They were asked the same questions about Ball B, and then they were asked how sure they were that the ball they had just named as going the farther distance would actually go farther. For example, if they said that they expected Ball A to travel 50 feet and Ball B to travel 60 feet, they were asked how sure they were that Ball B would travel farther. Again, in addition to rating their

sureness, they were asked to offer any reasons for their choices.

This series of questions was repeated for each of the 14 sets of data.

Results

Measures Participants rated how sure they were about their conclusions four times for each dataset: They gave a rating of their confidence in the initial conclusions drawn from the data, in the predictions they made of exactly how far each ball would go, and about which ball would go farther on the next trial. The ratings were assessed on a four-point scale, converted to a four-point variable, with 1 equivalent to not sure and 4 equivalent to totally sure.

In addition, participants offered reasons for their initial conclusions and final predictions of relative position. These reasons were coded for mention of any of the following factors: the precise proportion of times one ball went farther, a trend in the data, sample size, the magnitude of the difference between the two datasets, whether the datasets overlapped, a property of the ball that affected the results, a property of the robot or athlete that affected the results.

Participants also made specific numerical predictions of how far the two balls would go if launched one more time.

Levels of confidence Mixed model ANOVAs were used to examine the effects of condition, data size, overlap, and variability on ratings of sureness on the four-point scale. For each assessment, data size, overlap, and variability were within-subjects variables, and subject was treated as a random variable.

Across all age groups, there was no effect of condition (robot or athlete) on ratings of sureness. Therefore, on later analyses of these questions, we collapsed the data across conditions.

For conclusions about how sure the engineer or coach could be, based on the original data, that one athlete/ball threw/went farther, there were several notable effects. College students were highly sensitive to sample size, the sixth graders showed a small but not significant trend upwards, and the third graders showed a small but significant inverse trend. The data are summarized in Figure 1. Overall, there was a highly significant effect of grade ($F(2, 1448) = 56.38$, $p < 0.0001$), with third graders on average much more sure than sixth graders, who in turn were more sure than college students, across all sample sizes smaller than 6.

In addition to sample size, participants demonstrated a sensitivity to the presence of overlapping data points, such that they were less sure of conclusions when the data contained overlapping points. This effect was significant for all grades.

Similar patterns emerged on the assessments of participants' sureness about their predictions, both about the specific distance the balls would travel, and about which ball they expected to go farther on a repeated trial. The strongest relationships were for the college students, who appeared to always link their sureness rating to the number of data points, the proportion of overlapping data points, and occasionally to the level of variation in the data.

Although similar features in the data affected the level of sureness, there were striking differences in the sureness responses to the different questions. Participants were much less sure about specific predictions than about overall conclusions or about predictions of relative placement on a future trial. Overall, general linear models for each dataset, considering the measure of sureness as a repeated factor indicate a strong relationship between both grade and the specific question asked (i.e., the assessment of sureness for general conclusions, specific predictions and relative predictions). Across all grades, there was a relationship between the level of sureness and what question was asked, but it was weakest in the third graders and strongest in the sixth graders.

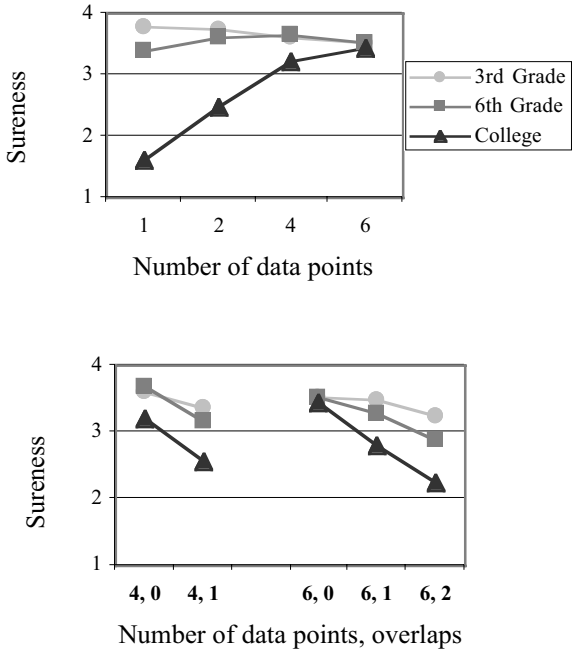


Figure 1: Average ratings of sureness by grade, sample size and number of overlapping data points

Reasons offered Participants offered justifications for confidence in their conclusions and for the predictions they made about which ball would go farther on a subsequent trial. We examined whether participants mentioned each reason at any point in response to a question about each of the fourteen datasets. We then ran Chi-square tests to look for grade differences in the frequency of participants mentioning each factor.

Table 2 presents a summary of these results. An overwhelming majority of the responses were in reference to the data and not to theoretical issues such as properties of the ball or robot/athlete. Young children were most likely to mention either the proportion of the data (e.g., "Five out of six times A went farther") or a trend in the data (e.g., "B generally went farther"). In general, college students used a much wider range of responses than younger children, with nearly all of them mentioning sample size at least once. Interestingly, despite the significant increasing trend, only a small percentage of the participants mentioned within-column variability, one of the factors we manipulated (variability in Table 2).

Table 2: Percentage of participants at each age who mentioned each justification for their sureness ratings

	3rd Gr.	6th Gr.	College
Data responses			
Precise proportion	92	86	86
Trend in data*	90	89	100
Sample size**	10	25	96
Overlap	56	64	72
Variability**	0	11	28
No Overlap**	5	7	58
Magnitude of diffs**	36	75	90
Outlier*	0	0	10
Theory responses			
Ball property	8	14	14
Robot/athlete property	18	18	22

Grade differences: *p<0.05; **p<0.01

Justifications of predictions of which ball would go farther in a new trial followed a very similar pattern as that described above, for all three age groups.

One area in which a condition difference might be expected is in use of justifications that refer to qualities of the ball, the robot, or the athlete. Mentioning the property of the robot or athlete did vary considerably by condition, with nearly all mentions in the athlete condition (i.e., participants sometimes said that a property of athlete was a reason for the outcome, but almost never attributed it to a property of the robot). This trend was even stronger when justifying predictions of future outcomes.

Predictions Participants predicted how far each ball would go if the experiment were repeated. The data from two third graders and one sixth grader were not included in this analysis because they included numbers that differed from the mean by more than twice the range of the data. These outliers skewed the data considerably, and suggested that these participants did not understand the prediction task.

Overall, however, the participants were very good at predicting how far the balls would go, and their predictions averaged close to the mean. Third graders averaged predictions that were 108.4% of the actual data means (SD =10.0); sixth graders averaged predictions that were 102.7% of the means (SD = 3.7); and college students averaged 100.1% of the means (SD = 1.4).

Discussion

Our overall conclusion is that in the absence of clear domain knowledge upon which to base theoretical explanations, children and college students paid attention to several features of data. At the same time, there were clear age differences in many responses, indicating changes over time that likely come from a combination of education, experience, and development.

In all age groups, participants were less confident about conclusions from datasets in which there were overlapping data points, indicating a sensitivity to variation in the patterns of the data. College students were significantly more confident when there were more data points, though third graders were actually slightly more confident with smaller sample sizes. It is possible that the third graders were overwhelmed by the variability and became more confused about drawing conclusions when there were more data points to consider.

Participants also showed an appreciation for some types of variability by differentiating their sureness ratings for different types of predictions. At all ages, they were more sure of conclusions about relative distances on a future trial than about specific distances, with the effect most pronounced in college students. This response pattern indicates an expectation that variation is more likely in precise measurements than in overall patterns of results. Participants seemed less attuned to within-column variability in the data, rarely citing it as a justification for either their conclusions or predictions, though older participants were still more likely to cite it.

The lack of major differences between conditions was an unexpected outcome. We had anticipated that the two different ways of framing the data would lead to different theoretical explanations of the data. However, most of the justifications offered for drawing conclusions from the data were based on the numerical evidence (e.g., a trend in the data, sample size), while very few were linked to mechanistic explanations such as a feature of the ball that might cause the outcome. There was a small but significant trend for those in the athlete condition to be more likely to justify their explanations and predictions by suggesting that the athletes may have varied in some way. However, a minority of participants at all age levels used such theoretical justifications. Some researchers have argued that children mistakenly justify conclusions that should be based on data by using their background theoretical knowledge (e.g., Kuhn, et al., 1995). In contrast, we suggest that in fact when children do not have background knowledge upon which to rely, they are likely to talk about the data in justifying conclusions.

In making distance predictions, overestimation was more common than underestimation, particularly among the youngest children. Similarly, third graders often claimed to be totally sure of conclusions they could draw after seeing only one pair of data, while college students tended to reserve their enthusiasm until seeing at least four consistent pairs of data. In general it appears that third graders often overestimate both their confidence in their ability to judge the quality of evidence, and their predictions of future performance. College students were very skilled at basing predictions of future events on the mean of observed events.

This study is a first step toward a clearer understanding of the many factors that influence the use of data in different contexts. Many other data manipulations could be examined to explore this question more thoroughly. For example, one could manipulate the size of within-column variation, the relative size of outliers, and the size of the means. In addition, one could consider multiple groups of data or a single set of data to be evaluated.

Our long-term goal is to better understand the interaction between features of data and theoretical framing, as people of all ages most often encounter data in contexts in which they have some background knowledge. This study was designed to tease apart some of the specific features of data that are used when there is not a lot of theoretical background knowledge influencing conclusions drawn. Future studies will continue to examine different characteristics of data within a range of contexts, to learn more about how they interact to affect reasoning.

Acknowledgments

This research was funded in part by grants to David Klahr from the McDonnell Foundation and from NIH. We owe many thanks to Anne Siegel and Jen Schnakenberg for assistance with data collection, data entry, and draft comments. Thanks also to David Klahr and three anonymous reviewers for comments on an earlier version of this paper.

References

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: The Free Press.

Gutheil, G., & Gelman, S. (1997). Children's use of sample size and diversity information within basic-level categories. *Journal of Experimental Child Psychology*, *64*, 159-174.

Jacobs, J. E., & Narloch, R. H. (2001). Children's use of sample size and variability to make social inferences. *Applied Developmental Psychology*, *22*, 311-331.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237-251.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Klahr, D., Chen, Z., & Toth, E. E. (2001). Cognitive development and science education: Ships passing in the night or beacons of mutual illumination. In S. M. Carver & D. Klahr (Eds.) *Cognition and instruction: 25 years of progress* (pp. 75-119). Mahwah, NJ: Lawrence Erlbaum Associates.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasonin*g. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientist*s. Psychological Review, 96*, 674-689.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, *60* (4), pp. 1-128.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*, 955-968.

Masnick, A. M., & Klahr, D. (2001). Elementary school children's understanding of experimental error. In J. D. Moore & K. Stenning (Eds). *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, (pp.600-605). Mahwah, NJ: Lawrence Erlbaum Associates.

Schlottman, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Development*, *72*, 103-122.

Shaklee, H., Holt, P., Elek, S., & Hall, L. (1988). Covariation judgment: Improving rule use among children, adolescents and adults. *Child Development*, *59*, 755-768.

Shaklee, H. & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development*, *52*, 317-325.

Shaklee, H., & Paszek, D. (1985). Covariation judgment: Systematic rule use in middle childhood. *Child Development, 56*, 1229-1240.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*, 99-149.

Reusable Templates in Human Performance Modeling

Michael Matessa[1], Alonso Vera[1], Bonnie John[2], Roger Remington[1], and Michael Freed[1]
({rremington, mmatessa, avera, mfreed}@arc.nasa.gov)

[1]Cognition Laboratory, Mailstop 262-4, NASA Ames Research Center
Moffett Field, CA 94035 USA

[2]School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Current computational modeling of human performance can benefit from reusable building blocks of human behavior. Using CPM-GOMS, a cognitively-based task analysis method used in HCI, we have been exploring the concept of reusable templates of common behaviors and their efficacy for generating zero-parameter *a priori* predictions of complex human behavior. This paper details the features we believe are important when moving from hand-crafted models of particular tasks to reusable building blocks of commonly occurring behavior. As this becomes common practice, proportionately more attention can be paid to the task analysis specific to each new domain.

Introduction

To model human behavior successfully one needs to be able to decompose a complex task into a set of primitive operations to which performance parameters may be assigned. These primitives represent the building blocks from which behavior will be constructed such that performance can be predicted for entire task sequences. Thus, the choice of primitives and the method of combining them to construct larger behavior sequences are critical if performance estimates are to be at all accurate. In this paper, we describe a method for combining basic cognitive, perceptual and motor operations into larger behavioral units. These larger units, which will be referred to as *templates*, are applications of psychological theory about the perceptual, motor and cognitive process underlying human performance. Two main points will be argued: 1) task-level reuse is important and 2) behavioral templates are a good way to achieve this reuse. We will present data for a simple HCI task and describe the templates we borrowed (reused) to model it.

Model reuse is a profitable avenue to explore for four reasons. First, consider mousing to a button and clicking on it as an example. People learn and use this skill in most interactions with computers. A model of mousing and clicking on a button should be applicable to many HCI tasks; a new model of mousing and clicking on a button should not be built from scratch for each new task. Second, reuse provides expertise beyond a single researcher. That is, the model for mousing and clicking on a button can be built by researchers with expertise on visual/motor behaviors and the complex model-builder can benefit from that expertise. Third, reuse provides external verification of the component models. If the model for mousing and clicking on a button predicts the behavior well in the context of a complex task, the data provides an independent test of the mechanisms of that model. Finally, reuse provides additional constraint on models of complex tasks. If the mousing and clicking on a button model predicts the behavior well, the HCI modeler should not change the basic mechanisms of the model simply to make it work in a new domain.

Cognitive architectures such as ACT-R (Anderson & Lebiere, 1998) and Soar (Newell, 1990) embody a large set of reusable constraints on behavior prediction. These, however, are mostly at the cognitive level rather than at a task level. We know how to reuse architectures but not content. Modeling performance on human-computer interaction (HCI) tasks is valuable but currently suffers from several problems. Models of task performance need to be handcrafted and typically take a long time to be created. Neither whole models nor their component parts tend to be reused, allowing little transfer of code from one task model to the next. It is difficult to incorporate psychological knowledge into models and cognitive/psychological expertise is needed to do so. Once a model is complete, it may be uninformative with respect to a next attempt at modeling.

Other disciplines that build complex systems, like engineering and computer science, have successfully employed reuse as an approach to tame complexity and this is an approach that cognitive science should explore as well. There have been numerous arguments about the benefits of using a unified cognitive architecture to provide power, structure and constraint (e.g., Anderson, 1983; John & Altmann, 1999; Newell, 1990), but fewer efforts to incorporate previously-built models of generalized capabilities into models of more complex tasks (e.g., Nelson, Lehman & John, 1994).

Cognitive Modeling Methods

Cognitive architectures such as ACT-R and Soar have attempted to solve some of these problems. Underlying mechanisms embody the psychological theories of the architectures, and so the theories do not have to be explicitly coded for each model. Attempts have been made to reuse models at the task level, but in general this is not a widely adopted practice (e.g., John & Lallement, 2000; Nelson, Lehman & John, 1994). As a result, models are mostly handcrafted and take a long time to create.

The field HCI uses cognitive models in several ways. It often is less interested in the process of modeling than in getting results quickly, and has therefore put emphasis on modeling frameworks that are easier to learn and use than the more complex cognitive architectures like Soar and ACT-R (John, 1998). In addition, that field has sought modeling procedures that reuse actual pieces of models as well as the framework. For instance, GOMS models (Card, Moran & Newell, 1983) are constructed by hierarchical goal decomposition with reusable, empirically-determined execution times assigned to particular goals.

The GOMS methodology has proven highly successful in predicting task completion times for skilled users in routine human computer interaction (HCI) tasks (e.g., Gray, John, & Atwood, 1993; John, Vera, & Newell, 1994). GOMS is really a family of analysis techniques in which performance predictions emerge by combining a task decomposition with estimates of completion times for steps in the decomposition. The task decomposition produces a representation of the task as a set of nested goal states that include an initial state and a final goal state. The user is assumed to move from one goal state to another by applying operators that represent actions, such as moving a mouse or reading a word. The set of nested goal states often resembles a hierarchy, but need not form a strict hierarchy. The iterative decomposition into goals and nested subgoals can terminate in leaf nodes (primitives) of any desired granularity, the choice of level of detail dependent on the predictions required. Times are assigned to the operators that transition between goal states, with additional times often assigned to subgoal completion (Kieras, 1994). Since GOMS is meant to model routine behavior, the user is assumed to have methods that apply sequences of operators and subgoals to achieve a goal. Selection rules are applied when there is more than one method to achieve a goal.

The CPM-GOMS extension to GOMS (John, 1990) adds psychological knowledge to the GOMS goal decomposition by expressing common HCI tasks (e.g. reading from a screen, typing) as patterns of Model Human Processor (MHP; Card, Moran & Newell, 1983) operations of its cognitive, perceptual, and motor processors. These patterns, or templates, can help cognitive modeling by grouping psychological knowledge into reusable chunks that can be organized into larger behavioral sequences. Although approaching modeling with reusable templates has been taught for a decade (John & Gray, 1992), it has not become widespread, possibly because CPM-GOMS was not in computational form (it had to be done by hand by drawing PERT charts) until recently (John, Vera, Matessa, Freed & Remington, 2002).

Template Structure

Templates are reusable applications of psychological theory that describe short behavioral sequences. Thirteen templates were offered to CPM-GOMS modelers in tutorials and classes in the early 1990s (John & Gray, 1992) and others have been added since then. One example of a CPM-GOMS template for mouse clicking (created by Gray & Boehm-Davis, 2000) can be seen in Figure 1. The template incorporates a psychological theory of the cognitive, perceptual, and motor components of mouse clicking and dependencies between these components. The template was developed in the context of a simple task of clicking on lit circles, but has successfully been reused in the context of clicking to operate a simulation of an automated teller machine (John, et al., 2002).

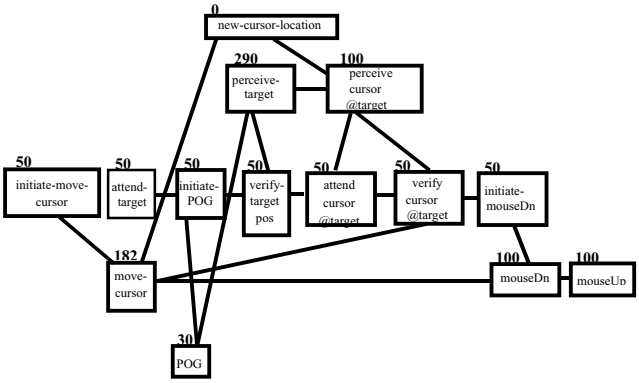


Figure 1: Model of carefully moving the cursor to a target and clicking the mouse button (adapted from Gray & Boehm-Davis, 2000).

The psychology in this template is in the form of the durations, the cognitive bottleneck, the logical dependencies, and the task dependencies. The *durations* for the cognitive, perceptual, and motor components of the template are empirically derived MHP-level estimates of these activities. CPM-GOMS assumes a *cognitive bottleneck,* where cognitive processes cannot execute in parallel. The cognitive resource stream is therefore serially scheduled based on the logical and task dependencies as described below.

According to CPM-GOMS, each motor activity must be preceded by a cognitive activity that initiates it. This is an example of a *logical dependency* — a motor action cannot take place unless a cognitive motor initiation activity has occurred. That you cannot click on a target

until you have moused over to it is an example of a *task dependency*. It is not a CPM-GOMS level logical dependency nor is it a necessary consequence of the cognitive bottleneck; it is true because the task requires it to be so. So, for example, a possible performance error might be to click before the mousing movement to the target is complete. In contrast, executing a motor action without a preceding cognitive initiation is not a possible performance error. Similarly, executing two cognitive activities in parallel is not a possible performance error. This combination of constraints embodies the unique application of psychological theory to the crafting of each template.

Interleaving Templates

Embodying psychological theory in separate templates is only the first step. The challenge is that CPM-GOMS templates need to be interleaved to fully capture the time course of behavior. Simply adding up the performance time predicted by a sequence of templates produces a time that is longer than that of human performance. A key CPM-GOMS assumption is that humans are able to perform certain components of the templates not in strict sequence but interleaved so that some components of a later template can occur in an earlier template.

A concrete example of this interleaving can be seen in a task as simple as hand-washing. While most eye fixations are related to immediate actions such as turning on the faucet, a small number are made to objects relevant only to future actions. Pelz and Canosa (2001), with the aid of an eye tracker, observed that subjects consistently made eye movements to the towel during earlier parts of the hand-washing process.

Part of the psychological theory embedded in templates is the knowledge of what components of one template can occur in parallel with components of another template. This interleaving theory has until recently only been implemented in paper-and-pencil. Details of how interleaving is implemented in our system are presented later. When cognitive components have a dependency on perceptual or motor components which take a relatively long time to complete, slack time may occur when the cognitive resource is not being used. Interleaving involves filling up this slack time with activity from the cognitive components of the next template.

ATM Study & Data

CPM-GOMS has been demonstrated to make accurate zero-parameter *a priori* predictions of skilled HCI behavior. Using templates, we created a CPM-GOMS model of a simple HCI task — withdrawing money from an ATM. We gave two users extensive practice with this task because CPM-GOMS models are expected to predict the performance of highly-skilled users (Baskin & John, 1998).

The ATM Task

The task was to make an $80 withdraw from a checking account on a Visual Basic simulation of an automated teller machine. Users interacted with the ATM by using a mouse to click on simulated keys or slots. The users were instructed to follow the following steps:

Insert card (click on the picture of the card slot)
Enter PIN (click on the 4, 9, 0, and 1 buttons in turn)
Press OK (click on OK button)
Select transaction type (click on withdraw button)
Select account (click on checking button)
Enter amount (click on 8 and 0 buttons)
Select correct/not correct (click on correct button)
Take cash (click on the picture of the cash slot)
Select another transaction (click on No button)
Take card (click on the picture of the card slot)
Take receipt (click on the picture of the cash slot)

This task was repeated 200 times by the users. This level of practice is comparable to that used by both Card, Moran, and Newell (1983) in a text editing task and Baskin and John (1998) in a CAD drawing task when they explored the effects of extensive practice on match to various GOMS models.

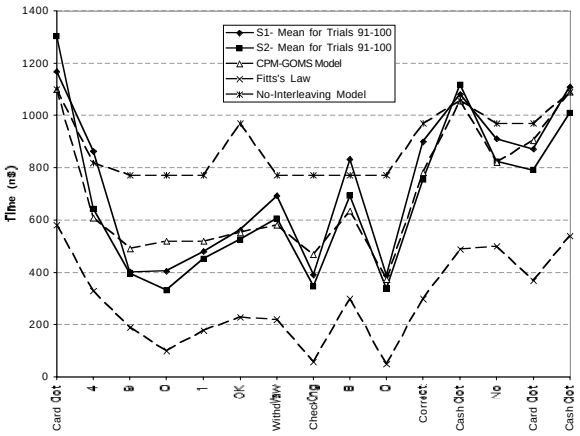


Figure 2: Click times for users and models

The CPM-GOMS Model

The CPM-GOMS model was created by first expressing the hierarchical goal structure of the ATM task. At the bottom of the hierarchy were two CPM-GOMS templates: Slow-Move-Click and Fast-Move-Click. The cognitive, perceptual, and motor components of these templates were taken directly from Gray and Boehm-Davis (2000) where predictions from these templates were compared to data from several variations of a simple target selection task.

The Slow-Move-Click template is shown in PERT chart form in Figure 1. Because there was uncertainty about where a target would appear in each trial, Gray and Boehm-Davis considered Slow-Move-Click to represent a careful selection of a visible target. Fast-

Move-Click represented a more confident selection of a target when the user knew where the target was to appear. In our model, we choose to use Fast-Move-Click for clicking on the ATM buttons because they were a reasonable size and Slow-Move-Click for clicking on the card and cash slots because these slots were thinner and difficult to hit unless the user was careful.

Comparing the Model to Data

Figure 2 shows a comparison of the CPM-GOMS model predictions of mouse click times and mean user click times. Because Baskin and John (1998) have found that CPM-GOMS models predict behavior well at around the 100th trial of a practiced procedure, the means of trials 91-100 for each ATM user are shown. To see the benefit of interleaved templates, predictions from two other models are also shown: a model using only Fitts s Law predictions (a motor time prediction based only on target geometry), and the CPM-GOMS model with sequential templates but no interleaving.

The first thing to notice is the good *a priori* fit of the model to user data. The degree to which interleaving contributes to this fit can be seen by comparing the interleaving and non-interleaving models. The *non-interleaving model* generally predicts a longer time for a mouse click and does not capture the variation of user click times. This variation is better captured by the *Fitts s Law only* model, but the model does not represent the perceptual and cognitive processes incorporated into templates and so predicts faster click times than users produce. These comparisons show that templates contain important predictions of perceptual, cognitive, and motor processes, and that the theory of template interleaving can capture the abilities of users to save time by performing parallel processes.

The important point to emphasize is that the templates used to make these predictions were developed for a very different task and were successfully reused in the present model. Also, note that the tool we have developed to implement CPM-GOMS, Apex (Freed, 1998b, Freed & Remington, 2000), allows us to easily generate alternative versions of the model (e.g., just Fitts s Law, templates with no interleaving). We will expand on the workings of the tool below.

Implementing Templates

How does an organism organize its resource allocation over an extended period of time over an extensive sequence of behaviors? Hierarchical task decomposition is a convenient formalism for an analyst to record and communicate their analysis of a task but it also seems to have some psychological reality in how people organize their environment and cognitive resources to approach a complex task. Many representations of human behavior use hierarchical task decomposition, from Scientific Management at the turn of the century, to HTA (Kirwan and Ainsworth, 1992) to GOMS, Soar and ACT-R. However, in these latter architectures, the hierarchical task decomposition bottoms out at the level of operations of the underlying cognitive, perceptual and motor processors.

It is relatively easy to perform a hierarchical task analysis for each new task domain, and necessary because each domain has its own objects, tasks, knowledge, and procedures. It is much more difficult to describe the cognitive, motor, and perceptual processes underlying every new task. However, many tasks bottom out at the same component behaviors constrained by human abilities. The component behaviors are actions like visual search, mouse movements, and typing. These types of actions are perfect candidates for re-use because they occur repeatedly in many tasks, and must be realized in operations closer to the architecture than is necessary to describe the task domain. Removing the burden of understanding and programming in the underlying cognitive architecture in the form of reusable templates could make cognitive modeling more accessible to a wider range of domain experts.

What does it take to create templates in a cognitive architecture?

- A systematic relationship between the bottom level of the hierarchical task decomposition, the templates, and their realization in the underlying architecture
- A systematic relationship at the boundaries between templates realized in the underlying architecture. This may be more complicated than simple serial execution, as we will describe in the context of CPM-GOMS, the Model Human Processor, and its embodiment in Apex.

Template Grain Size

Generalizeable behavior templates should exist at different grain sizes. From a practical modeling perspective, it is valuable to represent basic HCI behaviors in different size chunks. For example, one branch of a hierarchical goal decomposition may terminate at simply clicking on an item whereas another might terminate at a more complex action such as choosing an item from a pull-down menu. The latter action encompasses the first (twice, in fact), but both are useful as separate templates. Along with generalizability, a valuable constraint on the grain size of templates is the level at which hierarchical goal decomposition would bottom out. For example, a template that is just the reaching component of typing a key (i.e., putting your finger over it without pressing on it) would be very general; it would be Fitts s Law plus the associated cognitive and perceptual components. However, it would not typically be where a hierarchical goal decomposition would bottom out. A hierarchical goal decomposition is more likely to bottom out with activities such as click-on-button or press-key than just

the reaching component of these activities, and templates should reflect this.

Implementing Interleaving

The template-builder gets to deliberately program these dependencies. The relationship between templates in CPM-GOMS involves not sequencing, but interleaving of parallel operators on multiple resources. In highly skilled behavior, people attain a high degree of parallelism in their behavior. To illustrate this point, think about printing a document in the word processor you regularly use. It is not uncommon for people to select the Print command then move the mouse to the place where the OK button will appear well before the dialog box is visible. A hierarchical task analysis would list the steps (1) select Print , (2) wait for dialog box, and (3) click on OK , but the realization in the underlying architecture should allow the mouse movement associated with step (3) to precede the completion of step (2), whereas the clicking associate with step (3) must indeed wait for the completion of step (2) or the task will not be successfully completed. Thus, the relationship between templates is a complex system of interleaving of the architectural-level operations.

Apex allows manipulation of serial and parallel resource in exactly the form required by CPM-GOMS. Apex (Freed, 1998a) has a flexible resource management system that allowed the implementation of the constraints necessary to produce CPM-GOMS template interleaving. A complete description of the implementational rules guiding interleaving in CPM-GOMS as embodied in Apex is available elsewhere (Berkovich & Kwong, 2002; John et al., 2002). In general, we use three facets of the Apex architecture:

1. resource modules (i.e., cognition, hands, point of gaze, visual perception),
2. priorities on the templates, and
3. virtual resources to record a template s intention to use a resource.

Resource modules. To accomplish templates, operators consume resources at the architectural level. The resource modules act serially within themselves, so when they are occupied by an operator, other operators that require that resource must wait until that resource is free again. When more than one operator requires the same resource at one time, there needs to be a way to resolve the conflict and assign that resource to only one operator. Within a template, such conflicts are either avoided because of logical dependencies between the operators (e.g., perceiving a target cannot happen until the eyes have moved to that target) or resolved randomly because the template-builder has determined that it does not matter for the completion of the task whether one goes before the other. When there is contention for resources between operators in different templates, an additional mechanism for conflict resolution is needed.

Priorities. When there is competition for a resource between templates, priorities of the templates is used. Each template is assigned a priority associated with its order in the hierarchical task decomposition. In the printing example above, selecting the Print command (step 1) has the highest priority, waiting for the dialog box (step 2) has the next highest priority, and clicking the OK button (step 3) has the lowest priority. Both steps 1 and 3 will contend for the right hand to move the mouse to their respective targets, and step 1 will win by its priority, ensuring that the print command is selected before the mouse is moved to where the OK button will appear. However, step 2 (waiting) does not require the right hand, so the mouse is free to move to the location of the OK button before the dialog box appears.

Virtual resources. However, resources and priorities are not enough. For instance, the template to select the print command (step 1) requires the eyes to move to and perceive the menu title, and later within the same template to move to and perceive the desired menu item. However, while the hand is completing the move and click to the menu title and the menu is coming up, the eye resource is free. Therefore, step 3, clicking on the OK button wants the eye to move to the location of the OK button, and since the resource is free, there is no contention and priorities do not come into play. Thus, the eye resource could be assigned to the OK button even though the menu item will be appearing in a fraction of a second and the eye will be needed there. Therefore, to reserve the eye for the higher-priority template (in this case, the select-menu-item(Print)), we developed the idea of virtual resources. Virtual resources are analyst-defined entities that act like regular resources. Operators consume these resources and they have to be allocated using priorities. So at the beginning of the select-menu-item (Print) template s first eye movement, it reserves the eye-block virtual resource and only releases that virtual resource after its last visual perception is complete. Since all eye-movements and visual perception require this eye-block virtual resource, this technique blocks any lower-priority template from stealing that resource before a higher-priority template is finished with it.

Discussion

We have a computational system that implements and automatically interleaves reusable behavioral templates. While only two templates have been presented here in detail (Slow-Move-Click and Fast-Move-Click), several others have been implemented in our system for touch-typing and applied to a different computer aided design task, and several more are under development for interacting with the flight maintenance system found on commercial airliners. Previous work has shown that other currently existing templates in the literature are useful for simulations done by hand (Gray, John, &

Atwood, (1993) being a good example). We will work to implement these as well. As more templates are accumulated, issues such as coverage of possible behaviors can be addressed.

We are currently exploring the possibility of generating CPM-GOMS models from ACT-RPM. ACT-RPM imposes its own set of constraints on cognitive, perceptual and motor resources, many of them different from those of CPM-GOMS. This makes the generation of genuine CPM-GOMS behavior from ACT-RPM more challenging but also promises the possibility of extending the predictive scope of CPM-GOMS to a wider range of skill (i.e., from novice to expert) because of ACT-R s new production compilation mechanism (Taatgen & Lee, submitted)

In order for cognitive modeling to come into wider use in the design process, it is necessary to package the abundance of data on human perceptual, cognitive, and motor phenomena into a set of behavioral templates that can be directly incorporated into predictive, computational models. Templates reduce the amount of psychology and modeling methodology required to build models, compile the human performance data into templates, and allow the modeler to focus on task analysis.

Acknowledgments

This research was supported by funds from the NASA Aviation Operations Safety Program and the Intelligent Systems Program.

References

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates.

Baskin, J. D., & John, B. E. (1998). Comparison of GOMS Analysis Methods. *Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems (Summary)* 1998 v.2 p.261-262.

Berkovich, M., J., & Kwong, E. (2002). Apex template manual, working paper.

Card, S. K., Moran, T.P. & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freed, M. (1998a) Managing multiple tasks in complex, dynamic environments. In Proceedings of 15th National Conference on Artificial Intelligence, (Madison, Wisconsin,) Menlo Park, CA: AAAI Press/ MIT Press. pp. 921-927.

Freed, M. (1998b) *Simulating Human Performance in Complex, Dynamic Environments*. Doctoral Dissertation, Northwestern University.

Freed, Michael and Remington, R. (2000) GOMS, GOMS+ and PDL. *In Working Notes of the AAAI Fall Symposium on Simulating Human Agents*. Falmouth, Massachusetts.

Gray, W. D., John, B. E., & Atwood, M. E. (1993) Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction, 8,* pp. 237-309.

John, B. E. (1990) Extensions of GOMS analyses to expert performance requiring perception of dynamic visual and auditory information. In proceedings of CHI, 1990 (Seattle, Washington, April 30-May 4, 1990) ACM, New York, 107-115.

John, B. E., Vera, A. H., and Newell, A. (1994). Towards real time GOMS: A model of expert behavior in a highly interactive task. *Behaviour and Information Technology,* **13**, 4, pp. 255-267

John, B. E. (1996) TYPIST: A Theory of Performance In Skilled Typing. *Human-Computer Interaction , 11* (4), pp.321-355.

John, B. E. (1998) Cognitive modeling for Human-Computer Interaction. Invited paper in the *Proceedings of Graphics Interface 98* (Vancouver, British Columbia, Canada, June 18-20, 1998) Canadian Human-Computer Communications Society.

John, B. E. & Altmann, E. M. (1999). The power and constraint provided by an integrative cognitive architecture. Invited paper, Proceedings of the 2nd international conference on cognitive science and the 16th annual meeting of the Japanese Cognitive Science Society joint conference (July 27-30, 1999. Tokyo, Japan). pp. 20-25.

John, B. E. & Gray, W. D. *GOMS Analyses for Parallel Activities.* Tutorial materials, presented at CHI, 1992 (Monterey, California, May 3- May 7, 1992), CHI, 1994 (Boston MA, April 24-28, 1994) and CHI, 1995 (Denver CO, May 7-11, 1995) ACM, New York.

John, B. E. & Lallement, Y. (2000) A Demonstration of Integrative Modeling of a Complex Dynamic Computer-based Task. In *Proceedings of the 2000 AAAI Fall Symposium on Simulating Human Agents*, November 3-5, 2000.

John, B. E., Vera, A. H., Matessa, M., Freed, M., & Remington, R. (2002) Automating CPM-GOMS. *Proceedings of CHI*, *2002* (Minneapolis, April 20-25, 2002) ACM, New York.

Kirwan, B. & Ainsworth, L. K. (Eds.) (1992). *A guide to task analysis*. London, UK

Nelson, G. H., Lehman, J. F., & John, B. E. (1994) Integrating cognitive capabilities in a real-time task. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, August 1994. pp. 353-358.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press. Cambridge, Massachusetts.

Pelz, J.B. & Canosa, R. (2001) Oculomotor Behavior and Perceptual Strategies in Complex Tasks, *Vision Research*, 41:3587-3596.

Taatgen, N. A. & Lee, F.J. (submitted). Production Composition: A simple theory of Complex Skill Acquisition.

Collaborative Interactions: The Process of Joint Production and Individual Reuse of Novel Ideas

Mark U. McGregor (markmc@pitt.edu) Michelene T.H. Chi (chi@pitt.edu)
Department of Psychology and the Learning Research and Development Center,
University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260

Abstract

Collaborative problem solving involves the active exchange and interaction of ideas between two or more people and such interactive exchanges can result in the joint production of co-constructed ideas, some of which may be novel. We analyzed verbal data of pairs of students collaboratively solving problems posed by a computer workplace simulation (a banking business), and then individually solving two transfer problems, in order to examine the frequency of occurrence of co-constructed novel ideas, and the subsequent individual reuse of these co-constructed ideas. The results show that in collaborative interactions, about 20% of the task-relevant ideas were produced jointly, whereas about 80% of the utterances were produced individually (i.e., they were self-explanations). However, about half of these jointly produced ideas (or 10%) were novel. Moreover, individual collaborators were able to reuse these jointly constructed ideas to solve transfer problems. Finally, more interactive collaborative pairs produced a higher proportion of jointly constructed ideas than less interactive pairs, and individual members of more interactive pairs reused jointly constructed ideas more than low interactive pairs.

Introduction

In classrooms and workplaces, individuals frequently learn by collaborating with others, in tasks such as solving physics problem (Kneser & Plotzner, 2001), planning (Barron, 2000), and learning electricity (van Boxtel, van der Linden, & Kanselaar, 2000). Although operational definitions of collaboration vary widely both within and across various fields (e.g., Psychology, Education, Artificial Intelligence, CSCL), for the purposes of this paper, we define collaboration as the active exchange and interaction of ideas between two or more individuals attempting to discover solutions or create knowledge together (Damon, 1984). While some of the results of previous collaboration research are inconsistent, the majority support the conclusion that compared to solving a problem alone, collaborative problem solving is often more efficient, and in some conditions, more efficacious than individual learning (SCANS, 1991; Webb & Palinscar, 1996).

Most of the initial research on collaborative learning focused on the environmental conditions under which collaborative learning was more effective than individual learning. Some examples of such environmental factors are group composition, task features, context, and communicative medium (Dillenbourg et al., 1996). However, these mediating factors also interact in a highly complex manner, and this complexity has made the resulting examination of how these multiple interactions produce collaborative learning effects a very difficult pursuit.

In part due to this difficulty, an alternative approach to the study of collaboration focuses on the interactive processes that are thought to underlie successful collaborative learning. Examples of such processes are observing peers' strategies, engaging in productive argumentation, explaining one's own thinking, sharing knowledge, and providing critique (Azmitia, 1988; Bos, 1937; Coleman, 1998; Hatano & Iganaki, 1991; King, 1990; Phelps & Damon, 1989; Webb, Troper, & Fall, 1995). Many of these processes are captured more-or-less in Webb & Palinscar's (1996) 'Input-Process-Output' model of group (collaborative) processes.

The process component of the model contains four common collaborative learning processes: (a) resolving conflict and controversy, (b) giving and receiving explanations, (c) providing emotional and motivational support, and (d) co-constructing new ideas. The first two processes result in the generation of explanations, either to resolve a conflict, or to explain a problem or solution, and such explanation generation is known to produce learning gains (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Palinscar & Brown, 1984). The third process is generally comprised of personality factors (e.g., emotion regulation, motivation, social skills, and attitudes), each of which effects how collaborators interact with one another. The fourth process, co-construction, can potentially result in novel ideas, or ideas that no collaborator previously possessed explicitly.

While Webb & Palinscar's (1996) four collaborative processes do drive collaborative learning effects, the first three processes are not unique to collaboration -- that is, each can also be observed and implemented in individual learning environments. For example, conflict and controversy can arise within oneself when one thinks more deeply about, or attempts to integrate new information with, one's prior beliefs. The resolution of such conflict (or self-repairs) produces learning gains (Chi, 2000). Similarly, giving help or explanation is akin to self-explaining one's own thinking without interacting with another individual, and such self-explanations foster learning (Chi et al., 1989). Similarly, receiving explanations has always been shown to be helpful as well, but less so than giving explanations. Providing emotional and motivational support is also not unique to collaboration – individuals are capable of supporting their

own emotional and motivational needs. In the end, while all four of Webb & Palinscar's (1996) proposed collaborative processes do drive collaborative learning effects, co-construction is the only one that is unique to collaborative learning

Co-construction has been defined as the process of the joint production of ideas (by members of a group) that no individual group member is likely to produce on their own (Barron, 2000; Rafal, 1996). Thus, co-construction is a process that may differentiate collaborative and individual learning environments. That is, collaborators may co-create novel ideas that were not coherently present before they collaborated, or were unlikely to have been elaborated individually. For example, Kneser and Ploetzner (2001) administered qualitative and quantitative pre-test, intermediate-test, and post-test questions to individuals learning mechanics. One group of students was instructed on how to solve mechanics problems qualitatively, while another group was instructed quantitatively. After administering the intermediate-test, individuals from the two contrasting groups were placed in dyads and asked to collaboratively solve new mechanics problems. The authors found that individuals were able to answer questions (post-test) that they were unable to answer prior to collaborating (intermediate-test). However, the results of this study must be interpreted cautiously since the pairs differed in expertise, which suggests that one member of the dyad may have learned from the other member, and not that the dyads co-constructed the answers together.

Hence, the goals of this paper are to document the extent to which collaboration facilitates the joint production of novel ideas, and the extent to which individual members of collaborating pairs take ownership of jointly produced novel ideas such that they are able to assimilate and reuse those ideas later in individual (transfer) situations. To address these questions, we report a detailed verbal analysis of collaborative protocols collected in the context of dyads interacting with a computer simulation.

Methods

The verbal data that we analyzed were collected and described in Jeong, Taylor, and Chi (2000). Here we briefly summarize how the verbal data were collected and the measures that are relevant to our analysis.

Participants

Twenty-six high school students (juniors and seniors from local urban high schools) participated in this study, comprising 13 collaborative pairs (4 male and 9 female same-sex collaborative pairs). Data from one pair was excluded from analysis due to a collection error, resulting in 12 collaborative pairs. Individual participants brought a friend of the same gender to participate in the study with them, resulting in pairs where each collaborator had known the other for a mean of 4 years. All participants were compensated for participation in the study and all reported familiarity with computers. A majority of the students reported that they had used other computer simulations or games in the past.

Materials

Court Square Community Bank (CSCB). CSCB is a computerized workplace simulation (SIM) in which the user assumes the role of a new vice-president at a small local bank. The VP is required to solve problems arising at the bank that cover a variety of general business issues such as facilities upgrades or customer relations. The problems encourage the VP to employ a number of business management activities (see Ferrari, Taylor, and VanLehn, 1999; and McQuaide, Leinhardt, & Stainton, 1999; for more details on the simulation program). We will use the more general term SIM to refer to CSCB.

SIM Problems. The SIM is sub-divided into 14 different episodes, 8 of which were selected for this study so that diverse topics would be covered with a minimum amount of overlap. Sample episode topics were reinvesting profits into bank facilities, closing/relocating branches, approving mortgages, and selecting the best candidate for a position.

Measures. Several measures, such as definitions and general business knowledge questions, were designed to assess participants' overall pre-test and post-test performance. The one relevant here is the transfer task that assessed general and context-specific knowledge at a deeper level than the definitions and questions measures.

The transfer task consisted of two problems (Fresh Foods and Giant Gallery), each modeled after two problems discussed in two SIM episodes (episode 9 and 10). The transfer problems were designed to appear different on the surface, but shared the same deep structure as the associated SIM problems. During the transfer problems, participants acted as vice-president (VP) of a grocery store company that experiences some of the same problems that the bank experienced in the SIM. In the Fresh Foods problem, students decided how to allot available funds to various facilities improvement options for the grocery store – as they did for the bank in SIM episode 9. In the Giant Gallery problem, students decided whether to close a less profitable store and/or open a new store – as they did for bank branches in SIM episode 10.

Talk-aloud protocols were elicited from participants to obtain more detailed performance data. Participants were instructed to talk-aloud (Ericsson & Simon, 1984) during their collaborative (simulation) and individual (transfer) problem-solving sessions, both of which were audiotape. All collaborative SIM sessions were also video-taped

Procedure

Participants took part in four laboratory sessions, each separated by approximately four days, in the following order: (1) pre-test, (2) simulation session I, (3) simulation session II, and (4) post-test.

The pre-test and post-test were administered to participants individually, and both tests asked participants to respond to the definitions, transfer, and questions tasks – and in that order. Each simulation session consisted of four epi-

sodes. The first three episodes in each session were performed collaboratively in dyads, and the last problem in each session was performed individually. On the collaborative problems, the pairs were instructed to work as a team in discussing how to handle the problems and to reach a consensus before making any decisions.

Analyses and Results

We report the results of verbal analyses (Chi, 1997) of individuals' performance on the two transfer problems and their associated SIM episodes 9 and 10. This section describes each step of verbal analysis employed, followed by its result.

The Number of Task-relevant Ideas

Did collaborators produce task-relevant ideas? To assess the number of task-relevant ideas the pairs articulated, each pairs' verbal protocols were coded for any combination of utterances representing a meaningful concept. Hence, ideas could be produced either by one individual in one turn, by one individual over several conversational turns, or by both individuals over several turns on one topic. Ideas were also required to be task-relevant – that is, related either to business in general, the theme of the episode (i.e., facilities improvement), or a solution the pair considered. Task irrelevant ideas consisted of comments of several types, such as "I guess you just gotta ask", "We need to find out some more information", or "Read the memo", and they were not coded.

Example 1 shows an articulated task-relevant idea. Note that examples are verbatim excerpts from protocols with prior utterances (in brackets) added for additional context.

Example 1: Articulated Idea

A: [If we renovate the floors] we will not be able- [to maintain our leadership because…]
B: [subjects speaking simultaneously] new technology.
A: Exactly…because we will not have the new technology.

The basic idea expressed in example 1 is that renovating the floors will consume the majority of the funds available for facilities improvements, resulting in a lack of funds available to purchase new technology (ATMs) that would help the bank maintain its leadership position in the marketplace.

In general, the protocols were sparse in terms of the amount of substantive ideas produced. For example, the total number of lines available to code from two randomly chosen pairs were 2,131 and 1,320 (pairs 8 and 11, respectively, episode 9). Of those lines, only 3% and 1%, respectively, were substantive enough to warrant coding. This is consistent with evidence in the literature showing that in collaborative tasks in which concrete actions have to be taken (such as working with a simulation), the dialogues are action-oriented and less abstract and rich (Bennet & Duane, 1991; Pilkington & Parker Jones, 1996; van Boxtel et al., 2000) Nonetheless, in playing the two episodes of the simulation, collaborative pairs produced a total of 365 task relevant ideas (M=30.42, SD=9.23, per pair). The remaining analyses will be reported in the context of the total number of task relevant ideas.

The Number of Novel and Restated Ideas

Were the task-relevant ideas produced while playing the two simulation episodes novel or restated? To assess whether collaborators generated novel constructions, as opposed to merely restating information presented by the SIM, we compared the ideas produced by collaborators to the ideas explicitly presented by the SIM (SIM-ideas). SIM-ideas were identified through content analysis of the two relevant SIM episodes (9 and 10). This content analysis produced a transcript of the virtual conversations, interactions, and materials explicitly presented by the SIM. SIM-ideas were then identified in the SIM-transcripts following the same procedure employed to identify collaborative ideas. We then compared the collaborative ideas produced by a pair to the SIM-ideas that they were exposed to in order to determine whether the collaboration idea was a novel construction (no match), or a restatement of information embedded within the simulation (match). When comparing ideas we focused on their conceptual meaning rather than the literal vocabulary used.

Ideas coded as novel constructions were new reasons generated by students (most likely from prior knowledge or experience), inferences following from what was stated in the SIM episode, substantial paraphrases (paraphrases of more than one idea), and integration statements in which students combined ideas expressed in the SIM episode in ways that the SIM did not explicitly suggest. Restated ideas were unsubstantial paraphrases (one idea), or verbatim restatements of information presented by the SIM. Examples of a novel construction and a restated idea are given below:

Example 2: A Novel Construction (also jointly produced)

B: This proposal [New ATMs] helps the bank's profitability
A: help
B: helping the bank to run
A: more smoothly

Example 2 was coded as a novel construction because the SIM never explicitly relates the bank's profitability with new ATMs or with the bank running smoothly. (Note that example 2 is also a jointly produced idea, which will be described in the next section).

Example 3: A Restated Idea

B: [New ATM] cards will be easier to use
SIM: New ATMs are easier for the customer to use.

Example 3 was coded as a restated idea because what B articulates is an unsubstantial paraphrase of the SIM explicated idea that new ATMs are easier to use. The results of

this analysis show that collaborators produced just as many novel ideas (197) as restated ideas (168).

Individually and Jointly Constructed Ideas

Given that many novel ideas were produced, the question of interest is whether they were jointly produced (i.e., co-constructed), or individually produced, as compared to restated ideas. Thus, for each of the 197 novel and 168 restated ideas we determined whether they were individually constructed or jointly co-constructed. Co-constructed ideas were defined as those ideas that when taken together, across speakers, form a complete idea, but when taken individually, do not represent the same complete idea (Rafal, 1996; Barron, 2000). Thus, we decomposed each idea unit into 3 component parts: initiation, elaboration, and completion. Initiation was defined as the point at which the first utterance (word) of the idea occurred; completion was defined as the point at which a meaningful statement could be identified; and elaboration was defined as the collection of utterances between initiation and completion, where the content of the idea was articulated. We then examined how each idea was produced in terms of who articulated each component.

For each idea, if the same collaborator produced all three components, then the idea was coded as individually produced. Alternatively, if different collaborators produced any of the components of one idea, then that idea was coded as jointly produced. Hence, jointly produced ideas required the collaborators to display conversational moves such that they completed each other's ideas. The following examples illustrate this.

Example 4: An Individually Produced Idea

A: revenues and expenses at the downtown branch changed
B: Uhh….
A: revenues have just start like increase and decrease and then leveled off so…
B: Umm…yeah… [typing] how do you spell fluctuate
A: fluctuating, but now it's leveled off and…
B: How do you spell…
A: well, they generally decreased

Example 4 shows a complete idea (revenues and expenses fluctuated but decreased in general) initiated, elaborated, and completed by one collaborator (A), while the other (B) interjects task-irrelevant utterances. The initiation component in this case is the beginning of the statement: "revenues…" the completion component is the point at which a complete idea is identifiable: "…generally decreased"; while the elaboration component is the content between the initiation and the completion of the idea.

Example 5: A Jointly Produced Idea:

B: Okay, the new system would give the- give the employees…
A: more time to deal with the customers.

Example 5 shows an idea (of a new system) initiated by one collaborator (B), then elaborated and completed by the other collaborator (A). Example 2 also illustrates a jointly-produced idea.

Table 1 shows the number of novel and restated ideas that were either individually or jointly produced. Not surprisingly, roughly four times (81%) as many ideas were individually (297) rather than jointly (19% or 68) produced. Proportionately, jointly produced ideas were just as likely to be novel (59%, 40/68) as restated (41%, 28/68).

Table 1: Total Number of Jointly and Individually Produced, Novel and Restated Ideas for All Collaborative Pairs

	Novel	Restated	
Individually Produced	157	140	297
Jointly Produced	40	28	68
	197	168	365

In sum, collaborative pairs produced ideas jointly about 20% (or 68) of the time; individually about 80% (or 297) of the time. Given that jointly produced ideas were equally likely to be novel as restated, about 10% (or 40/365) of the ideas were jointly produced novel ones.

Reuse of Ideas During Transfer

Did individuals reuse the jointly produced novel ideas on transfer problems? In other words, were the individual collaborators able to reuse the jointly produced novel ideas, to indicate that they have, to some extent, taken ownership of or assimilated the ideas? Each individual idea stated while solving a transfer problem was compared to each idea originally produced by the pair when they solved the associated SIM episodes. If an idea articulated while solving the transfer problem matched one that was produced while playing the SIM, then the idea was coded as a reused idea.

In order to make more sensitive comparisons between collaborative pairs during the SIM and individuals at transfer, each individual's transfer performance was averaged with the individual transfer performance of the other member of their original collaborative pair. This averaging procedure resulted in equal *n* in each condition (collaboration vs. transfer).

In general, the transfer transcripts were sparser than the collaborative transcripts in terms of the total number of ideas produced (114; M=9.5, SD=3.75). Of these, 32% (36/114) were ideas that were originally produced during their collaborative session. Overall, individual collaborators reused more ideas that were originally produced individually (25) than ideas that were produced jointly (11). However, recall that a significantly greater number of ideas were originally produced individually (297, see Table 1), rather

than jointly (68). Thus, proportionately, a larger percentage of jointly produced ideas were reused (16% or 11/68) than individually produced (8%, or 25/276), although this difference is not significant. Basically, jointly and individually produced ideas were equally likely to be reused.

Not surprisingly, of the individually produced ideas, participants tended to reuse those that they generated on their own (64% of the times, 16/25) more than those that were generated by their partner (36% of the time, 9/25; t (11) = 2.86, $p < .05$). In contrast, in the reuse of jointly produced ideas, there was no preference for self-initiated or partner-initiated (59% or 6.5/11 versus 41%, 4.5/11). Taken together, these results suggest that jointly produced novel ideas were equally shared by each partner, regardless of who initiated them, whereas individually produced ideas were not as well assimilated by the partner.

In sum, these results show that about one-third (32%) of the ideas individuals stated while solving the transfer problems were originally produced during collaboration, and these reused ideas were equally likely to have been individually produced as they were to have been jointly produced. However, collaborators had a definite preference to reuse self-initiated, individually produced ideas, but had no such preference when reusing jointly produced ideas. This gives the jointly produced ideas a special status, as if the ideas were truly shared and owned by both partners.

High and Low Collaborative Pairs

We hypothesized that pairs who were more interactive would produce more co-constructed ideas. To test this hypothesis, we determined whether specific pairs were more or less collaborative based on the number of conversational turns taken by each pair while they solved SIM episodes 9 and 10. High and low collaborative groups were then formed based on a median split (Mdn = 933 turns), and excluded two pairs extremely close to the median. The mean number of turns for high and low groups was significantly different (1186.4 vs. 763.8 turns, respectively; t (8) = 5.171, p < .01,), and there was no significant difference in the total number of ideas produced overall. (Note that from this point forward high and low collaborative pairs will be referred to as such, while individual members of high and low collaborative pairs will be referred to as high and low collaborators).

Once the groups were established, we compared the proportion of ideas produced jointly versus individually by high and low collaborative pairs. For the high collaborative pairs, the proportion of ideas that were jointly produced was greater (27% vs. 9%; t (8) = 3.77, p < .01); while the proportion of individually produced ideas was lower (73% vs. 91%, t (8) =3.82, p < .01). Additionally, the proportion of jointly produced and novel ideas was also greater for high collaborative pairs (17% vs. 6%); (t (8) =3.10, p < .05).

In sum, being a member of a high collaborative pair resulted in a redistribution of the types of ideas produced during collaboration. That is, high and low collaborative pairs produced equal numbers of ideas overall, but high collaborative pairs produced more ideas jointly and fewer ideas individually than low collaborative pairs; and high collaborative pairs produced a greater proportion of novel co-constructed ideas than low collaborative pairs

High and Low Individuals' Reuse of Ideas

While solving the transfer problems alone, individual members of high and low collaborative pairs produced roughly equivalent total numbers of ideas (52 vs. 46.5, respectively), as well as roughly equivalent numbers of reused ideas (19 vs. 13; see Table 2). However, the types of reused ideas were again differentially distributed - high collaborators reused more jointly produced ideas than low collaborators (18% vs. 8%, 8.5/47 vs. 1/12**;** t (8) = 2.434, $p < .05$), and high collaborators were more likely to reuse jointly produced novel ideas than low collaborators (20% vs. 11%, 5.5/28 vs. 1/9), although this difference was not significant ($p = 0.10$).

In sum, individual members of high collaborative pairs reused more jointly produced ideas, and had a greater tendency to reuse jointly produced novel ideas. Taken together, this pattern of results suggests that if collaborators engage in more interaction then they are more likely to produce co-constructed ideas, these co-constructed ideas are likely to be novel, and both the co-constructed and novel co-constructed ideas are likely to be reused. Hence, collaboration has the advantage of producing co-constructed ideas and co-constructed novel ideas that are reusable.

Table 2: Number of Ideas Reused by Individuals Who Participated in High and Low Collaborative Pairs

	High	Low
Ideas at Transfer	52	46.5
Ideas Reused	19 (36%) [1]	13 (28%) [1]
Individually Produced	10.5 (8%) [2]	12 (9%) [2]
Jointly Produced	8.5 (18%) [2]	1 (8%) [2]
Novel Ideas	13.5 (13.5%) [2]	7.5 (10%) [2]

[1] *% of ideas produced at transfer that were reused (e.g., 19/52 = 36%).*
[2] *% of ideas reused at transfer given the number produced during collaboration (see Table 1); (e.g., Reused Individually Produced = 10.5; Individually Produced during collaboration = 128; 10.5/128 = 8%).*

Discussion

These results suggest that one advantage of collaboration may arise from the co-construction of novel ideas. Overall, we found that collaborators tended to produce more ideas individually than jointly, confirming the overall benefit of self-explaining (Chi, et al, 2000). However, joint production did occur close to 20 percent of the time, and jointly pro-

duced ideas were just as likely to be novel as restated. Overall, more novel ideas were produced individually than jointly (again, confirming the benefit of self-explaining). However, individuals reused fewer novel ideas that were individually produced by their partner than were jointly produced with their partner, suggesting that listening to novel ideas produced by another was not as effective as co-constructing novel ideas together. These results suggest that one does not assimilate knowledge produced by a partner as well as knowledge co-constructed by both partners. Finally, the more interactive collaborators reused a greater percentage of jointly produced ideas, as well as a greater percentage of jointly produced novel ideas, thus being more interactive provided more opportunities to co-construct and reuse novel ideas. Thus, we may conclude, (cautiously since the numbers are small), that collaboration is an effective form of learning in part because about 10% of collaborative efforts result in the production of co-constructed novel ideas, a portion of which individuals take ownership of and reuse subsequently.

References

Azmitia, M. (1988). Peer interaction and problem solving: When are two heads better than one? *Child Development, 59,* 87-96.

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *Journal of Learning Sciences, 9 (4),* 403-436.

Bennett, N. & Dunne, E. (1991). the nature and quality of talk in co-operative classroom groups. Learning and Instruction, 1, 103-118.

Bos, M.C. (1937). Experimental study of productive collaboration. *Acta Psychologica, 3,* 315-426.

Chi, M.T.H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M.T.H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In Glaser, R. (Ed.). *Advances in Instructional Psychology*. (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.

Coleman, E. (1998). Using explanatory knowledge during collaborative problem solving in science. *Journal of Learning Sciences, 7,* 387-427.

Damon, W. (1984). Peer education: the untapped potential. *Journal of Applied Developmental Psychology, 5,* 331-343.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In P.Reimann & H. Spada, *Learning in Human and Machines: Towards an Interdisciplinary Learning Science.* (pp. 189-211). Oxford: Elsevier Science.

Ferrari, M., Taylor, R. & VanLehn, K. (1999). Adapting work simulations for schools. *Journal of Educational Computing Research, 21 (1),* 25-53.

Hatano, G. & Inagaki, K. (1991). Sharing cognition through a collective comprehension activity. In R.L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 331-348). Washington, DC: American Psychological Association.

Jeong, H., Taylor, R., & Chi, M. (2000). Learning from a Computer Workplace Simulation. *Proceedings of the 22nd annual meeting of the Cognitive Science Society*, (pp 705-710). Mahwah, NJ: Lawrence Erlbaum Associates

King, A. (1990). Facilitating elaborative learning in the classroom through reciprocal questioning. *American Educational Research Journal, 27*, 664-687.

Kneser, C. & Ploetzner, R. (2001). Collaboration on the basis of complementary domain knowledge: Observed dialogue structures and their relation to learning success. *Learning and Instruction, 11,* 53-83.

McKendree, J., Stenning, K., Mayes, T., Lee, J., & Cox, R. (1998). Why observing a dialogue may benefit learning: The vicarious learner. *Journal of Computer Assisted Learning*, *14 (2)*.

McQuaide, J., Leinhardt, G., & Stainton, C. (1999). Ethical reasoning: Real and simulated. *Journal of Educational Computing Research, 21 (4),* 425-466.

Phelps, E., & Damon, W. (1989). Problem solving with equals: Peer collaboration as a context for learning mathematics and spatial concepts. *Journal of Educational Psychology, 81, 639*-646.

Pilkington, R.M. & Parker-Jones, C.H. (1996). Interacting with computer-based simulation: The role of dialogue. Computers and Education, 27, 1-14.

Rafal, C. (1996). From co-construction to takeovers: Science talk in a group of four girls. *Journal of Learning Sciences, 5,* 279-293.

Secretary's Commission on Achieving Necessary Skills. (1991). What *work requires of schools: A SCANS report for America 2000.* Washington, DC: U.S. Department of Labor.

van Boxtel, C., van der Linden, J., & Kanselaar, G. (2000). collaborative learning tasks and the elaboration of conceptual knowledge. Learning and Instruction, 10, 311-330.

Webb, N.M., & Palinscar, A.S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology (pp. 841-873).* New York: Macmillan.

Webb, N.M., Troper, J.D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87,* 406-423.

A Strong Schema Can Interfere with Learning: The Case of Children's Typical Addition Schema

Nicole M. McNeil (nmmcneil@students.wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Martha W. Alibali (mwalibali@facstaff.wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

This study investigated whether children's schema for typical addition interferes with their ability to learn about mathematical equivalence. In a pretest, elementary school children (1) solved a set of math equivalence problems (e.g., 3 + 4 + 5 = 3 + __), (2) reconstructed equivalence problems after viewing them briefly, and (3) provided definitions of the equal sign. Children were categorized according to the number of measures (out of 3) on which they exhibited the typical addition schema. Children then received one of four interventions that presented new information about equivalence problems. Finally, children completed a posttest similar to the pretest. From pretest to posttest, children who exhibited the addition schema on all three measures were the least likely to change their strategy for solving the problems, followed by children who exhibited the schema on two or one of the measures. All of the children who did not exhibit the schema on any of the three measures changed. It is important to note that all children used incorrect strategies at pretest, so it was the addition schema in particular that was associated with change resistance. Thus, a strong schema can interfere with learning. Furthermore, children's addition schema may put them at risk for difficulties in learning higher-level mathematics.

Some behaviors and ideas are particularly difficult for individuals to acquire, even after numerous attempts have been made by the individuals to learn them or by instructors to teach them. Consider, for example, a woman who has always counted on her fingers when solving addition problems. She does not wish to count on her fingers; in fact, she finds it to be an infuriating habit. However, no matter how hard she tries to memorize the addition facts, she always seems to resort to counting on her fingers. Consider also a father who is trying to learn a second language as an adult. The father is frustrated because his daughter, who is only five years old, is learning so much more quickly than he is. Why is it so difficult for the woman to memorize the addition facts and for the father to learn a second language? One reason may have to do with the organization and strength of the behaviors and ideas they already possess. The woman's knowledge about addition may be organized into a strong schema that interferes with her memorizing of the addition facts, and the father's knowledge about his first language may be organized into a strong schema that interferes with his learning of the second language.

A schema can be defined as a higher-level assembly of knowledge that is "unitized" in that it supersedes its constituent parts and acts as a whole (cf. Smolensky, 1986; see also Hayes-Roth, 1977; Hebb, 1949). Once established, schemata can serve as selective mechanisms that determine how environmental stimuli are encoded, interpreted, and stored in memory. In general, schemata enable fast and efficient processing of environmental stimuli. However, this efficiency can come with a price, especially if a particular schema is inaccurate or lacking in some way.

When a schema is strong, it resists change (e.g., Allport, 1954; Bartlett, 1932; Bruner, 1957; Schutzwohl, 1998). Individuals who have a strong schema have been shown to resist learning new information when it is more specific than (Adelson, 1984; Thorndyke & Hayes-Roth, 1979), not applicable to (Voss, Vesonder, & Spilich, 1980), or discrepant with (Marchant, Robinson, Anderson, & Schadewald, 1991; Markus, 1977) their current schema. Indeed, individuals who have a strong schema often actively resist change by modifying or distorting environmental input so that it corresponds to their schema (Bartlett, 1932; Bruner & Postman, 1949; Guion, Flege, Akahane-Yamada, & Pruitt, 2000; Hannigan & Reinitz, 2001). Importantly, the stronger a schema is, the more resistant it is to changing (see Luchins, 1932).

Although the strength of a schema has clear behavioral consequences, it is not always obvious how to operationalize schema strength apriori. Some theories suggest that strength may be determined by how practiced particular action procedures are (Luchins, 1942), while others suggest that it may be derived from tightly organized perceptual (Flege, Bohn, & Jang, 1997; Intraub & Bodamer, 1993) or conceptual (Wellman & Gelman, 1992) information. These accounts need not oppose one another, though, because any given schema is likely constructed out of various related sub-schemata (Smolensky, 1986). We propose that schema strength depends on the relationship

between the three knowledge sources (action procedures, perceptual information, and conceptual information). According to this view, the strongest schema is one in which all three knowledge sources converge.

In the current study, we focused on a particular schema that most elementary school children possess. We refer to the schema as the *typical addition schema*. This schema includes action procedures for solving typical addition problems (e.g., 4 + 5 = __, 9 + 2 + 6 = __), such as counting up all of the numbers in the problem or summing all of the numbers by retrieving addition facts from memory. It also includes particular perceptual patterns, such as the pattern of having "= __" at the end of problem (Baroody & Ginsburg, 1983). The schema also includes a conceptualization of the equal sign as an operational symbol that means "the total" (Kieran, 1981; Rittle-Johnson & Alibali, 1999). When an addition problem is presented, the typical addition schema can be activated as a whole, enabling fast and accurate processing of the problem (McNeil, 2001).

The typical addition schema is adaptive because children need to use their knowledge of addition to learn other math concepts, such as multiplication and division. However, it may come with a price. Specifically, it may interfere with children's ability to learn about novel mathematics problems, such as mathematical equivalence problems, which are problems that have addends on both sides of the equal sign (e.g., 3 + 4 + 5 = 3 + __; Perry, Church & Goldin-Meadow, 1988).

Past work has shown that some children have an especially strong typical addition schema that leads them to distort environmental input about mathematical equivalence to correspond to their schema. Consider the problem 3 + 4 + 5 = 3 + __. In solving this problem, some children add up all the numbers and put 15 in the blank (McNeil & Alibali, 2000). In reconstructing the problem after viewing it briefly, some children reconstruct it in terms of the "= __ at end" perceptual pattern and write "3 + 4 + 5 + 3 = __" (McNeil, 2001). In defining the equal sign in the problem, some children say that it means to "add up all the numbers" (McNeil & Alibali, 2002).

Once children gain more experience with or are instructed about mathematical equivalence problems, they are bound to generate new ways of thinking about the problems. However, these new ways of thinking may be in conflict with their already established typical addition schema, and the two may compete for precedence (cf. Siegler, 1999). Accordingly, we hypothesize that the strength of the typical addition schema should be directly related to children's tendency to resist changes in their thinking after an intervention that provides new information about equivalence problems.

Method

Participants

The sample consisted of 67 third-, fourth-, and fifth-grade children (29 boys and 38 girls), all of whom solved a set of mathematical equivalence problems on the experimental pretest incorrectly. Children attended public or parochial schools in the greater Madison, Wisconsin area.

Measures

Problem Solving The problem solving measure elicited action procedures for solving the problems. It consisted of three mathematical equivalence problems of the form a + b + c = a + __. For each problem the experimenter placed the problem on an easel and said, "Try to solve the problem as best as you can and then put your answer in the blank." After children wrote a solution, the experimenter said, "Can you tell me how you got x?" After explaining each solution, children were asked to rate how certain they were about their "way of doing" the problem on a 7-point scale that ranged from "It's definitely wrong" to "It's definitely right," with "I'm not sure if it's wrong or right" as the midpoint.

Problem Reconstruction The problem reconstruction measure elicited perceptual representations of the problems. Two tasks made up the measure. The first was taken from Rittle-Johnson and Alibali (1999). Children were asked to reconstruct three equivalence problems of the form a + b + c = a + __ after viewing each for five seconds. The second task also included three problems and was a recognition version of the first task. Children were given a sheet of paper face down with seven problems on it. One of the problems was an equivalence problem in its correct form. The other six problems depicted errors children typically make when reconstructing equivalence problems, one of which was the typical addition foil a + b + c + a = __. After viewing an equivalence problem for five seconds, children were instructed to turn the sheet of paper over and find the problem that they just saw.

Equal Sign Definition The equal sign definition measure was used to elicit conceptual understanding of the equal sign. Two tasks made up the measure. Both were taken from Rittle-Johnson and Alibali (1999). Children were first asked to define the equal sign. Then, they were asked to rate the smartness of six fictitious students' definitions as not so smart, kind of smart, or very smart. The definitions were "the answer to the problem," "repeat the numbers," "the end of the problem," "something is equal to another thing," "two amounts are the same," and "the total."

Procedure

Children participated individually in one experimental session that was videotaped. In the pretest, children first completed the problem-solving measure, followed by the problem reconstruction and equal sign definition measures presented in random order. After the pretest, children were randomly assigned to intervention conditions in a 2 (reconstruction intervention or no reconstruction intervention) x 2 (equal sign definition intervention or no equal sign definition intervention) factorial design. During the intervention, children in all four conditions were presented with an equivalence problem that had the correct solution written in the blank (3 + 4 + 5 = 3 + 9). In the control condition, children were shown the correctly solved problem and were told that it was a correctly solved problem. They were then encouraged to think about the problem for one minute. Children in the other three conditions also were presented with the correctly solved problem, were told that it was a correctly solved problem, and were encouraged to think about it. In addition, children who received the reconstruction intervention were encouraged to notice the equal sign in the problem and were asked to point to it. Children who received the equal sign definition intervention were told that the equal sign means "that the things on one side of it have to be the same as the things on the other side of it" and were asked to repeat the definition. All children spent a total of one minute in the intervention. After the intervention, children participated in a posttest in which they first completed the problem reconstruction and equal sign definition measures in random order, followed by the problem-solving measure.

Coding

Problem Solving Problem-solving strategies were coded using a system developed by Perry, Church and Goldin-Meadow (1988). Strategies were assigned based on children's problem solutions and verbal explanations.

Problem Reconstruction Each reconstruction was examined for conceptual errors. Conceptual errors were errors that reflected inaccurate reconstructions of the structure of the equation, such as omitting the equal sign or one of the plus signs. Errors in reconstructing the particular numbers or order of the numbers were not counted as conceptual errors (e.g., for the problem 3 + 4 + 5 = 3 + __, writing 4 + 3 + 5 = 3 + __). Each recognition response was scored as correct or incorrect based on whether children correctly identified the equivalence problem on the sheet provided.

Equal Sign Definition Definitions were coded as expressing the concept of equivalence or not. None of the children gave a definition that expressed the concept of equivalence on the pretest. Children's ratings of each of the fictitious student's definitions of the equal sign were coded. Two points were given for "very smart" ratings, one point was given for "kind of smart" ratings, and zero points were given for "not so smart" ratings. The sum of the ratings for the two definitions "the total" and "the answer to the problem" were subtracted from the sum of the ratings for the two definitions "two amounts are the same" and "something is equal to another thing" to yield a difference score. A positive difference score indicates that definitions expressing the concept of equivalence were rated as smarter than definitions such as "the answer" and "the total."

Typical Addition Schema Children were categorized according to whether they exhibited the typical addition schema on the pretest measures. They were coded as exhibiting the schema on the problem-solving measure if they (1) used the "add-all-the-numbers" strategy on at least two of three equivalence problems (as shown in Table 1) and (2) gave that strategy an average certainty rating greater than four (on the 7-point scale). Recall that ratings of less than four indicate children think their strategy is incorrect. Children who use the add-all-the-numbers strategy but rate it as incorrect are likely using the strategy because they cannot come up with any alternatives (see Siegler, 1983), rather than because they are operating according to a strong schema per se.

Children were coded as exhibiting the schema on the problem reconstruction measure if they showed evidence of converting at least two problems to typical addition problems (either on the reconstruction task or on the recognition task, as shown in Table 1).

Children were coded as exhibiting the schema on the equal sign definition measure if they showed evidence of thinking that the equal sign means “the sum” or “the total.” Children could show this in one of two ways. They could express the idea of adding or totaling in the definition they provided (as shown in Table 1). Or, they could rate the definition "the total" as "very smart."

Children were categorized according to the number of pretest measures (out of three) on which they exhibited the typical addition schema. Thus, children were placed into an overall typical addition schema category of 0, 1, 2, or 3. The number of measures on which the schema was exhibited was considered to be a reflection of schema strength.

Table 1 presents examples of schema-based and non-schema-based responses on each measure. Notice that children’s responses are incorrect whether they exhibit the typical addition schema or not. Moreover, there is no reason to believe that the thinking of children without the typical addition schema is "closer to correct" than is the thinking of children with the schema. For example, defining the equal sign as "the answer" is just as incorrect, if not more so, than defining it as "the total." Thus, outside of the present framework, there is no reason to expect learning differences between children who do or do not exhibit the typical addition schema.

Table 1: Example Schema-based (SB) and Non-schema-based (NSB) Responses for the Problem 3+ 4 + 5 = 3 + __.

	Strategy (Solution || Explanation)	Reconstruction	Equal Sign Definition
SB	15 || "I added 3 plus 4 plus 5 plus 3."	3 + 4 + 5 + 3 = __	"Add up all the numbers together."
	14 || "I added them all up."	3 + 4 + 5 + 3	"The total of the problem."
NSB	4 || "4 comes after 3 in the pattern."	3 + 4 + 5 = 3 =	"Put your answer."
	24 || "3 and 4 and 5, times 2 is 24."	3 + 4 + 5 = + 3 __	"It's like where you end the problem."

Results

Manipulation Check

We examined pretest to posttest changes in children's performance on the problem reconstruction and equal sign definition measures as a check on whether our interventions provided children with new ways of thinking about the equivalence problems, as they were designed to do. A 2 (problem reconstruction intervention or no problem reconstruction intervention) x 2 (equal sign definition intervention or no equal sign definition intervention) ANOVA was performed with pretest to posttest change in number correct on the reconstruction measure (out of 6) as the dependent variable. As expected, the analysis revealed a significant main effect for problem reconstruction intervention, $F(1, 63) = 4.35$. Children who received the problem reconstruction intervention improved their performance on the reconstruction measure from pretest to posttest ($M = +1.78$, $SD = 1.31$) more so than did children who did not receive the intervention ($M = +1.00$, $SD = 1.65$). Neither the main effect for equal sign definition intervention nor the interaction was significant (both $Fs < 1$).

A similar 2 X 2 ANOVA was performed with pretest to posttest change in difference score on the ratings portion of the equal sign definition measure as the dependent variable. Recall that a positive difference score indicates that definitions such as "the same as" and "equal to" were rated as smarter than definitions such as "the answer" and "the total." As expected, the analysis revealed a significant main effect for equal sign definition intervention, $F(1, 63) = 25.62$. Children who received the equal sign definition intervention improved their difference score from pretest to posttest ($M = +1.94$, $SD = 2.0$) more so than did children who did not receive the equal sign definition intervention ($M = -.12$, $SD = 1.23$). Neither the main effect for reconstruction intervention nor the interaction was significant ($F < 1$ for both).

The preceding analyses indicate that the interventions provided new information about the equivalence problems and that children were, in general, able to take in the presented information in its specific form. The main question at hand is how this new information affected the way children solved the equivalence problems. The interventions themselves did not predict pretest to posttest changes in problem solving, $\chi 2$ (3, N = 67) = 3.89. This is not surprising given that we predicted that children would be differentially affected by an intervention depending on the strength of their typical addition schema.

Effects of Addition Schema

Recall that all children solved the problems incorrectly at pretest. Similarly, only three of the children responded correctly to all six problems on the reconstruction measure at pretest, and none of the children provided an equal sign definition that expressed the concept of equivalence at pretest. It is also important to note that children's pretest schema category (0 to 3) was independent of whether they participated in the control intervention or in one of the experimental interventions, $\chi 2$ (3, N= 67) = 3.89. Our main question was whether the strength of the typical addition schema influenced children's tendency to resist changes in the way they solved the equivalence problems after an intervention that provided new ways of thinking about the problems.

Table 2: Number of children in each typical addition schema category who changed or did not change their problem-solving strategy from pretest to posttest.

Number of Pretest Measures Reflecting Addition Schema	Change	No Change
0	5 (100%)	0
1	14 (64%)	8
2	13 (42%)	18
3	1 (11%)	8
$\chi 2$ (3, N = 67) = 12.88		

Children were classified as changing their problem-solving strategy if they solved any of the three, posttest problems using a different strategy than they used to solve the pretest problems. Table 2 displays the number

of children who changed or did not change their strategy from pretest to posttest in each of the typical addition schema categories. As shown in the table, children who exhibited the addition schema on all three measures were highly unlikely to change their strategy from pretest to posttest. Again, all children used an incorrect strategy at pretest, so it was the typical addition schema in particular that was associated with change resistance. All of the children who did not exhibit the schema on any of the measures changed their strategy from pretest to posttest. Such a high proportion of change is surprising, given that the intervention lasted only one minute.

Although we were primarily interested in how a strong schema influences change after an intervention, we were also curious about the correctness of children's strategies at posttest. Results were similar when correctness of posttest strategy was used in the analysis in place of pretest to posttest strategy change. Children were classified as having a correct problem-solving strategy if they solved any of the three, posttest problems using a correct strategy. For example, children would be classified as having a correct strategy for the problem 3 + 4 + 5 = 3 + __ if they put the solution "9" in the blank and said that they added 4 plus 5 to get 9. Of the 9 children who exhibited the schema on all three measures, only 1 (11%) used a correct strategy on the posttest. Of the 31 who exhibited it on two measures, 10 (32%) used a correct strategy on the posttest. Of the 22 who exhibited it on one measure, 8 (36%) used a correct strategy on the posttest. All 5 (100%) who did not exhibit the schema on any of the measures used a correct strategy on the posttest. The analysis revealed a significant relationship between the strength of children's addition schema at pretest and whether or not they used a correct strategy on the posttest, $\chi 2$ (3, N = 67) = 11.52.

Discussion

The results of the current study indicate that the strength of the typical addition schema can interfere with children's ability to change their way of solving mathematical equivalence problems after an intervention that provided new ways of thinking about the problems. Moreover, children who did not exhibit the typical addition schema did not merely change their strategy for solving the problems after an intervention, but actually changed to using a correct strategy.

These results complement previous work that has suggested that strong schemata resist change (e.g., Allport, 1954; Bruner, 1957; Luchins, 1932; Schutzwohl, 1998). The present study indicates that individuals who have a strong schema resist changes in their thinking even after an intervention that supplies new ways of thinking. This finding may provide a potential avenue for investigation into individual differences in learning. When a group of students is presented with a particular instruction, why do some fail to learn or change, while others succeed? Although it is only speculation at this point, it may be the case that individuals who have difficulty learning new ideas are less flexible and more resistant to change because they develop strong, inaccurate schemata more readily than do individuals who are precocious learners.

The present study also extends previous research about schemata by introducing a new way of operationalizing schema strength. Some accounts have defined schema-like structures according to well-practiced action patterns (e.g., Luchins, 1942), while others have emphasized perceptual (e.g., Guion et al., 2000) or conceptual (e.g., Hannigan & Reinitz, 2001) information. In the present study, children exhibited the typical addition schema in their action procedures (i.e., problem-solving strategies), perceptual encodings (i.e., problem reconstructions) and conceptual knowledge (i.e., equal sign definitions). Results provide support for the notion that the strongest, most change-resistant schemata are ones in which all three of the knowledge sources converge on the same idea. Thus, if educators wish to build strong, accurate schemata in their students, they should not focus on building up one aspect of knowledge at the expense of the other two.

The current research has additional implications for educators. Specifically, findings suggest that a strong typical addition schema carries a heavy price and may put children at risk for difficulties in later years when algebraic equations become the focus of mathematics instruction. Thus, educators may want to consider expanding and varying the context in which they present the operation of addition and the equal sign so that children are less likely to form an inaccurate schema from their experience. More generally, results suggest that children's existing knowledge can interfere with the ability to learn new information. Thus, educators should be cautious about what they infer about children's abilities based on proficiency with today's topic.

Acknowledgments

This work was supported by National Science Foundation Child Learning and Development grant #BCS-0096129 to Martha Alibali. We thank members of the Cognitive Development Research Group at the University of Wisconsin for helpful discussions, Jerry Haeffel for comments on a previous version of this paper, and Maureen Kaschak for help with data collection. We also thank the students, parents, educators at Frank Allis, Van Hise, Marquette, Sacred Hearts, Edgewood, and Our Redeemer schools. We would especially like to thank Mrs. Burwell's third-grade class at Sacred Hearts school for their generosity and enthusiasm. A preliminary report based on a subset of the data described herein was presented at the October 2001 meeting of the Cognitive Development Society, Virginia Beach, VA.

References

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10*, 483-495.

Allport, G. W. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Baroody, A. & Ginsburg, H. (1983). The effects of instruction on children's understanding of the "equals" sign. *Elementary School Journal,* 84, 199-212.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, England: Cambridge University Press.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*, 123-152.

Bruner, J. S., & Postman, L. J. (1949). On the perception of incongruity: A paradigm. *Journal of Personality, 18*, 206-223.

Flege, J., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 437-470.

Guion, S. G., Flege, J., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107,* 2712-2724.

Hannigan, S. L., & Reinitz, M. T. (2001). A demonstration and comparison of two types of inference-based memory errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 931-940.

Hayes-Roth, B. (1977). Evolution of cognitive structures and processes. *Psychological Review, 84,* 260-278.

Hebb, D. O. (1949). *Organization of behavior.* New York: Wiley.

Intraub, H. & Bodamer, J. L. (1993). Boundary Extension: Fundamental Aspect of Pictorial Representation or Encoding Artifact*? Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1387-1397.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12,* 317-326.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs, 54*, (6, Whole No. 248).

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 2*, 63-78.

Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1991). Analogical transfer and expertise in legal reasoning. *Organizational Behavior & Human Decision Processes, 48*, 272-290.

McNeil, N. M. (2001). Mental sets and flexibility in the development of mathematical skill. Poster presented at the Biennial Meeting of the Society for Research in Child Development, Minneapolis, MN.

McNeil, N. M., & Alibali, M. W. (2000). Learning mathematics from procedural instruction: Goals influence learning from the "outside in." *Journal of Educational Psychology, 92*, 734-744.

McNeil, N. M. & Alibali, M. W. (2002). Charting the path of conceptual development: Not all contexts are created equal. Manuscript in preparation.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359-400.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*, 175-189.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Science, 3*, 430-435.

Siegler, R. S. (1993). How knowledge influences learning. *American Scientist, 71*, 631-638.

Schutzwohl, A. (1998). Surprise and schema strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1182-1199.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), Parallel Distributed Processing. Volume 1: Foundations (pp. 194-281). Cambridge, MA: MIT Press.

Thorndyke, P. W., & Hayes-Roth, B. (1979). The use of schemata in the acquisition and transfer of knowledge. *Cognitive Psychology, 11*, 82-106.

Voss, J. F., Vesonder, G., & Spilich, H. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior, 19*, 651-667.

Changes in Learners' Exploratory Behavior in a Simulated Psychology Laboratory

Kazuhisa Miwa, Norio Ishii, Hitomi Saito, and Ryuichi Nakaike
{miwa, ishii, hitomi, nakaike}@cog.human.nagoya-u.ac.jp
Graduate School of Human Informatics, Nagoya University
Nagoya, 464-8601 JAPAN

Abstract

We constructed a virtual psychology laboratory (called VPL) on a computer. VPL simulates the process of pair subjects collaboratively solving Wason's 2-4-6 task, which has been traditionally used in the field of the psychology of discovery science. Participants were required to study collaborative problem solving while repeating experiments and hypothesis revisions using VPL. We conducted three experimental sessions using VPL. As a result, we confirmed, across the sessions, the improvement in various types of participant' performance, such as the organizational construction of experimental design, the degree of correctness of hypotheses the participants formed, and the generality of findings they discovered.

1. Introduction

It is one of the most important objectives in scientific research to understand the behavior of complex systems, such as physical, chemical, and biological systems. For example, psychologists, regarding humans as complex systems, try to identify the factors that determine the behavior of the systems (humans) through psychological experiments. Various types of knowledge are needed to organize psychological experiments. The ability to control experimental factors, CVS (the Control of Variables Strategy), is regarded as one of the most important skills. Klahr et al. have empirically studied the CVS ability of various types of subjects, such as elementary school students, university undergraduates, and graduates majoring in psychology, by analyzing the discovery process for programming grammar to manipulate a toy vehicle called BigTrak (Klahr, 2000). Moreover, they tried to apply the findings on CVS ability obtained in their laboratory studies to a real educational environment (Klahr, 2001).

Schunn and Anderson constructed a simulated psychology laboratory, called SPL, on a computer. Using SPL, they conducted an experiment in which university students and professional psychologists participated, and analyzed their abilities for designing and interpreting experiments (Schunn & Anderson, 1999). In their analysis, they discussed the difference between the general domain-independent and domain-dependent skills used by each participant for planning psychological experiments.

Additionally they proposed that SPL could be used as a learning environment for tutoring in experimental planning skills (Schunn & Anderson, 2001). However, in SPL, two ad hoc theories were given to the participants; and the participants were required to plan experiments that determined which of those two theories was valid. The process of forming theories (hypotheses) was ignored. Additionally, SPL did not actually simulate the human cognitive process, but simply output subjects' performances, using a previously installed function, as numeral values of the parameters input by the subjects. The process through which the output was obtained was not considered. In the present study, we construct a more realistic and complex experimental environment called VPL (Virtual Psychology Laboratory). Using VPL, we let university undergraduates experience conducting psychological experiments that lasted for several hours.

Schunn and Anderson were mainly interested in how the participants' behavior changed based on their degree of expertise in the research domain concerned. Our interest, on the other hand, is to show changes in the participants' behavior, such as in the formation and verification process of hypotheses including the stage of experimental planning, as a function of their training. We are also interested in the effect of VPL as a simulated psychology laboratory on the training and the improvement of learner's experimental behavior.

2. Experimental environment

2.1 VPL: Virtual Psychology Laboratory

In VPL, two production systems collaboratively solve a traditional discovery task: Wason's 2-4-6 task (Wason, 1960). The mission given to participants was to study factors determining the systems' performance. We can think of the factors determining the performance as, for instance, the degree of difference between the two systems' strategies, the interaction between those strategies and nature of targets, and the capacities of the systems' working memory.

It should be noted that this research theme being used for VPL is a highly realistic subject that has been

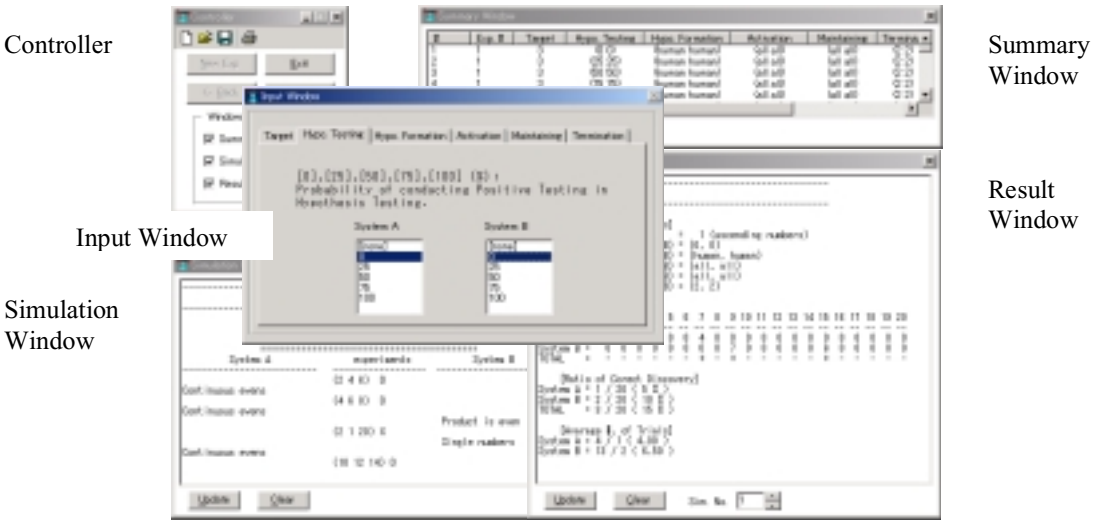


Figure 1: System's interface.

discussed, through recent decades, by psychologists studying human collaborative discovery in laboratory studies (Gorman, 1992; Laughlin, et al., 1997). Moreover, it is also important to note that the psychological validity of this simulator has been tested by our several experiments. The system's performance is determined according to an actual simulation of solving the task. We have already confirmed that the performance of this simulator reflects that of humans well (Miwa, 2001). Figure 1 shows the interface of VPL.

The "Controller" manages the starting and ending of simulations and the appearance of each window. The participants set up experimental factors in the "Input Window". The "Simulation Window" presents a real time process of two production systems solving the Wason's 2-4-6 task. The "Result Window" shows the final result of each simulation. The "Summary Window" summarizes the experimental results obtained by the preceding simulations.

Table 1 shows the experimental factors that the participants can manipulate. Five of the six factors (excluding "Target") are specified in each of the two production systems. In the following experiment, the values of two parameters (# of activated instances and # of maintained hypotheses) were fixed at "all"; the participants could manipulate only the other four parameters. The performance of the simulator is determined by various factors. The fundamental nature of its behavior, such as the existence of interaction between the generality of the targets and the hypothesis-testing strategy (Klayman & Ha, 1987) and a main effect of the working memory capacity (# of activated instances and # of maintained hypotheses), is thoroughly consistent with the findings that several psychologists have reported in real psychological experiments.

2.2 Experiment

Participants: Twenty undergraduate students, not majoring in psychology, participated in the experiment as a part of a university class.

Background knowledge: Prior to the experiment, the participants learned the experimental procedure of Wason's 2-4-6 task, and also the research objectives and motivations of laboratory studies using this kind of simple task. In a preliminary class, the participants read a research paper, which was experimental material prepared by the authors. The paper indicated the experimental result when a single subject solved the task. The result showed that there was interaction between the hypothesis-testing strategy and the nature

Table 1: Factors determining the simulator's

Factors	Levels
Target	[#1] - [#35] Thirty-five kinds of targets used in the experiment. For example, Target #1 is "ascending numbers"; Target #35 is "three different numbers".
Hypothesis testing strategies	[0], [25], [50], [75], [100] The probability of conducting positive tests in generating instance [100] and [0] mean that the simulator always conducts positive tes and negative tests, respectively.
Hypothesis formation strategies	[human], [random], [specific], [general] [human] means that the simulator generates hypotheses as human do. [random]: generating hypotheses randomly. [specific]: generating specific hypotheses prior to general ones. [general]: generating general hypotheses prior to specific ones.
# of activated instances	[all], [6], [5], [4], [3] The number of instances that can be activated at once in the working memory when generating hypotheses.
# of maintained hypotheses	[all], [5], [4], [3], [2] The number of previously rejected hypotheses that can be maintained in the working memory.
Condition for terminating the search	[all], [5], [4], [3], [2] The number of continuos confirmations when the simulator terminates the search. [2] means when a hypothesis is continuous confirmed two times, the simulator recognizes the hypothesis as th final solution, and terminates the search.

of the targets. The participants took part in the experiment after understanding this finding.

Procedure: Three experimental sessions were conducted at intervals of a week. Each session lasted for one hour and a half. At the end of each session, the participants were required to report the findings they had obtained from a series of experiments in the session.

The participants' behavior in each experimental session basically repeated the following procedures. First, the participants entered, in the experimental sheet, (1) the objectives of the experiment they would perform (what are they investigating?), (2) the prediction of the experimental result, and (3) the experimental planning used for controlling experimental factors (which factors are focused on and which levels of each factor are searched?); then they performed the series of experiments planned in the experimental sheet, by manipulating the simulator. After obtaining the experimental result, they entered (4) the interpretation of the experimental result. The participants repeated this series of procedures until the end of the session.

Pre- and Post- tests: Before and after the three experimental sessions, pre- and post- tests were conducted to measure the subjects' fundamental ability to control experimental factors.

3. Experimental results

3.1 Chunking behavior

We define a set of organized experiments as a chunk. Thus, we think of a more sophisticated construction of experimental planning as a process of constructing higher chunks (Miwa, 2000).

The participants conducted their experiments by searching the experimental space as depicted in Figure 2. As mentioned before, two factors, # of activated instances and # of maintained hypotheses, were fixed at the value "all". The participants manipulated the simulator and obtained experimental results after filling in the experimental sheet. A set of experiments planned in a piece of the experimental sheet is regarded as a unit of experiments. Almost all experimental planning (about 96%) entered in a piece of the experimental sheet was constructed based on the factorial experiment design. So, for example, when p levels and q levels in each of two factors were searched, a total of p x q experiments was completely performed according to the experimental planning. We excluded, from the following analysis, units of experiments (4%) which violated this factorial experiment design.

We regard this set of experiments planned in a piece of the experimental sheet as the most basic chunk. We call this basic chunk a "Unit". The participants combine multiple Units to construct a higher chunk. We propose the following two types of chunking, Type A and Type B, as methods for constructing a higher chunk.

See Figure 3 in which a Unit is constructed by the set of experiments where some levels of Factor *n* and Factor *m* are searched. The first type of chunking is Type A (Figure 3(a)) where the searcher shifts a searching level of Factor *k* one by one, while maintaining the search of Factor *n* and Factor *m*. The set of these experiments can be grouped as a chunk of experiments in which three factors, Factor *n*, Factor *m*, and Factor *k*, are simultaneously controlled. The important point is that factors other than the controlled

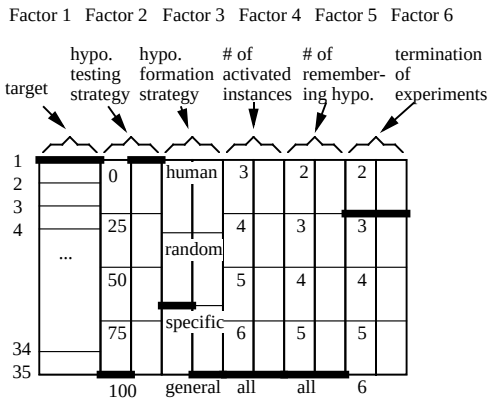


Figure 2: The experimental space searched by Ss.

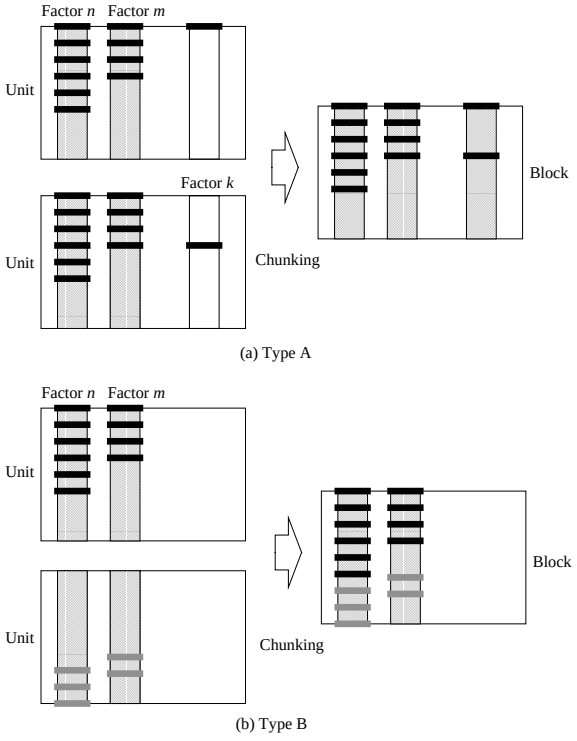


Figure 3: Two types of chunking behavior.

three factors are fixed at an identical level.

The second type of chunking, Type B (Figure 3(b)), occurs when it is impossible that all levels involved in the focused factors, such as Factor *n* and Factor *m*, can be searched at the same time; the search is divided into multiple Units. In this case, the set of multiple Units can also be seen as a chunk. The point is that factors other than Factor m and n are fixed.

By using these two types of chunking, bigger chunks can be constructed from multiple basic Units. We call these higher chunks "Blocks". Here we define the compression ratio of chunking based on the number of individual experiments constructing a single Block. For example, in Figure 3 (a), one Block is constructed from 48 experiments [= 6 (Factor n) x 4 (Factor m) x 2 (Factor k)]; so the compression ratio of chunking is 0.021 (= 1/48). On the other hand, in Figure 3(b), as 30 experiments (= 6 x 4 + 3 x 2) construct a Block, the compression ratio of chunking is 0.033 (= 1/30). The smaller ratio of chunking means that the participants are able to construct a bigger chunk in their experimental behavior. Consequently, the compression ratio of chunking reflects the degree of participants' organizational experimental behavior.

Figure 4 shows, for each of the three experimental sessions, the average compression ratio of chunking of the 16 out of 20 participants, who participated in all of the three experimental sessions. (Similarly analyses of these 16 subjects' results are shown in sections 3.2, 3.3, and 3.4.) As the experimental sessions proceeded, the compression ratio of chunking decreased. As a result of ANOVA, a main effect of the experimental sessions was significance ($p < .01$). This result confirms that the participants learned to construct bigger chunks, i.e., exhibited more organizational behavior, through repeating experimental activities.

3.2 Controlled factors

We can also confirm the process of constructing bigger chunks by analyzing the transition of the number of controlled factors by the participants across the three experimental sessions. Figure 5 shows the average ratios of the number of Blocks, in which one, two, and three or more factors were controlled, to the number of all Blocks. As the experimental sessions proceeded, the ratio of Blocks manipulating more than three factors increased, whereas the ratio of Blocks manipulating one factor decreased. As a result of ANOVA, there was interaction between the experimental sessions and the number of controlled factors ($p < .05$). A simple main effect of the experimental sessions at each of the two single levels, one and more than three, in the number of controlled factors was significance ($p < .05$ and $p < .01$ respectively). The result above shows that the participants learned, during the progress of the experimental sessions, to conduct experiments in which a greater number of various factors were manipulated.

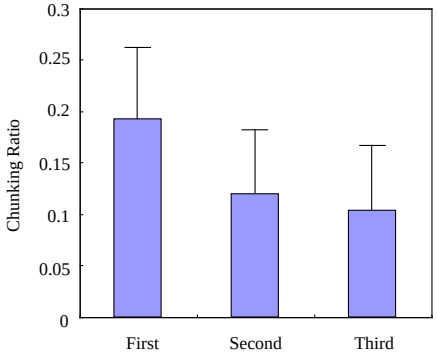


Figure 4: Transition of the ratio of chunking.

3.3 Hypotheses

We also focused on the hypotheses formed by the participants.

The participants entered their prediction of the experimental results in the experimental sheet before executing a series of experiments. At that time, they also estimated the degree of confidence in the prediction on a 1 to 5 scale. Additionally, after executing the experiments with the simulator, they entered their interpretation of the experimental results. At that time, they also estimated the degree of correctness of their prediction on a 1 to 5 scale.

Figure 6 shows the average degree of confidence estimated before executing experiments and the average degree of correctness estimated after the experiments. The figure indicates that the degree of correctness was improved from the first to third sessions while the degree of confidence was almost constant. As a result of ANOVA, there was interaction between the experimental sessions and the two kinds of participants' estimation (the degree of confidence and correctness) ($p < 0.01$). A simple main effect of the experimental

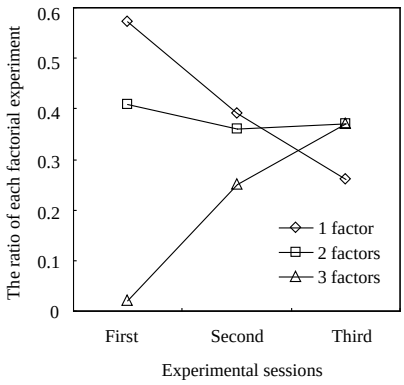


Figure 5: Transition of the number of controlled factors.

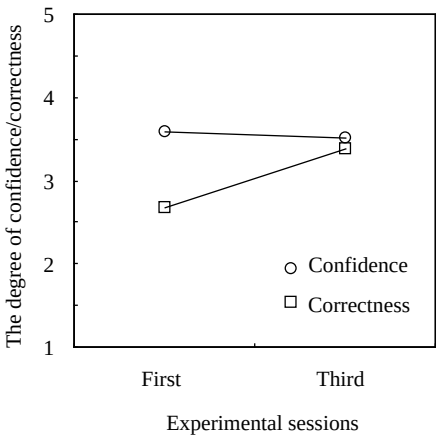


Figure 6: Transition of the degree of confidence/correctness of hypotheses.

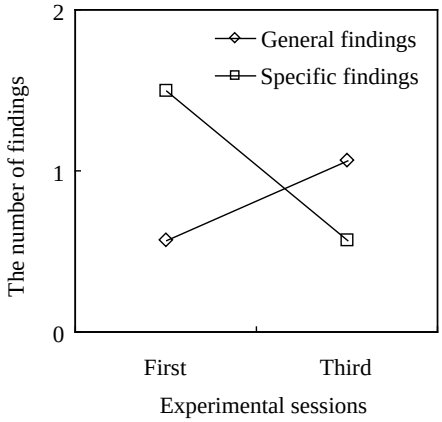


Figure 7: Transition of the number of specific/general findings.

sessions at the degree of correctness revealed significance ($p < 0.05$) whereas an effect at the degree of confidence did not.

The degree of correctness reflects the objective validity of the participants' hypotheses whereas the degree of confidence reflects the participants' subjective estimation of the probability of their hypotheses. The invariant of the degree of confidence implies that the change in the complexity of the participants' hypotheses was not so marked between the former and latter parts of the experimental sessions. On the other hand, the improvement in the degree of correctness confirms that the participants learned to form more accurate hypotheses during the progress of their experiments even though the complexity of the hypotheses was almost constant.

3.4 Findings

Next we move to an analysis of the findings that the participants discovered. As mentioned before, the participants were required to report their findings at the end of each experimental session.

We categorize the findings from the viewpoint of their generality. We define participants' general conclusions mentioning the relation between an experimental factor (or factors) and the system's performance as general findings. For example, the conclusion, "positive testing is effective in finding the specific targets whereas negative testing is effective in finding the general targets", is an example of a general finding because the participants mention the relation between the two factors, the nature of targets and the hypothesis-testing strategies, and the system's performance. On the other hand, we define restricted conclusions mentioning a factor (or factors) determining the system's performance only in a specific situation as specific findings. For example, the conclusion, "in terms of target #27, negative testing is effective ", is an example of a specific finding because this conclusion mentions a restricted finding for a specific case: target #27.

Figure 7 shows the average number of specific and general different findings in the first and third experimental sessions. The figure shows that the number of general findings increased across the sessions while the number of specific findings decreased. As a result of ANOVA, there was interaction between the experimental sessions and the nature of findings (specific and general) ($p < 0.01$). Simple main effects of the experimental sessions at both levels of specific and general in the nature of findings revealed significance ($p < .01$ and $p < .05$ respectively). This confirms that the participants gradually came to discover general findings during the progress of the experimental sessions.

3.5 Improvement from Pre test to Post test

Lastly, we discuss whether the participants learned general procedural knowledge on experimental planning by analyzing the pre- and post- tests that were carried out before and after all of the experimental sessions.

In the pretest, the participants were required to plan an experiment that identified the factors (temperature and/or humidity) responsible for the growth of bacteria. In the posttest, an isomorph of the problem in the pretest was used where the participants were required to identify the factors causing the growth of plankton. The participants' solutions in each test were categorized into two types: (1) for identifying the factors determining the growth of bacteria or plankton, first varying one factor while fixing the other factor then manipulating that other factor (that is, first varying humidity while fixing temperature then varying temperature while fixing humidity); and (2) simultaneously controlling both two factors. We call

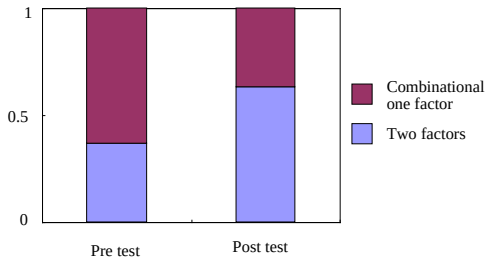


Figure 8: The comparison of experimental planning in the pretest and the posttest.

the former planning a "combined one factor experiment", and the latter a "two factors experiment". The latter planning is more sophisticated because it can detect interaction between the two factors, but the former cannot. Figure 8 shows the comparison of the solutions of 19 participants in the pretest and in the posttest. One of the 20 participants was excluded from the analysis because the subject indicated a confusing answer. Fisher's exact analysis supported a tendency in the increase of the two factors experiment in the posttest compared to in the pretest ($p < .1$).

The above result confirms that some of the participants successfully acquired general procedural knowledge on conducting appropriate experimental planning through repeatedly performing experiments using VPL.

4. Discussions and conclusions

In this experiment, the participants were not given any instruction from a tutor. The participants experienced the three experimental sessions receiving the feedback from the simulator while repeatedly performing their experiments by themselves without any instruction from others. However, the various types of participants' performance, such as organizational designing of experiments, the degree of correctness of formed hypotheses, and the generality of findings, were remarkably improved. This implies that this kind of exercise using a simulated research environment, such as VPL, could be effective for providing tutoring in psychological activities to students who begin to learn experimental psychology.

We understand that it was still not clear that these improvements were brought about by the learning of general experimental skills such as CVS or simply by the increase of information on the problem space searched during the progress of the experiments. However, we believe that the improvement of the scores from the pretest to the posttest confirms that some of the participants had learned something related to general skills on experimental planning because the contents of those tests were independent from those dealt with in the exercise using VPL. At any rate, the experimental results support the possibility of achieving "learning by doing" without instructions through this sort of relatively short-term exercise by using a VPL-like learning environment (Anzai, 1979).

In our future work, we will examine the usage of VPL as an experimental microworld. We could clarify, for instance, the difference between Novices' and Experts' experimental processes and the effects of background knowledge on the processes. We will also further discuss on the possibility of using VPL as a tutoring system. For example, it might be possible to activate the participants' learning process by giving informative feedback to learners based on the idea of constructing higher chunks.

References

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. Psychological Review, 86, 124-140.

Gorman, M. (1992). Simulating science: heuristics, mental models, and technoscientific thinking. Indiana university press.

Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, Mass.: MIT Press.

Klahr, D., Chen, Z., & Toth, E. (2001). From cognition to instruction to cognition: a case study in elementary school science instruction. In C. Crowley, et al. (Eds.), Designing for science: implications from everyday, classroom, and professional settings. Mahwah, NJ: LEA.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94, 211-228.

Laughlin, P. R., Magley, V. J., & Shupe, E. I. (1997). Positive and negative hypothesis testing by cooperative groups, Organizational behavior and human decision processes, 69, 265-275.

Miwa, K. (2000). Human Discovery Processes Based on Searching Experiments in Virtual Psychological Research Environment. LNAI, 1967, 225-239.

Miwa, K. (2001). Emergence of effects of collaboration in a simple discovery task. Proceedings of the 23rd annual conference of the cognitive science society, 645-650.

Schunn, C. D., & Anderson, J. R. (2001). Acquiring expertise in science: explorations of What When and How. In C. Crowley, et al. (Eds.), Designing for science: implications from everyday, classroom, and professional settings. Mahwah, NJ: LEA.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. Cognitive Science, 23, 337-370.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12, 129-140.

Learning to Solve Complex Propositions: Does knowledge of truth-values bootstrap modal operators?

Bradley J. Morris (bjmorris@pitt.edu)
University of Pittsburgh, LRDC, 3939 O'Hara St.
Pittsburgh, PA 15260 USA

David Klahr (klahr@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Evaluating complex propositions requires evaluating truth-values and assigning modal operators. Previous research suggested that evaluating truth-values may be the key to assigning modal operators. This study placed 111 third and fifth grade children in one of three training conditions: no training, training truth-value assignment, and training truth-value and modal operator assignment. The results indicate that truth-value assignment training is sufficient to significantly improve children's evaluations of complex propositions.

Reasoning with complex propositions (statements using AND, OR, NOT, IF) forms the basis of much higher-order thinking. There are three reasoning classes associated with processing propositions: evaluating a proposition as true or false (truth-values), evaluating whether a conclusion follows from the premises (validity), and judgments about possibility and necessity (modal operators). While much research has focused on judgments about validity (for a recent review see Markovits & Barrouillet, 2002), we will focus on a less-researched area: evaluating truth-values and assigning modal operators.

The assignment of truth-values entails determining the truth or falsity of a statement (Johnson-Laird, 1983). The complexity of assigning truth-values depends on the number of elements being evaluated (i.e., how many items need to be evaluated) and the number of states under evaluation (the number of combinations and their associated truth-values). To better illustrate this point we will provide examples of two representative tasks. The first is sentence verification. In a sentence verification task, subjects typically are given a simple proposition (e.g., the star is white) to evaluate either with their existing knowledge or with some reference materials (e.g., picture of a white star). There are two possible values for each proposition: true or false. Because there is only one element under consideration, the evaluation is based on semantic properties (Roberts, Wood, & Gilmore 1994). The second type of task, evaluating complex propositions such as conjunctions and disjunctions, is more complicated than sentence verification because it requires the evaluation of two elements and four possible states. For example, when evaluating a conjunction (e.g., the star is white and the circle is blue), each single proposition has its own truth-value (white star; blue circle). Additionally, the statement as a whole is only true if both single propositions are true, thus there is only one of the four possible resulting combinations that results in an assignment of "true" for the entire statement.

Assigning modal operators is determining when a statement is possible or necessary (Johnson-Laird, 1983). Like the assignment of truth-values, the assignment of modal operators differs in complexity depending on the nature of the task. Modal operators can be assigned on statements such as "A brother is a boy" in which by definition the statement is necessarily true (Miller, Custer, & Nassau, 2000). In this task the assignment is based on purely semantic factors. A more difficult task is assigning modal operators for complex propositions such as contradictions and tautologies. To determine that a tautology is always true (possible) and that a contradiction is always false (impossible) requires evaluating the semantics and syntax of the statement. That is, one must consider the truth-value of the connective and whether the semantic elements match any of the possible truth-values. Thus, a contradiction will always be false because one of the two propositions will always be false and an AND statement requires both elements to be true for the entire statement to be true.

Development of Truth-values and Modal Operators

Very little attention has been given to how children coordinate assigning truth-values and assigning modal operators. That is, are these processes related and if so, how? Perhaps gains on one phase do not correspond to gains on the other, thus we will call this possibility the *separate phase hypothesis*. Much previous research on either process has focused on a single process without examining the other typically reporting performance in one without respect to changes in the other (Ruffman, 1998; Braine & Rumain, 1981; Osherson & Markman, 1976; Paris, 1974). Thus, perhaps the two are not related.

The only theoretical position that has examined both processes, mental logic, states that the two processes are part of a single inferential schema that is acquired with language (Braine & O'Brien, 1997). Once activated, these schemas fire a series of inferential rules that produce a

correct conclusion for all types of inferences (truth-values and modal operators). Thus this position suggests that children should either err on assigning truth-values or modal operators or produce correct evaluations for both, but there should be no inconsistency. We will call this approach the *compiled phase hypothesis*. Previous research demonstrates differences in young children's ability to assign truth-values and modal operators (Morris & Sloutsky, 2002; Miller, Custer, & Nassau, 2000; Braine & Rumain, 1981; Osherson & Markman, 1976), however, most of this research did not look at the consistency of responses within individual children. Two studies that examined within-participant consistency (Morris & Sloutsky, 2002; Morris & Klahr, in review) found large differences in children's performance specifically, that children produced correct responses on assigning truth-values before modal operators. These differences may suggest that such coordination in these types of inferences may occur late in development (see Morris & Sloutsky, 2002, for a discussion).

It is also possible that the two types of inferences are related and that knowledge about one is related to the development of the other. We will call this the *dependent phase hypothesis*. Morris & Sloutsky (2002) and Morris and Klahr (in review) examined children's assignment of truth-values and modal operators and observed that many children erred on one while giving correct responses on the other, providing evidence against the compiled phase hypothesis. Further, in each case, children who correctly assigned modal operators also assigned correct truth-values while the converse was not true suggesting that the ability to assign correct truth-values *preceded* the ability to assign correct modal operators. The *dependent phase hypothesis* suggests that children must first learn to assign correct truth-values before they can assign modal operators because the mappings that allow truth-value provide sufficient information to infer modal operators.

The present study was designed to compare the three approaches and to test the dependent-phase hypotheses. A training study was conducted in which children were given instruction on assigning truth-values and modal operators. Children were randomly assigned to one of three conditions. The first instructional condition ("mapping") provided explicit instruction on evidence evaluation only. The second instructional condition ("necessity") gave children explicit instruction on evidence evaluation and evidence requests. The third condition ("control") gave children no instruction.

Our first prediction is that mere exposure to logical statements is insufficient to improve performance on any processing phase. If all conditions show equal improvement on the post-test, then this supports the compiled phase hypothesis because learning was not required, only familiarity with the type of problem. The second prediction is that training is necessary for improvements in performance. If the mapping and necessity conditions show significant improvement from pre- to post-test and children's performance in these conditions are significantly better than the performance of children in the control condition, this provides evidence against the compiled phase hypothesis (because training was required to improve performance) and supports the separate and related phases hypotheses. The third prediction is that improvement is specific to the phases on which training has been given. If children in the mapping and necessity condition are not be significantly different in their post-test performance levels then this supports the dependent phase hypothesis, however, if the mapping and necessity conditions are different on the post-test, then this supports the separate phase hypothesis.

Method

Participants

The participants were 111 children: 60 third (mean age 8,8; 29 boys, 31 girls) and 51 fifth (mean age 10,8; 25 boys, 26 girls) graders enrolled in two public or two private schools located in Pittsburgh, Pennsylvania. Participants were chosen on the basis of returning a parental consent form. Third and fifth grade children were selected because cross-sectional studies (e.g., Morris & Sloutsky, 2002) have demonstrated significant improvements in logical reasoning between these ages. Children in each grade were randomly assigned to one of the three conditions.

Design

We used a 3 (condition) x 2 (grade) x 4 (session) design with session as within-participants measure. The three conditions (control, mapping, necessity) differed in the amount of explicit instruction on how to evaluate different types of logical statements. The control group was given a series of 12 problems formally identical to those in the pretest in both training sessions without any instruction. Children in the mapping condition were given explicit instruction about rules used to evaluate evidence in logical statements. Children in the necessity condition were given the same evaluation rules as those in the mapping condition and were given explicit instruction about when evidence was necessary and unnecessary.

Each child participated in four sessions: a pre-test, training 1, training 2, and a post-test. In the pre-test, children were asked to evaluate the truth-status of a series of 16 statements. In the second and third sessions children were given one of the three training conditions. In the fourth session each child was given a post-test formally identical to the pre-test.

Procedure

The procedure was divided into four sessions over four days: (1) Day 1- Pre-test, (2) Day 2- Training 1, (3) Day 3- Training 2, (4) Day 4- Post-test. Each session was separated by approximately one week (M= 8.2 days). Each child was interviewed individually in a quiet location in his or her school. The interviewer recorded each child's responses for all four sessions.

Session 1- Pre-test

All instructions for the experimental segment were read to each participant and repeated if requested. The pre-test consistent of 4 warm-up statements and 16 actual statements. Two cards were placed in front of each child: a statement card (face up) and an evidence card (face down). The order of presentation of statement and evidence pairs was counterbalanced across participants. Each child was presented a total of 16 statements corresponding to 4 of each of the following types: conjunctions, disjunctions, tautologies, and contradictions. The child then read the statement card aloud. After the child read the statement, Question phase 1 (*a priori* evaluation) was then asked ("Is the sentence true, not true, or can't you tell?"). After the answer was recorded, Question phase 2 was asked (evidence request: "Do you need to see the picture to find out?"). If evidence was requested, the evidence card was turned over (all were pictures) and placed in front of the participant. Question phase 3 was then asked (evidence evaluation) in which the child was asked to evaluate the initial statement using the evidence requested ("Now that you have seen the picture, was the sentence true, not true, or can't you tell?"; Asked only if evidence was requested). The task took approximately 15 minutes.

Session 2- Training 1

Control Condition. The procedure used in the control condition was identical to the procedure used in the pre-test. Children in the control condition were given a set of 12 statements structurally identical to those in the pre-test but with altered content. As in the pre-test children were asked up to three questions for each statement corresponding to the three processing phases. Children were given no feedback or instructions after their responses. This procedure lasted 10 minutes.

Training Warm-up Segment. Note the training warm-up was used only with the mapping and necessity conditions. The training warm-up segment was a brief session in which children were given basic rules for evaluating evidence with the connectives AND or OR. The training consisted of three parts: (1) explanation of sentence parts, (2) explanation of evaluation rules, and (3) rule use/feedback. Children were first given an explanation of the "parts" of each sentence in the task. Each sentence was divided into two parts (clauses separated by the connective) and an "important word" (the logical connective). Children were told that the important word indicated which rule was used to evaluate each type of statement.

Children were then asked to demonstrate the parts and important words on a new sentence with feedback provided for errors. Next, children were given brief instruction on how to evaluate evidence with each type of important word. Children were given simple rules for each connective. Children were told that the evidence had to match both parts for an AND sentence to be true, otherwise it is false. Children were told that evidence had to match only one part of the statement for an OR to be true otherwise the statement is false. The decision to explain an OR as an exclusive OR was made to make a clean conceptual distinction between OR and AND.

Once the rules for important words were explained, children were given a statement and a series of evidence cards, presented one at a time, and asked to identify the parts and important word then to evaluate the statement as a whole. In total, two statements (1 OR, 1 AND) and four evidence cards (1 true, 1 false for each statement type) were given over the training warm-up. Feedback was provided for incorrect responses giving the correct answer and re-explaining the evaluation rule. This procedure took 5 minutes. The warm-up was only given before Training 1.

Mapping Condition. The mapping condition provided explicit instruction about evaluating evidence with statements. Each child was given four statements (one of each type) and three different pieces of evidence for each statement. The mapping condition provided instruction for each statement in three stages over which the scaffolding provided by the interviewer was gradually reduced: (1) explicit instruction, interviewer-led solution, (2) probe questions, scaffolded solution and (3) probe questions, child-led solution.

In the first stage of the instruction, the interviewer placed a statement card on the table and asked the child to evaluate the statement as true, not true, or can't tell. The interviewer then placed the first evidence card on the table and asked the child to evaluate the statement based on the evidence card. The interviewer then provided explicit instruction about the parts and important words of each statement and the correct conclusion regardless of the child's response. The second evidence card was then placed on the table beside the first evidence card. The interviewer asked the child to repeat the rule for the important word, match evidence to each part of the sentence, and then to evaluate the sentence as a whole. If any evaluation was incorrect, children were given immediate feedback. The third evidence card was placed on the table beside the first two cards. The interviewer then asked the child evaluate each part of the sentence and evaluate the sentence as a whole. Feedback was provided only if the child answered incorrectly. This procedure was repeated for each statement type and took approximately 10 minutes per child.

Necessity Condition. The procedure in the necessity condition was identical to the mapping condition with the addition of two probe segments after evaluating the second and third evidence cards. After the second evidence card, the two evidence cards (and the truth-values associated with them) were reviewed to determine if their contents changed the truth-values of the statement as a whole. After the third evidence card, all cards were reviewed. The child was then asked whether the evidence changed the statement's truth-value. If the evidence was necessary, then the child was told that it was necessary to first see evidence before evaluating "*this type"* of statement. If evidence was unnecessary, then

the child was told that no evidence would change the truth-value of "*this type*" of statement because, for example, the two clauses in a contradiction can never match the same evidence (in this case the same picture).

Session 3: Training 2

The training 2 was identical to that used in Training 1.

Session 4: Post-test

The post-test procedure was identical to the pre-test.

Materials

Pre- and Post-Test Materials. Pre- and post-test materials were logically identical but had slightly different content. Each session required the evaluation of four statement types: tautology, contradiction, disjunction, and conjunction. Each child saw four instances of each statement type. The materials consisted of 16 unlined 3 x 9" cards with one statement and 16 3 x 5" cards with a corresponding piece of evidence. An additional six cards (3 statement, 3 evidence) were used as materials for the warm-up items.

Control Condition. Two sets of materials (one for each training session) were created for the control condition. Each set contained 12 statements (3 of each type) printed on 3 x 9" cards and 12 3 x 5" evidence cards. Each set also included three statement cards and three evidence cards for the warm-up segment.

Training Warm-up Condition. The training warm-up (for mapping and necessity conditions only) included two statement cards used to demonstrate the parts of the sentence, plus two statement cards and four evidence cards for articulating the AND and OR rules.

Mapping and Necessity Conditions. The mapping condition required two sets of materials, one for each training session. Each set of materials included four statement cards (one of each statement type) and twelve evidence cards, three evidence cards corresponding to each statement card.

Results

The results will be presented in two sections: (1) aggregated analyses of the effects of training across each processing phase separately and (2) individual strategy analyses comparing changes in the consistency of response patterns across processing phases. The first set of analyses will separately examines training effects on a priori evaluations, evidence requests, and evidence evaluations. The second analysis examines changes in inter-phase consistency before and after training.

Aggregated analyses

Question Phase 1: A Priori Evaluations (AP). For all *a priori* evaluation phases, possible responses were "true," "not true," or "can't tell." Correct responses were coded as follows: contradictions- not true; tautologies-true; conjunctions and disjunctions-can't tell. Correct responses were scored as 1 while incorrect responses were coded as 0.

To determine the effectiveness of training on the number of correct a priori evaluations, a 3 (condition) x 2 (age) x 2 (pre-test vs. post-test) ANOVA was performed with session as a within-subjects variable. The analysis reveals a main effect for condition, F (2, 110) = 6.1, p <.003, and age F (1, 110) = 8.9, p <.003 and no interaction between condition and age F (2, 110) = .42, p >.65. Children in the mapping and necessity conditions gave significantly more correct responses in the post-test than in the pre-test. The performance of children in the control condition did not differ significantly from pre-to post-test. Fifth graders gave significantly more correct requests in the post-test than third graders.

Question Phase 2: Evidence Requests (ER). For the evidence request phase, possible responses were "yes" or "no." Correct responses were coded as follows: contradictions-and tautologies-No; conjunctions and disjunctions-Yes. Correct responses were scored as 1 while incorrect responses were coded as 0.

To determine the effectiveness of training on the number of correct evidence requests, a 3 (condition) x 2 (age) s 2 (pre-test vs. post-test) ANOVA was performed with session as a within-subjects variable. The analysis reveals a main effect for condition, F (2, 110) = 3.9, p <.01, and age F (1, 110) = 10.6, p <.001 and no interaction between condition and age F (2, 110) = .42, p >.65. Children in the mapping and necessity conditions gave significantly more correct responses in the post-test than in the pre-test. The performance of children in the control condition did not differ significantly from pre-to post-test. Fifth graders gave significantly more correct requests in the post-test than third graders.

Question Phase 3: Evidence Evaluations (EE). For the evidence evaluation phase, possible responses were "true," "not true," or "can't tell." Correct responses were coded as follows: contradictions-not true; tautologies-true; conjunctions (2 true, 2 false) and disjunctions (2 true, 2 false). Correct responses were scored as 1 and incorrect responses were coded as 0.

To determine the effectiveness of training on the number of correct evidence evaluations, a 3 (condition) x 2 (age) x 2 (pre-test vs. post-test) ANOVA was performed with session as a within-subjects variable. The analysis reveals a main effect for condition, F (2, 110) = 20.8, p <.001, and age F (1, 110) = 9.2, p <.003 and no interaction between condition and age F (2, 110) = .25, p >.77. Children in the mapping and necessity conditions gave significantly more correct responses in the post-test than in the pre-test while children

in the control condition did not differ significantly from pre- to post-test. As in previous conditions, fifth graders gave significantly more correct requests in the post-test than third graders.

The aggregated analysis demonstrated that experience with statements was not sufficient to improve performance, at least not in the limited exposure provided during the training period. Training effectively improved performance for children in the mapping and necessity conditions. The next series of analyses will examine the structure of change.

Individual Analysis

The individual analysis examined the consistency of a child's correct response patterns within each processing phase and across all processing phases. For example, although the aggregated data demonstrates that fifth graders generally outperformed third graders, these data do not indicate the extent to which an individual fifth grader produced correct answers for the evidence request phase or for all question phases. We considered a pattern in which 75% of responses were correct as *consistently correct*. A pattern below 75% was considered inconsistent. Tables 1 and 2 display the number of children coded as consistently producing correct responses for each processing phase in the pre- and post-tests.

Table 1- Percentage of Third Grade Children Giving Consistent, Correct Responses Within Each Processing Phase by Condition

Condition	AP	ER	EE
Control	0 (10)	0 (19)	19 (19)
Mapping	5 (30)	8 (40)	5 (85)
Necessity	5 (37)	5 (63)	11 (79)

Note. Posttest scores are presented in parentheses.

Table 2- Percentage of Fifth Grade Children Giving Consistent, Correct Responses Within Each Processing Phase by Condition

Condition	AP	ER	EE
Control	10 (14)	19 (29)	14 (10)
Mapping	15 (59)	10 (60)	25 (80)
Necessity	26 (62)	37 (79)	26 (95)

Note. Posttest scores are presented in parentheses.

There was a significant difference between the number of consistent, correct responses in pretests and in the post-tests for third and fifth graders. Third graders produced significantly more consistent evidence evaluations $\chi 2$ (2, 40.8, $p < .001$) in the post-test than the pre-test. Fifth graders produced significantly more evidence requests $\chi 2$ (2, 10.5, $p < .003$) and significantly more consistent evidence evaluations $\chi 2$ (2, 14.5, $p < .001$) in the post-test than the pre-test.

These data were used to code the child's overall response pattern (i.e., consistency measure on each of the three processing phases) as either consistently correct (consistent, correct responses on all phases) or mixed (inconsistent on one or more phases). The results are displayed in Table 3. Children in both training conditions produced significantly more consistent correct responses than children in the control condition, third graders $\chi 2$ (2, 4.2, $p < .03$), fifth graders $\chi 2$ (2, 15.1, $p < .001$).

Table 3- Percentage of Children Giving Consistent Correct Responses by Grade and Condition

	3rd		5th	
Condition	Pre	Post	Pre	Post
Control	0	9.5	13	13
Mapping	0	20	0	59
Necessity	0	37	11	61

Discussion

The results support the dependent phase hypothesis in which changes in performance are due to gains on knowledge of assigning truth-values, which then lead to improvements in assigning modal operators. We tested three predictions with a training study in which children were given varying levels of instruction. These data indicate that (1) contra the compiled phase hypothesis, a small amount of experience with logical statements is not effective for improving performance. Exposure to the control condition was not related to increases in performance. Training was related to significant increases in performance. Specifically, (2) both training conditions were effective in improving performance but (3) training on assigning truth-values was sufficient to produce consistent, correct responses at levels roughly equal to those of the necessity condition (in which both truth-values and modal operators were trained).

Morris and Klahr (in review) examined the order in which children make consistent correct responses on each processing phase. In all cases, correct evidence evaluation always preceded correct evidence requests and *a priori* evaluations. The authors suggested that correctly evaluating evidence provided sufficient knowledge from which children could make further inferences about problem classes.

The results can be explained by extending mental models theory. In Mental Models theory (Johnson-Laird, 1983), for each problem, models are created and searched for possible solutions. Thus, to derive a valid conclusion, a set of tokens is created and searched for possible and impossible solution states. This constitutes a solution for a single instance. As currently formulated, Mental Models makes no provisions for within-subject change in processing due to learning (e.g., a child's performance on trial 100 should be different than their performance on trial 1). The *dependent phase hypothesis* suggests that this type of change should be demonstrated by improvements on each phase and in the structure of change between processing phases. As currently formulated, mental models suggests that developmental

change is a function of increases in working memory. Although this is likely important, such a factor would not explain improvement derived from experience.

We suggest two possible mechanisms that may facilitate changes in performance: *discrimination* and *compilation*. After seeing a sufficient number of instances, a reasoner should form expectations about a class of similar types (e.g., determinate statements). In this way, a reasoner does not approach each instance as a new problem type. Rather, increased reasoning efficiency results from (1) eliminating redundant or unnecessary steps (Crowley, Shrager, & Siegler, 1997) and (2) making additional inferences about the class of similar problem types (Eisenstadt & Simon, 1997; Morris & Klahr, in review). For example, when reasoning about two propositional classes: indeterminate and determinate formal types, children initially confuse the two, treating both as indeterminate (Fay & Klahr, 1996; Morris & Sloutsky, 2002). By sixth grade, children begin to distinguish the two forms, but often fail to correctly determine when evidence is necessary and when unnecessary (Morris & Sloutsky, 2002).

On first exposure to a formal type, a reasoner may make errors at all phases. In evaluating a contradiction, for example, a reasoner may assert that they cannot assign a truth-value to the statement *a priori*, may request evidence and then fail to evaluate this evidence correctly. Once evidence is correctly evaluated and after several correct conclusions have been drawn, our hypothetical reasoner may then be able to infer that no evidence will change the truth-value of this particular statement. Once this inference is drawn, the reasoner may then also assert that the statement is false *a priori*. The processes described may illustrate (1) compilation, or eliminating a redundant processing step (evidence request), and (2) discrimination, or creating a new conceptual category for "statements that do not require evidence." Once this inference is made about one statement, the reasoner may generalize this to other propositions of this type: a statement of the form (A & ~A) is false.

Acknowledgements

This work was supported in part by a postdoctoral fellowship from NICHHD (HD08550) to the first author and grant HD25211 to the second author. Special thanks to Jen Schnakenberg, Anne Siegel, and Jolene Watson for data collection and coding and Corrine Zimmerman and four anonymous reviewers for editorial suggestions.

References

Braine, M., & O'Brien, D. (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182-203.

Braine, M. & O'Brien, D. (Eds.). (1998). *Mental Logic*. Mahwah, NJ: Erlbaum.

Braine, M., & Rumain, B. (1981). Development of comprehension of "Or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31,* 46-70.

Cheng, P., & Holyoak, K. (1985). Pragmatic Reasoning Schemas. *Cognitive Psychology, 17*, 391-416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-316.

Crowley, K., Shrager, J., & Siegler, R. S. (1997). Strategy discovery as a competitive negotiation between metacognitive and associative mechanisms. *Developmental Review, 17,* 462-489.

Eisenstadt, S. A., & Simon, H. A. (1997). Logic and thought. *Minds and Machines, 7,* 365-385.

Fay, A. L., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: Preschoolers' understanding of indeterminacy. *Child Development, 67*(2), 689-716.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1999). Deductive Reasoning. In J. Spence (Ed.), *Annual Review of Psychology,* Vol. 50 (pp.109-135). Palo Alto, CA: Annual Reviews.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review, 22*, 5-36.

Miller, S. A., Custer, W., & Nassau, G. (2000). Children's understanding of the necessity of logically necessary truths. *Cognitive Development, 15*, 383-403.

Morris, B.J., & Klahr, D. (in review). From Empirical to Formal: The Role of Evidence in Children's Reasoning Strategies

Morris, B. J., & Sloutsky, V. (2002). Children's solutions of logical versus empirical problems: What's missing and what develops? *Cognitive Development, 76*, 1-22.

Osherson, D., & Markman, E. (1975). Language and the ability to evaluate contradictions and tautologies. *Cognition, 3*(3), 213-226.

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology, 16*, 278-291.

Roberts, M.J., Wood, D.J., & Gilmore, D.J. (1994). The sentence-picture verification task: Methodological and theoretical difficulties. *British Journal of Psychology, 85*, 413-432.

Ruffman, T. (1999). Children's understanding of logical inconsistency. *Child Development, 70* (4), 872-886.

Logical Strategery

Bradley J. Morris (bjmorris@pitt.edu)
Christian Schunn (schunn@pitt.edu)
University of Pittsburgh, LRDC, 3939 O'Hara St.
Pittsburgh, PA 15260 USA

Abstract

We propose a conceptual framework for explaining logical reasoning in terms of competing strategies. Contrary to previous approaches in which a single theory is suggested to explain all logical reasoning, this framework suggests that the core elements of existing theories constitute strategies, each of which have unique processing demands. A strategy is more likely to be used when its processing demands match a problem's task demands. The framework specifies how each strategy may be distinguished theoretically and empirically.

Overview

The study of logical reasoning has typically proceeded as follows: researchers (1) discover a response pattern that is either unexplained or provides evidence against an established theory, (2) create a model that explains this response pattern, then (3) expand this model to include a larger range of reasoning situations. For example, researchers typically investigate a specific type of reasoning (e.g., deduction) using a particular variant on an experimental task (e.g., the Wason selection task). The experiments uncover a specific reasoning pattern, for example, that people tend to match terms between the premises and conclusions rather than derive valid conclusions (Evans, 1972). Once a reasonable explanation is provided for this response pattern, researchers typically attempt to expand this explanation to encompass related phenomena, such as, the role of 'bias' in other reasoning situations such as weather forecasting (Evans, 1989). Eventually, this explanation may be used to explain all performance on an entire class of reasoning phenomena (e.g. deduction) regardless of task, experience, or age. We term this the *unified approach*.

Unified approaches have traditionally suggested that logical reasoning is either rule-based (application of transformation rules that draw valid conclusions once fired; Rips, 1994) or model-based (creating and searching veridical representations of premises and possible conclusions; Johnson-Laird, 1999). It seems possible, however, given the range of problem types, task demands, experience, and cognitive resources of the reasoner, that there may be more than one approach to solving a class of reasoning phenomena.

Logical reasoning tasks are quite varied, ranging from simple tasks such as statement evaluation (e.g., "Is my cat black?") to complex tasks such as predicate syllogisms (e.g., Some A are not B, Some B are C, Some C are not D, Are some D A?). There is no evidence that logical reasoning occupies a particular region of the brain (cite), that it is a coherent process distinct from the rest of cognition (Johnson-Laird, 1999), or that it involves the same cognitive processes across different tasks. Moreover, cognitive psychology has identified a variety of general cognitive processes that are used to solve a wide variety of tasks (e.g., analogy, retrieval, guessing). Thus, it should not be controversial to suggest a multiple strategy approach- that several, general processes might be used to solve logical reasoning problems in at least some situations, even if there are special logical reasoning processes.

We propose an alternative to the *unified approach* in which a series of simple strategies may be used rather than a single complex theory. Thus, we propose a new framework for explaining logical reasoning performance by incorporating simplified versions of existing approaches as possible strategies for solving logic problems. We list a variety of alternatives that seem highly likely to be used in at least some logical reasoning situations. It is not crucial to the argument here that all strategies are actually used. We will propose conditions under which various strategies may be used, thus proposing a framework through which strategies can be distinguished theoretically and empirically.

The multiple strategy approach has been suggested in the domains of judgment and decision-making (see Bettman, Johnson, Luce, & Payne, 1993, and Todd & Gigeranzer, 1999). Like the logical strategy model, these approaches suggest competition between various strategies in which selection is accomplished, in part, from an evaluation of effort-accuracy tradeoffs. However, the logical strategy model differs from these approaches in their function and goals: (a) the function of the LSM is the creation and evaluation of knowledge (inference) rather than selection and evaluation of knowledge for decision making (judgment), and (b) the goal of LSM is to establish (either through production or evaluation) valid inferences while the goal of other approaches is to return the most adaptive decision (Todd & Gigeranzer, 1999).

The Strategy Approach to Logical Reasoning

What does it mean to think in terms of strategies? A glance at the existing literature suggests that unified approaches have difficulty accounting for differences in performance between individuals and across tasks (for a review see Rips, 1994; Johnson-Laird, 1999). We suggest that differences in performance across individuals and

situations are due to the selection of a strategy and that strategy selection is a function of the history of success with each strategy and the match between processing demands of the strategy and the task demands of the problem. We will describe this match between the processing demands and history of success of the strategy and the situation/task demands as the *situational niche*.

As stated earlier, unified approaches have suggested that a single theory can account for the range of human performance. We suggest that the strategy selection approach may explain the same phenomena by proposing a series of specified approaches, each relegated to explaining a subset of the total range of human deductive performance.

What do we gain from this approach? The strategy approach allows established explanations to be incorporated into a single model in which all are possible explanations of behavior differing only in the extent to which the particular strategy has been used in similar situations and matches the task demands. The match between a particular strategy and its situational niche may not be rational, but may help explain individual and situational differences between theories. For example, it is well established that familiar content tends to improve performance on the Wason selection task (Wason & Johnson-Laird, 1972). Hundreds of experiments have investigated this phenomenon and have led to the introduction of a variety of explanations (e.g., matching rules), which differ widely in the range of their applicability. That is, some are specific to a particular set of materials while others seek to explain a larger range of reasoning behavior. What has rarely been investigated is the influence of the situational niche on performance, or, more specifically *how is the task itself contributing to the response pattern*? For example, take two contrasting approaches: in one, specific knowledge is required to solve a problem, and in the other, abstract rules (excluding the influence of knowledge) are used to solve problems. Problem A is given in which a substantial amount of content knowledge is given that is relevant to a solution. In this case, we would expect the knowledge-based approach to be better suited to solving this problem. If however a different problem were given in which no background information is proved, then we would predict that the second solution is more likely to be used. We suggest that a strategy approach allows flexibility in explanation by allowing for an individual to display a range of possible approaches to a problem set.

The logical strategy approach is in stark contrast to unified approaches in which explanations for processing logical statements are confined to one approach. A possible criticism at this point is that that a unified approach is more parsimonious than a strategy approach. That is, why suggest a series of competing strategies, each of which demands cognitive resources, when one approach would suffice? We provide two responses. First, current unified approaches have been unable to account for a range of performance without many ad hoc additions. For example, Mental Logic theory suggests that logical inferences are derived nearly automatically by a set of content-free inferential rules (Rips, 1994; Braine & O'Brien, 1998). In order to explain the effect of familiar content, Mental Logic theory incorporated an additional step in the reasoning process, a pragmatic filter, which determines if a statement is to be considered logical or conversational. In the former, logical inferential rules are applied, while in the latter case less formal conversational inferential rules are applied. The result is that the theory postulates an approach to reasoning that undermines its own primary thesis, that logical inference is a set of content-free rules applied as automatically as syntax.

Second, the strategy selection approach does not require a series of additional resources but can be accounted for by a small set of general-purpose cognitive mechanisms. Within the strategy selection framework a series of strategies can be derived within minimal effort on the part of the cognitive system from the experiences with the environment. For an example, let us return to the example given above. To account for reasoning in situations in which statements are presented without familiar content and reasoning with statements presented with familiar content, a strategy selection framework could account for empirical findings by the use of two strategies. The first does not use formal rules but uses the content to derive a plausible conclusion related to the specific content. In the second (without the presence of familiar knowledge), inferential rules are used because the most salient property of the problem is the relations between elements, not the content of elements. The use of each is specific to the situation in which both are presented.

In the sections that follow we will first outline seven strategies and their corresponding task demands. We will then examine the influence of task demands on their selection. We first state that the following is an incomplete model. There are many possible strategies and we are suggesting a small number in this paper. Second, the description of each is limited by space, thus does not cover the full range of possible applications. Given the nature of these limitations, the model is reasonably articulated for the purposes of the paper.

Token Based (Mental Models)

Overview. The token based reasoning strategy has the following characteristics: 1) information is represented as tokens derived from natural language which correspond to perceptual or verbal instantiations of possible states and 2) "logical" reasoning is achieved not through the application of formal rules but by the creation, inspection, and manipulation of tokens (Johnson-Laird, 1983; Johnson-Laird, Byrne, and Schaeken, 1992).

Outline of Processing/Processing Demands. Logical inferences are derived from manipulations of models rather than by using inferential rules. There are three steps in processing token-based propositions: propositional analysis, models generation, and model use. **Propositional analysis**

refers to language processing and is largely analogous to representing the surface structure of a statement and requires sufficient verbal/spatial working memory to encode and parse language. **Model generation** refers to the creation of tokens derived from the propositional analysis and relevant information in the existing knowledge base and in the environment. Generation requires verbal/spatial working memory space to create and hold tokens. **Model Use** is the process of searching and evaluating the set of models created by the procedures outlined above and requires a sufficient processing capacity to create veridical models of the necessary information, creation and search for counterexamples, and evaluations of truth-values. The primary limitation on processing is the working memory space required to create and search models for a solution.

The token-based strategy seems particularly useful in the solution of problems in which there are spatial relations because token-based representations can encode such relations more easily than propositional representations (Johnson-Laird, 1983, 1999). For example, in the transitive problem "Bill is to the right of Fred, Fred is to the right of Sam, Is Sam to the right of Bill?") A token based-representation would easily encode the relevant dimensions as follows:

[Sam] [Fred] [Bill]

An obvious limitation of this strategy is that the number of models needs to remain within the current working memory limitations of the reasoner.

Verbal (Mental Logic)

Overview Verbal logic approaches explain logical reasoning as the result of content-free, logical transformation rules applied to linguistically derived mental structures (Rips, 1994; Braine & O'Brien, 1998).

Outline of Processing in Verbal Theories The core elements of verbal theories share basic processing characteristics. Input is represented and processed in a verbal form (e.g., predicate-argument structures; Braine & Rumain, 1983). Sufficient verbal working memory is required to extract the formal elements and hold the representation in a predicate argument structure. **Application of Transformation Rules** are content-free rules represented as either condition-action pairs (Rips, 1994) or as inferential schemas (Braine & O'Brien, 1997). Once verbal input is represented, the content matches a series of transformational rules that produce an output that is either in the form of a conclusion, a statement that will be operated upon by additional rules, or a statement that does not match additional rules. Errors in processing are attributed to a failure in applying the appropriate rule to the statement.

The verbal strategy is most useful in solving abstract statements in which the focus is on relationships between elements. For example, in a version of the Wason selection task the card content is related only by formal structure, not by content (e.g., If there is a vowel on one side, there is an odd number on the other side).

Knowledge Based Heuristics (KBH)

Heuristics are rules that do not utilize logical algorithms. Such a strategy does not generate a valid conclusion but may generate "logic-like" performance (Cheng & Holyoak, 1985). KBH are easily implemented processing rules that use *content* as the basis for deriving a conclusion. Unlike algorithmic approaches (e.g. verbal strategy), these conclusions are not necessarily valid (often violating logical inference rules), yet are often pragmatically supported. An example is Pragmatic Reasoning Schemas (PRS; Cheng & Holyoak, 1985) in which social (permission rules) and physical (causality) regularities form the basis of a series of inferences schemas.

Outline of Processing in KBH

There are three processing steps in KBH: **parsing sentence**, **detection of relations**, and **solution output**. Sentence parsing refers to sentence comprehension and includes verbal and nonverbal information. The detection of relations occurs when the present content is similar to content for which there are established rules. For example, in permission relations, there are established rules (typically phrased as conditionals) that suggest appropriate responses. Matching content allows rules to be accessed. Once rules are accessed, they are applied to the specific situation and a solution output is produced.

The detection of specific relations determines if a statement matches an existing schema. Cues such as temporal sequence suggest obligatory or causal relations between elements. For example, in the statement "Mow the lawn and I will give you five dollars" the condition is set in the first clause while the consequent is set in the second clause. Previous knowledge of other exchanges (in which transactions are made on the basis of obligations) forms the basis of these inferences.

Matching Heuristics

Overview- Matching heuristics are selective processing strategies in which solutions are derived based on superficial elements such as terms or common elements (rather than on content as in KBH). Two well-known examples of matching heuristics are Matching biases and Atmosphere effects (Evans, 1989; Woodsworth & Sells, 1935).

Outline of Processing in Matching Heuristics

Matching heuristics specify rules of *selective processing* differ from all previous strategies in that no specific inferential content is accessed. These rules follow a basic processing model as follows: 1) **encode surface structure**, 2) **find key elements**, and 3) **match key elements**. For

example, in the Wason selection task, subjects prefer to choose cards named in the rules rather than cards that are not named in the rules (Evans, 1972). In the first processing step, the subject encodes the surface structure focusing on the elements in the rule. Most likely this involves encoding an IF→ THEN rule. The key elements are identified. For example, given "If an odd number on one side, then a vowel on the other side" the subject may focus on "odd number" and "vowel" as key elements. Then, when searching possible solution states, the subject will attend to those solution states that contain the key elements. Continuing with the example above, the subject may be more likely to select a card with an odd number and a card with a vowel because they match elements in the rule. This general processing model also applies to atmosphere effects.

Task-Specific Procedures

Overview- Like heuristic strategies, Task-specific procedures are non-logical procedures that achieve correct solutions on logical tasks without the use of formal inferential rules. Task-specific procedures are reasoning "short-cuts" that produce procedural solutions without declarative understanding. The limitation of TSP procedures is that they do not generalize beyond the specific type of reasoning format in which they were induced. Logical training (education and training studies) may produce these procedures leading to an understanding of logical reasoning analogous to the understanding of Chinese attributed to the occupant of the Chinese room (Searle, 1990).

Outline of Processing in Task-Specific Procedures

The processing demands in task specific procedures can be defined as two steps: (1) encoding the relevant problem features and (2) implementing the appropriate algorithm. For example, in a syllogism evaluation subjects concluded that any syllogism with two "somes" in the premises was invalid (Gallotti, Baron, & Sabini, 1986). Implementing a solution algorithm requires sufficient working memory to hold the encoded premises and to fire the appropriate algorithm.

Pragmatic Acquiescence

Overview- Pragmatic acquiescence (PA) refers to response patterns that are attempts to match the expectations of the questioner. In a situation in which someone has little prior knowledge, they may be inclined to seek social cues from the questioner as to how to respond to a novel situation. Rather than matching the conceptual features of the problem as in matching heuristics, PA-based solutions are based on the pragmatics of the problem/testing situation.

Outline of Processing PA- The PA strategy is used when (1) the pragmatic cues are most salient or (2) other strategies fail to produce a definitive solution.

The first step is encoding relevant cues. We suggest at least four such cues: a) speaker status, b) language cues, c) intonation cues, and d) gesture cues. Speaker status should influence acquiescence in the following ways: the validity of the response should increase as the authority of the speaker increases. This also suggests an informal metric for calculating the status of self and speaker. Language cues may be the most obvious and suggest the type of response that is expected (e.g., "don't you agree"). The second step is inferring possible solutions based on relevant cues. Selecting the PA strategy should occur when other strategies fail to match or when the pragmatic cues are most salient. In both cases, this suggests that the reasoner lacks the knowledge necessary to solve the problem at hand.

The final step is producing a solution. In this case, the reasoner has encoded relevant cues and determined the cued response. This response is given under any of the following conditions: 1) if no other strategy matches, 2) if a strategy produces a solution that is in conflict with the cued response and fails to override this solution, or 3) if the cued response is so highly activated that it overrides all other strategies.

Retrieval

Overview- Retrieval is accessing a previous solution from long-term memory. Retrieval differs from all other proposed strategies in that it is the only strategy that does not create an on-line solution. We include this strategy because solutions, once discovered, can be accessed from memory rather than creating a new solution each time the same problem is presented. Access to solutions will vary by the time interval between discovery and access (recency), the number of times the solution is accessed (frequency), and the degree to which the current problem state is similar to the problem state associated with the solution (fit). Guessing is a loosely constrained form of retrieval in which a response is produced on the basis of inaccurate or irrelevant information.

Outline of Processing in Retrieval- As suggested above, retrieval of previous solutions depends on a variety of factors. The most crucial is the number of possible matches to the current problem. IF there is only one match, then retrieval is simple. Because there are often several possible solutions to a particular problem, in order to retrieve a solution, there must be a mechanism to determine which of these possible solutions will be accessed at any given time, or conflict resolution. We suggest three mechanisms. The first is recency, or the time between when a solution has been discovered and the time it is accessed (Anderson & Lebiere, 1998). The second factor is the frequency of access. The number of times a solution is accessed increases the base activation level of the solution. The higher the base level of activation, the more likely it is that a particular solution will be accessed. The third factor is the degree to which the problem state linked to a solution state is similar to the current problem state, or fit. The degree of fit will determine which of a set of possible solutions is most similar to the current problem state.

Task Characteristics and Situational Niches

The previous section outlined the processing steps for each strategy. As stated in the introduction to the paper, the probability of a particular strategy being used is a function of the processing demands of the strategy and the situational niche. The following section will outline possible task demands, processing demands and how these factors may be related to the application of specific strategies.

The situational niche is similar to Todd & Gigeranzer's (1999) notion of ecological rationality. Both approaches are derived from Simon's (1957) concept of bounded rationality in which reasoning proceeds on the basis of limited information and both are content-sensitive, in that the type of reasoning response is a function of the task demands. While ecological rationality seeks the most adaptive decision/judgment within an open system (i.e., one in which a correct decision is indeterminate), a situational niche represents the current context in which reasoning is occurring and is a match between the processing demands of the system and the task demands of the problem within a closed reasoning system. Thus in a situational niche a correct solution is possible.

The degree to which a problem is familiar will influence the use of a particular strategy. Familiarity is contrasted on two dimensions, the familiarity of the content and experience with a particular problem type. The degree to which content is familiar should increase the probability of knowledge-based strategies. For example, it is a well-documented finding that an invalid syllogism with a believable conclusion is more likely to be accepted as true than a valid syllogism with an unbelievable conclusion (Evans, Barston, & Pollard, 1983). In this case the familiarity of the conclusion may be the most salient element, thus the element most likely to elicit a strategy match. In the case of less familiar materials, for example a syllogism with two "somes" in the premise, a reasoner may rely on a task-specific heuristic to derive a conclusion (Gallotti, Baron, & Sabini, 1986). When given a series of unfamiliar, abstract materials a reasoner may rely strictly on the formal elements of inference. For example given the abstract version of the Wason task, a reasoner may be unable to derive a series of valid conclusions (e.g. modus tollens) but only able to infer modus ponens (Wason & Johnson-Laird, 1972). In each of these cases the familiarity of content may change the problem's situational niche, resulting in different probabilities of matching a given strategy.

In the second sense of familiarity, strategy selection may also depend on the degree of experience a reasoner has with the specific problem type. If a reasoner has a great deal of experience with the specific problem type, they are more likely to retrieve a solution or use the same strategy as used on previous trials. As experience decreases, strategy selection is more likely to be a function of other factors in the situational niche (e.g., presentation format). Strategy selection will be influenced by previous experiences with the problem type and the nature of their outcome associated with the use of priori strategies.

The presentation format also may influence the strategy selected for a particular problem. Presentation formats may be verbal, written, or visual. The type of representation may illustrate or obscure problem characteristics crucial to a correct solution (Larkin & Simon, 1987). Perhaps differences in solutions differ as a function of both the situational niche and the strategy that matches this niche.

In order to illustrate possible links between strategy selection and situational niche, we present the following example. Imagine a transitivity problem in which the basic instructions are given as follows:

> Four people are waiting in line at a movie theater with a new seating policy. The new policy states that in order to allow everyone to see the screen, all patrons have to be seated by height. That is, shorter patrons are seated near the front while taller patrons are seated near the back. Five people, Homer, Marge, Bart, Lisa, and Maggie are going to the theater. Based on their relative height (including hair), place them in proximity to the screen.

Knowledge of the source material may influence the type of strategy used. A reasoner with a great deal of knowledge of the source material (The Simpsons®) may simply retrieve a solution (high content familiarity). One who cannot simply retrieve a solution may use their knowledge to match the task constraints as in a knowledge-based heuristic. In this case, the reasoner may be able to place a few members of the family in order without a transitive inference.

Those with no knowledge of the television show may need to solve the problem using different strategies that may depend on the presentation format. If presented pictorially, the representational format reduces the amount of information in working memory and allows a solution to be derived from scanning the relative heights from the visual array (see Figure 1).

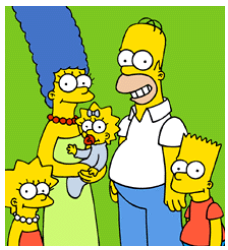


Figure 1

If presented verbally, the task can be simplified by ordering the information to align with task demands. For example, if presented in an ordered format (as in Example 2), a simple scan of relations may allow a solution to be derived. In this case, a matching heuristic may be devised in which the tallest person is only on the left side of the text. From here relations are structured after the tallest has been identified.

Example 2
Homer is taller than Lisa.
Marge is taller than Bart.
Homer is taller than Bart.

Bart is taller than Lisa.
Lisa is taller than Maggie.
Marge is taller than Homer

When terms are randomly distributed in text (i.e., ordering is not aligned with task demands) then each element must be encoded and compared to all other components requiring greater working memory resources (see Example 3). Such a presentation format may best match a token-based strategy, in which pictures are easily represented spatially as a series of tokens.

Example 3
Homer is taller than Lisa. Bart is taller than Lisa. Marge is taller than Bart. Homer is taller than Bart. Lisa is taller than Maggie. Marge is taller than Homer

The previous examples suggest a link between task demands and processing resources. But how do task demands and processing demands produce strategy selection? Differences in processing demands and task demands will lead to differences in the salience of problem elements. Previous strategy use influences the probability that a given strategy will be used. Using this framework, we may be able to explain both inter- and intra-individual change. For information about competition between existing strategies, we will examine data one example of errors in logical reasoning: performance differences based on the presence of familiar content. In verbal strategies, errors in processing are attributed to a failure in applying the appropriate rule to the statement. There are at least two conditions under which a rule is unavailable for processing. **Failure to retrieve** a rule suggests that although the rule is present in long-term memory, it is not retrieved for processing the current information. **Failure to match** a rule is typically explained by the presence of content effects (Braine & O'Brien, 1998). That is, when the content is either familiar or supports an inference beyond that of the statement's form, then rule matching is either suppressed or may match a different rule (Rips, 1994). Although failure to match has been cited as a condition under which abstract rules fail to apply, it is plausible that under these conditions knowledge-based heuristics are more likely to be applied, resulting in slightly different conclusions. Conversely, knowledge-based heuristics rules often fail to fire when given abstract elements (e.g., If A, then B) and are restricted to induced relations (i.e., obligation and permission) (Cheng & Holyoak, 1985; Rips, 1994). In both cases, there are ranges of results that cannot be explained by each unified theory, however, viewing each as a strategy allows the inclusion of the seemingly conflicting empirical findings into a single model.

The strategy approach maintains the explanatory power of each theory while increasing the scope of explanation. The strategy approach accounts for a range of results by suggesting that each strategy possesses distinct processing demands that are likely to match the task demands of specific problem. By allowing multiple approaches within individuals across time the strategy approach is maximally flexible allowing the possibility of explaining differences across tasks and individuals.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Bettman, J.R., Johnson, E.J., Luce, M.F., & Payne, J.W. (1993). Correlation, conflict, and choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 931-951.

Braine, M. & O'Brien, D. (Eds.). (1998). *Mental Logic*. Mahwah, NJ: Erlbaum.

Braine, M., & Rumain, B. (1981). Development of comprehension of "Or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31*, 46-70.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic Reasoning Schemas. *Cognitive Psychology, 17*, 391-416.

Evans, J. St. B. T. (1972). On the problems of interpreting reasoning data: Logical and psychological approaches. *Cognition, 1*, 373-384.

Evans, J. St. B. T. (1989). The psychology of deductive reasoning. London: Routledge.

Evans, J. St. B. T., Barston, J.L., & Pollard, P. (1983). On the conflict between logical and belief in syllogistic reasoning. *Memory and Cognition, 11*, 295-306.

Galotti, K.M., Baron, J., Sabini, J.P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental rules? *Journal of Experimental Psychology: General, 115,* 16-25.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1999). Deductive Reasoning. In J. Spence (Ed.), *Annual Review of Psychology*, Vol. 50 (pp.109-135). Palo Alto, CA: Annual Reviews.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99,* 418-439.

Larkin, J.H., & Simon, H.A., (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65-100.

Rips, L. J. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Searle, J. R. (1984). *Minds, brains, and science*. Cambridge, MA: Harvard

Gigeranzer, G., & Todd, P.M. (1999). *Simple heuristics that make us smart*. Oxford University Press: Oxford, UK.

Wason, P.C., & Johnson-Laird, P.N. (1972). *The psychology of reasoning: Structure and content*. London: Batsford.

Woodworth, R.S., & Sells, S.B. (1935). An atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology, 18*, 451-460.

Commonalities and Distinctions in Featural Stimulus Representations

Daniel J. Navarro and Michael D. Lee
{daniel.navarro, michael.lee}@psychology.adelaide.edu.au
Department of Psychology
University of Adelaide SA 5005, Australia

Abstract

This paper evaluates four featural models of stimulus similarity using data collected for a set of 16 nations. Algorithms are developed for finding stimulus representations, and the important issue of balancing data-fit against model complexity is addressed by using the Geometric Complexity Criterion. Although the data clearly incorporate both common and distinctive features, Tversky's (1977) Contrast Model seems unable to express these regularities in an appropriate manner. However, we show that a new version of the Contrast Model that treats each feature as either being common or distinctive is better able to capture the essential aspects of the similarity judgments.

Featural Representation

A fundamental issue in psychology regards the appropriate manner in which to represent stimuli in a model of human cognition. As argued by Brooks (1991), it is important to constrain representations to those justified by empirical data, and avoid the questionable practice of specifying representations "by hand" . One well-established technique for pinning down mental representation involves measuring the similarity between pairs of stimuli. The assumption underlying this approach is that the decision process involved in judging similarity is a simple one, and thus the data can be considered to reflect the underlying mental representation to a large extent. While this is not without theoretical difficulties (e.g., Goodman, 1972; Goldstone, Medin, & Halberstadt, 1997), it is substantially superior to the alternative approach of hand-tuning representations, which may not reflect human representational structures in any regard. Goldstone's (1999) recent review identifies four main approaches to similarity modeling: geometric, featural, alignment-based and transformational. In this article we discuss current approaches to featural representation, and provide experimental evidence to support a new approach for modeling featural similarity.

The featural approach to mental representation describes an object in terms of the attributes it possesses. Features may be either perceptual or conceptual in nature: for example, a tiger might possess the features "four legged", "orange", and "predatory". The task of deriving featural representations from similarity data can be stated as follows: if n denotes the number of stimuli in the domain, then given an $n \times n$ matrix of similarity judgments $\mathbf{S}$, find a set of m features that explain these judgments. We can denote this set of features by the $n \times m$ feature matrix $\mathbf{F} = [f_{ik}]$, where f_{ik} is 1 if the ith stimulus possesses the kth feature, and 0 if it does not.

Four Models of Featural Similarity

One well-established approach for extracting featural representations from similarity data involves using additive clustering algorithms (e.g, Shepard & Arabie, 1979; Tenenbaum, 1996). The similarity between two stimuli is estimated as the sum of the weights of their common features (i.e., those that they both possess). That is,

$$\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk} + c, \tag{1}$$

where w_k denotes the saliency weight of the kth feature, and c is a positive-valued constant added to all similarity estimates. Thus an m-feature common features representation consists of the feature matrix $\mathbf{F}$, the vector of saliency weights $\mathbf{w} = [w_1, w_2, \ldots, w_m]$ and the additive constant. As noted above, additive clustering relies on a purely common features model. This means that the stimuli become more similar only to the extent that they share features.

An alternative featural model is the distinctive features model, under which similarity is measured according to the differences between the features that stimuli have. This means that if one stimulus has a feature and another does not, they become less similar. This can be written as

$$\hat{s}_{ij} = c - \frac{1}{2}\sum_k w_k f_{ik}\,(1 - f_{jk}) - \frac{1}{2}\sum_k w_k\,(1 - f_{ik})\,f_{jk}, \tag{2}$$

which is identical to the symmetric distance metric proposed by Restle (1959), and closely related

to discrete multidimensional scaling (Clouse & Cottrell, 1996; Rohde, in press).

A general framework that interpolates between these two models is Navarro and Lee's (2001) adaptation of Tversky's (1977; Gati & Tversky, 1984) Contrast Model (TCM), consisting of a weighted sum of the common features similarity (Eq. 1) and the distinctive features similarity (Eq. 2). If we let $0 \leq \rho \leq 1$ denote the weighting given to the common features component, then this model is given by

$$\begin{aligned} \hat{s}_{ij} &= \rho \sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2} \sum_k w_k f_{ik} (1 - f_{jk}) \\ &\quad - \frac{1-\rho}{2} \sum_k w_k (1 - f_{ik}) f_{jk} + c. \end{aligned} \tag{3}$$

The common features model corresponds to the extreme case $\rho = 1$, and the distinctive features model to the other extreme case $\rho = 0$.

However, this model is not the only way of striking a balance between common and distinctive features. Alternatively, we propose a new featural similarity model in which *each individual feature* is declared to be either a common feature (which increases the similarity of pairs of stimuli that share it) or a distinctive feature (which decreases the similarity of a pair of stimuli if one has it and the other does not). This Modified Contrast Model (MCM) is thus:

$$\begin{aligned} \hat{s}_{ij} &= \sum_{k \in CF} w_k f_{ik} f_{jk} - \frac{1}{2} \sum_{k \in DF} w_k f_{ik} (1 - f_{jk}) \\ &\quad - \frac{1}{2} \sum_{k \in DF} w_k (1 - f_{ik}) f_{jk} + c, \end{aligned} \tag{4}$$

where $k \in CF$ implies that the sum is taken over common features, and $k \in DF$ means that only distinctive features are considered.

Psychologically speaking, the argument is that a feature embodies some kind of regularity about the world, which may be that a set of stimuli all have something in common, or alternatively, that two groups of stimuli are in some way different from each other. A common feature instantiates the idea of "similarity within", whereas a distinctive feature represents the notion of "difference between". While it may be the case that the saliency of a feature can change, a commonality does *not* suddenly become a distinction, nor vice versa. In the MCM, the overall balance between commonality and distinctiveness emerges as a function of the relative number and saliency of common and distinctive features, rather than being specified by the parameter ρ, as it is in the TCM. That is, where the TCM assumes that common and distinctive features are weighted during the decision process, the MCM considers the commonality or distinctiveness of a feature to be a regularity inherent in the environment, and so embeds it in the representation itself. In this way, the MCM assumes that featural regularities can be either commonalities or distinctions, but never a bit of both. When a group of stimuli have both common and distinctive aspects, the MCM models these two aspects as two distinct featural regularities.

Model Fitting

It is useful to distinguish between the psychological problem of modeling featural similarity and the numerical problem of finding features (Shepard & Arabie, 1979). The psychological problem is: given a set of features $\mathbf{F}$, how should similarities be estimated? This is the question addressed by the four featural models discussed in the previous section. The numerical problem is a data fitting problem: given a set of similarity data $\mathbf{S}$, and assuming a particular psychological model, what set of features $\mathbf{F}$ most probably gave rise to the data? A variety of approaches have been adopted in fitting the additive clustering model, ranging from mathematical programming (Arabie & Carroll, 1980) to expectation maximization (Tenenbaum, 1996) and stochastic hillclimbing (Lee, in press). The process by which such algorithms operate is relevant to the psychological problem of similarity modeling only inasmuch as we require that they derive good answers to the numerical problem. While none of the above methods is perfect, it is fair to say that each approach performs well enough to allow interpretation and discussion of the derived representations. The representations derived here used a stochastic hillclimbing approach to fit the featural models similar to that adopted by Lee (in press) and Navarro and Lee (2001).

The fitting procedure adopted the Geometric Complexity Criterion (GCC: Myung, Balasubramanian, & Pitt, 2000) as the measure to be minimized by the successful representation. As has been remarked upon previously (e.g., Myung, 2000; Roberts & Pashler, 2000), achieving a good data-fit is not the sole criterion of a good model. Other considerations such as generalizability, simplicity and interpretability must be taken into account. From a quantitative standpoint, one can operationalize the trade-off between fit and complexity in a kind of formal version of Ockham's razor. The GCC is based on the notion that the complexity of a model is given by the number of distinguishable parametric distributions indexed by the model. Informally, this can be thought of as a measure of how many different

similarity matrices could be produced by a given feature structure under all possible choices of saliency weights. The more distributions a model indexes, the more complex it is. This measure is superior to the Akaike Information Criterion (Akaike, 1977) or the Bayesian Information Criterion (Schwarz, 1978), which estimate model complexity by counting the number of free parameters. As Lee (2001) has pointed out, the number of parameters is not a good indicator of the complexity of featural representations, since the way in which features are assigned to stimuli has a considerable influence on model complexity. Furthermore, Navarro and Lee (2001) have demonstrated that common features representations are more complex on average than distinctive features representations. These systematic differences in what Myung and Pitt (1997) call the functional form complexity of a model require a more discriminating measure such as the GCC. The derivation of GCC measures for the four featural models is straightforward, and follows the approach outlined by Navarro and Lee (2001).

Experiment

In order to provide an empirical test of the four featural similarity models, similarity data were collected for a set of 16 nations identified by name. The nature of this domain made it less than satisfactory to present people with a pair of countries and ask them to rate their similarity. It seems likely that this task would be ambiguous, in that the initial reaction of participants may be to ask, "Compared to what?" Even when the similarity between a pair of nations does not need a context, participants are unlikely to bring to this task a pre-existing numerical scale of nation-similarity upon which to rate it. An alternative approach is to provide participants with a context in which to make judgments. The task we used was to present people with a list of four countries, and ask them to select from the list the pair of nations most similar to one another.

Method

Participants Participants in the study were 16 university students (4 male, 12 female) aged 17 to 36, with a median age of 24, who took part in the experiment for course credit.

Materials The list of nations used was: China, Cuba, Germany, Indonesia, Iraq, Italy, Jamaica, Japan, Libya, Nigeria, the Philippines, Russia, Spain, United States, Vietnam, and Zimbabwe. They were chosen to suggest a variety of possible classification schemes (e.g., political system vs geographical location), and involve a variability in overall saliency (e.g., Italy and Germany were better known to most of our participants than Zimbabwe and Nigeria).

Procedure On each trial a list of four countries was displayed (via computer) to the participant, who was asked to pick out the two countries most similar to each other. The 16 nations yield $\binom{16}{2} = 120$ distinct pairs of nations, and a total of $\binom{16}{4} = 1820$ possible lists of four. Given that the similarity ratings are sensitive to all four presented stimuli, it was important to exhaust exactly the set of 1820 quadruples. To that end, the 1820 items were partitioned into 20 subsets of 91 quadruples. Most participants provided responses to one of these subsets, though a few of the participants provided responses to multiple subsets. Since each quadruple involves the presentation of 6 of the 120 nation-pairs, each pair appeared a total of $\frac{1820 \times 6}{120} = 91$ times across the entire data set.

Results

Calculating the mean empirical similarity involved operationalizing the similarity of a pair of countries as the expected probability of selecting that pair in an arbitrary trial containing both stimuli. Using a standard result in Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995, p. 31), if a particular pair is chosen k times out of n (n being 91), then the empirical similarity is given by $s_{ij} = \frac{k+1}{n+2}$. In using the GCC to control model complexity, it is important to know the precision of the similarity values (Lee, in press), which is basically a measure of the extent to which participants agreed in their judgments. Precision is important because more precise data justify more complex models. We estimated the precision to be moderate, by using the full distribution of the similarity judgments. Details of this estimation procedure, as well as that used to calculate similarity values, are given by Navarro (2002).

Using our stochastic hillclimbing algorithms, representations of the nations similarity data that minimized the GCC were found for each of the four similarity models. Of these four representations, the GCC values for the common features, distinctive features, and MCM representations are virtually indistinguishable, with the TCM performing slightly better. However, the qualitative characteristics of these representations are important in terms of model interpretability, and we discuss each in turn.

The best common features representation is shown in Figure 1, and contains seven features that explain 78.1% of the variance in the data. The features are highly interpretable, containing features for western European nations, Caribbean nations, south-

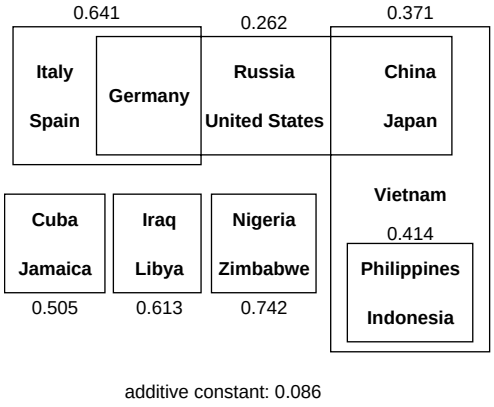


Figure 1: Common features representation of the nations similarity data, accounting for 78.1% of the variance (GCC=41.2). Common features are depicted by rectangles that encompass the nations that possess them.

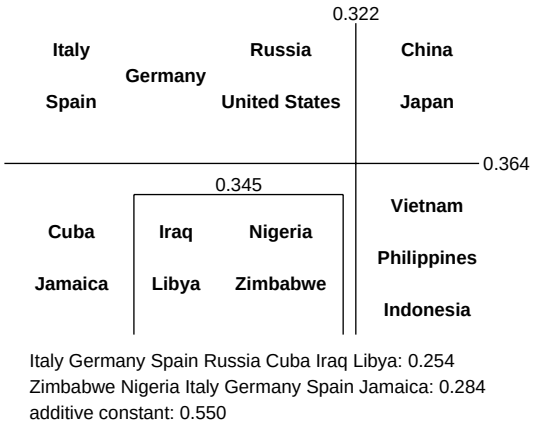


Figure 2: Distinctive features representation of the nations similarity data, accounting for 71.0% of the variance (GCC=41.5). Distinctive features are depicted by lines that partition the stimulus set.

ern African nations and Asian nations. It seems probable that the feature containing only the Philippines and Indonesia reflects perceived cultural and geographical similarities, and that the Iraq–Libya feature reflects the fact that both are Arabic nations and that both have been considered "rogue states". The final feature consists of Germany, Russia, United States, China and Japan, and could be said to denote the "world powers" among the stimulus set.

The best distinctive features representation is shown in Figure 2, and contains five features that explain 71.0% of the variance in the data. Three of the five features have a natural interpretation: one feature separates the African and Middle-Eastern nations from the rest of the world, and another separates the Asian nations from the others. The top-weighted feature in Figure 2 makes an intuitively plausible distinction that might be labeled "developed vs undeveloped". It is interesting to note that the placement of China within the developed nations is equivocal, since the GCC increases only marginally when China is placed in the other category. This makes sense, given China's status as a quickly developing nation. Significantly, however, the remaining two features do not appear to reflect any kind of interpretable structure. From a psychological standpoint, this is highly undesirable, since a central aim of representational modeling is to find interpretable structures in the data.

Table 1 displays the six feature representation derived using the TCM with $\rho = 0.7$, which explains 80.8% of the variance. Since the TCM specifies a balance of common and distinctive features, there is no simple way of depicting this representation graphically. The high ρ value indicates that commonalities are weighted more heavily than differences. All of the clusters in Table 1 appear in either the common features or distinctive features representations, which is not surprising. The features themselves make sense, although it is not easy to see exactly what $\rho = 0.7$ means when providing an overall interpretation. It is noteworthy that the distinctive feature that separated the developing from developed world in Figure 2 does not appear in this representation, despite being the most heavily weighted of the distinctive features. The reason for this may be that the feature does not make sense as anything other than a purely distinctive feature, since any common features component makes one half (either developed or developing) more salient than the other.

The seven feature MCM representation shown in Figure 3 explains 81.2% of the variance, and picks out a number of features from the common features representation: the western Europe, Caribbean, southern Africa and Asian features are all present, as is the "world powers" feature. The feature containing Cuba, Iraq and Libya is interesting, in that the inclusion of Cuba is a marginal case as judged by the GCC. With Cuba included, the feature has a "rogue states" interpretation, whereas without Cuba it would reflect the Arabic nations. Finally, the model also includes the "developed vs developing" regularity from the distinctive features representation. The comparison between this distinctive feature and the "world powers" common feature is also worth making. The fact that these two related but

Table 1: TCM representation of the nations similarity data, employing a common features bias ($\rho = 0.7$), accounting for 80.8% of the variance (GCC=36.9).

Stimuli in Cluster	Weight
Germany, Italy, Spain.	0.682
Nigeria, Zimbabwe.	0.495
China, Indonesia, Japan, Philippines, Vietnam.	0.453
Indonesia, Philippines.	0.374
China, Germany, Japan, Russia, United States.	0.316
Iraq, Libya, Nigeria, Zimbabwe.	0.288
additive constant	0.236

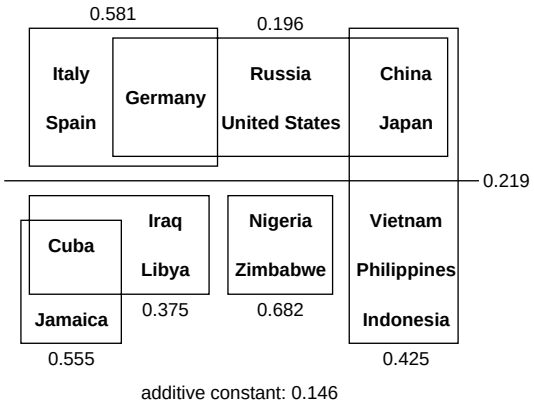


Figure 3: MCM representation of the similarity-condition countries data, accounting for 81.2% of the variance (GCC=41.0)

different features emerge in these different frameworks suggest that people make a distinction between the developed nations and the undeveloped nations, but that there is also something shared by the world powers of the stimulus set. Not all developed nations share this status, so it is appropriate that both features emerge. Moreover, the fact that two nations are developed does not necessarily imply that they are alike, but if one is developed and the other developing, it does make them different, so the "developed vs. developing" feature *should* be a distinctive feature. Correspondingly, two nations *are* alike if they are major world powers, but this does not say anything about their similarity if they are not.

Discussion

As previously noted, a major aim of featural similarity modeling is to capture simple, interpretable regularities present in the data. In this case, the common features representation is easily interpreted, yet the distinctive features representation is not. Furthermore, both the TCM and MCM representations evidence a bias towards commonalities. Overall, therefore, it appears that the participants' judgements were more heavily influenced by common features than distinctive features. Nonetheless, both the TCM and MCM representations include a distinctive features component, suggesting that distinctive features are not irrelevant in the data.

It is worth noting that the "developed vs developing" feature included in the distinctive features and MCM representations is the single most prominent regularity in the data set: when restricted to a single feature, all featural models except the common features model (which is incapable of expressing this feature) yield this feature. It accounts for more variance and has a substantially lower GCC than any other single feature. Though we do not wish to draw overly strong conclusions from a single experiment, the prominence of this regularity calls into question a central assumption of the TCM. As argued in the results section, this "developed vs developing" feature *only* makes sense as a purely distinctive feature. The TCM could only incorporate it by setting $\rho = 0$. However, as observed above, common features are more prominent than distinctive features in the data, and thus a high ρ value is preferred. In doing so, the TCM is able to give a good account for the data in a quantitative sense, but only by discarding this qualitatively important regularity. Notably, since the TCM assumes that all features are subject to the single decision variable ρ, it cannot simultaneously accommodate a commonality and a distinction under *any* parameterization. In comparison, the MCM is not merely capable of accommodating this phenomenon; it is typical behavior for the model.

Future Work

As noted above, a single data set provides only mild evidence: it is important to demonstrate that similar effects may be observed in other data sets. Though space does not permit further analyses here, work currently underway in this area seeks to generalize these findings in precisely this way. Further research could also extend the MCM in order to allow other types of distinctions. A distinctive feature in the current model partitions the stimulus set into two groups, such as "male" and "female". It could be argued that there are regularities that have the hallmarks of a distinction (in that differences-between matter, but similarities-within do not) but involve

more than two groups. For example, the notion that a thing is "animal", "mineral" or "vegetable" could be considered to be distinctive feature that partitions the stimuli into three groups. Moreover, there is a case to be made for representational formalisms that involve both discrete aspects (such as features) and continuous aspects (such as spatial dimensions). Accordingly, another avenue for research would be to pursue hybrid models that involve spatial as well as featural components.

Acknowledgements

This research was supported by Australian Research Council Grant DP0211406, and by a scholarship to DJN from the Australian Defence Science and Technology Organisation. We wish to thank several referees for helpful comments.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of Statistics* (p. 27-41). Amsterdam: North-Holland.

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, *45*(2), 211-235.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139-159.

Clouse, D. S., & Cottrell, G. W. (1996). Discrete multi-dimensional scaling. In *The 18th Cognitive Science Conference* (p. 290-294). San Diego, CA.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology*, *16*, 341-370.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Goldstone, R. L. (1999). Similarity. In R. Wilson & F. C. Keil (Eds.), *MIT encylopedia of the cognitive sciences* (p. 763-765). Cambridge, MA: MIT Press.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, *25*(2), 237-255.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects* (p. 437-447). Indianapolis: Bobbs-Merrill.

Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, *45*, 131-148.

Lee, M. D. (in press). Generating additive clustering models with limited stochastic complexity. *Journal of Classification.*

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190-204.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170-11175.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian aproach. *Psychonomic Bulletin and Review*, *4*(1), 79-95.

Navarro, D. J. (2002). *Representing Stimulus Similarity.* Unpublished phd thesis, University of Adelaide.

Navarro, D. J., & Lee, M. D. (2001). Clustering using the contrast model. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 686-691.

Restle, F. (1959). A metric and an ordering on sets. *Psychometrika*, *24*(3), 207-220.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

Rohde, D. L. T. (in press). Methods for binary multidimensional scaling. *Neural Computation.*

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Thinking by Doing?
Epistemic Actions in the Tower of Hanoi

Hansjörg Neth (NethH@Cardiff.ac.uk)
Stephen J. Payne (PayneS@Cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff CF10 3YG, Wales, United Kingdom

Abstract

This article explores the concept of epistemic actions in the Tower of Hanoi (ToH) problem. Epistemic actions (Kirsh & Maglio, 1994) are actions that do not traverse the problem space toward the goal but facilitate subsequent problem solving by changing the actor's cognitive state. We report an experiment in which people repeatedly solve ToH tasks. An instructional manipulation asked participants to minimize moves either trial by trial or only on the last three of six trials. This manipulation did not have the predicted effect on the trial-by-trial move counts. A second, device manipulation provided some participants with an "exploratory mode" in which move sequences could be tried then undone without affecting the criterion move count. Participants effectively used this mode to reduce moves on each trial, but there was no clear evidence that they used it to learn about the problem across trials. We conclude that there is strong evidence for one sub-type of epistemic action (acting-to-plan) but no evidence for a second sub-type (acting-to-learn).

Introduction

How do we learn to solve a problem? The most popular view within the Cognitive Science community is that we do so by solving the problem. Anzai and Simon's (1979) theory of 'learning by doing' marks a major breakthrough in research on learning through problem solving. They proposed an adaptive production system which mirrored the strategy transformations of a human participant as she solved the Tower of Hanoi (ToH) problem, and in so doing provided the impetus for many subsequent theories of the mechanisms by which problem solving leads to learning (e.g. Klahr, Langley & Neches, 1987).

All learning-by-doing accounts share the assumption that learning about a particular problem occurs as an automatic by-product of problem solving activity. However, in many problem solving situations learning may be more deliberate than the learning-by-doing account implies. We suggest that problem solvers may sometimes orient themselves to *learning goals* rather than *solution goals* (O'Hara & Payne, 1998; Trudel & Payne, 1995).

In relation to the ToH task, this position is encouraged by VanLehn's (1991) re-analysis of the original Anzai & Simon (1979) protocol, in which he notes that the participant was "acting like a scientist" (p. 16) and repeatedly suspended her problem solving activity to acquire new strategic knowledge.

Further general support for a deliberate learning mode nested within problem solving activity can be derived from the work of Kirsh and colleagues (1995, Kirsh & Maglio, 1994), who have explored a distinction between goal-directed *pragmatic actions* and *epistemic actions* whose primary purpose is to improve cognition by changing an agent's computational state. Although epistemic actions are not immediately goal-directed, they may improve subsequent performance through their cognitive effects.

The primary goal of this article is to seek experimental evidence for the use of epistemic actions in problem solving with the ToH puzzle. Identifying epistemic actions in ordinary problem solving activity is difficult, because they are only distinguished by their cognitive motivations and consequences rather than directly observable characteristics (and not all actions that do not successfully move toward the goal are epistemic!). We use two manipulations that may allow participants to utilise epistemic actions, and at the same time facilitate their detection. The first manipulation is *instructional*: participants were asked either to optimize their performance on every problem solving trial, or on trials 4, 5 and 6 of a series of six repeated problems. We hypothesize that delaying the enforcement of the performance criterion will encourage a learning orientation, and the use of epistemic actions, during the early first trials. The second manipulation is to provide *device support* that enables participants to separate pragmatic from epistemic actions. Thus, our computer-based version of ToH allowed participants to switch into an "exploratory mode" in which they could make move sequences that were later undone and were not counted towards the performance criterion.

These twin manipulations allow us to refine Kirsh's formulation of pragmatic and epistemic actions by distinguishing between two kinds of epistemic action: those that have only immediate within-problem effects (*acting-to-plan*) and those that have longer-term cognitive consequences (*acting-to-learn*). If the exploratory mode is used merely as an external support for look-ahead or planning, motivated by questions such as 'Is this a good sequence of moves?', we would regard such usage as acting-to-plan. On the other hand, if additional actions on earlier trials are shown to lead to better problem solving on later trials we would have evidence for acting-to-learn.

To anticipate our conclusions, we find strong support for acting-to-plan, but no decisive support for acting-to-learn.

Method

Participants

Forty-four Psychology undergraduates (with a mean age of 20.7 years) took part in the experiment to receive course credit. Participation was restricted to first year undergraduate students who reported no prior exposure to the task. All participants were familiar with graphical user interfaces and did not suffer from any perceptual or cognitive impairments.

Apparatus

The experiment used a graphical software version of the ToH problem which was programmed in Visual Basic 6 and displayed on a 17" screen. A disk could be transferred between towers by indicating its source and target locations using a drag-and-drop procedure. In case of an illegal move there was an auditory warning signal and the selected disk slid back to its original position. A counter showing the current number of pragmatic moves was displayed in the top right hand corner of the screen.

Materials

Participants had to solve a sequence of 5-disk ToH puzzles in the standard tower-to-tower version. To prevent improvements due to superficial rote memorization we used six simple isomorphs, which were created by systematically switching the source and target towers.

Design

As we wanted to test participants' *spontaneous* use of epistemic actions we did not want to specifically encourage them to explore the problem, but rather provide subtle opportunities that may be used or ignored.

The *instructional manipulation* consisted of two levels. Participants were either instructed to optimize their performance (i.e., minimize the number of pragmatic moves needed to solve the puzzle) on each of several problems, or asked to optimize their performance on the last three of six problems. Hence, whereas the first group of participants was implicitly discouraged from using epistemic actions by the instruction to be *performance oriented* throughout an unspecified number of trials, the second group was presented with an opportunity to be *learning oriented* in the first three of six trials.

The second experimental manipulation consisted in withholding or providing *device-support* for epistemic actions. Two different versions of the device were distinguished:

In the standard *pragmatic moves only* condition, each move of a disk on the screen counted towards the performance criterion of minimizing the number of (pragmatic) moves.

	Device-support	*Instruction*
1.	pragmatic mode only	'minimize on trials 1–6'
2.	pragmatic mode only	'minimize on trials 4–6'
3.	pragmatic+epistemic mode	'minimize on trials 1–6'
4.	pragmatic+epistemic mode	'minimize on trials 4–6'

Table 1: Overview of the two experimental factors and four groups.

In a second *pragmatic plus epistemic moves* condition two different device modes were introduced to the participants. Whilst having to solve the puzzle in so-called "solution mode", participants had the option to switch into an "exploration mode" at any point by pressing and holding down the Shift key. Whereas in both modes disks could be moved in an identical fashion, moves made in exploration mode were not added towards the total performance score and always reversed when switching back into "solution mode" by releasing the Shift key.

Note that the specific design of exporatory mode addresses the difficulty of detecting epistemic moves by effectively creating an *operational definition*: Since participants are aware of the mandatory reversal of all moves made in exploratory mode, entering the mode signals the use of epistemic moves.

One way of characterizing both the instructional and device manipulation is that they do not prevent learning by doing, but provide additional opportunities for *learning by not solving* the puzzle. A combination of both experimental factors yielded the four experimental groups shown in Table 1.

As each participant had to solve a total of six ToH puzzles the experiment employed a mixed design, with *device-support* and *instruction* as between-subjects manipulations and *trial* as a within-subjects factor.

Procedure

Each participant was assigned to one experimental group according to the order of arrival at the laboratory. After reading a generic description of the Towers of Hanoi puzzle participants were introduced to the graphical user interface. To demonstrate that they had understood the task constraints and to familiarize themselves with the user inteface they solved a simple two-disk version of the puzzle.

Participants then received their respective minimization instructions and were told that the experiment normally takes around 45 minutes regardless of their individual performance.

After each successful completion of a problem, participants received a brief message reminding them of their respective minimization instruction before starting the next trial.

On average, participants completed the experiment within 40 minutes.

Predictions

Our primary predictions refer to comparisons between and within experimental groups (rather than assuming a 2x2 factorial design; in particular Group 4 plays a subsidiary role in the study, and will only be analysed in relation to first-order findings).

The main predictions concern the number of pragmatic moves needed to solve the puzzle. As we expect all groups to learn throughout the course of the experiment, we predict a gradual reduction of the mean number of pragmatic moves required to solve the puzzle. This familiar practice effect constitutes the baseline which we expect to be modulated by the experimental factors of instructional goal orientation and device support.

If the instructional manipulation encourages members of Group 2 to invest additional moves in trials 1–3 and this in turn results in better learning, they ought to need fewer moves on trials 4–6. Thus, we predict an *interaction* of instruction and trial for Groups 1 vs. 2.

Next, if participants are spontaneously capable of using the exploratory device mode to improve their performance, Group 3 should need fewer pragmatic moves than Group 1 in all trials. Hence, we predict a *main effect* of device support on the number of pragmatic moves for Groups 1 vs. 3.

Our secondary predictions involve Groups 3 and 4 and address different possible motivations for epistemic moves:

If the exploratory device mode is primarily used for *learning* purposes (acting-to-learn) we should find an instant use of epistemic moves in both Group 3 and 4. If learning actually occurs, the frequency of epistemic moves should decrease over time. If, on the other hand, epistemic moves are used primarily for *online planning* within a trial (acting-to-plan) we expect a more opportunistic use due to the instructional manipulation. In this case, we expect the frequency of epistemic moves in Group 3 and 4 to display an *interaction* over trials.

Finally, if the use of epistemic mode is unselective, and predominantly due to *affordances* created by the design of our device we should find a constant use of epistemic moves throughout all trials and similar usage patterns in Groups 3 and 4.

Results

Numbers of Moves

As all groups were instructed to minimize the number of moves to solve the ToH puzzle our comparative analysis of their performance will be based on the number of pragmatic moves per trial.

Overall learning effects Before we consider the comparisons between individual groups, we will examine the expected overall effects of learning. Figure 1 displays the mean number of pragmatic moves for each group from trials 1 to 6. A mixed ANOVA of pragmatic moves with *group membership* as between-subjects factor and *trial* as within-subjects factor confirms the overall learning effect [$F(5,200)=16.3$, $p=.000$, $MSE=260.4$] and indicates that there were differences between groups [$F(3,40)=5.8$, $p=.002$, $MSE=489.2$], but the interaction between the two factors did not reach significance [$F(15,200)=1.5$, $p=.13$, $MSE=260.4$].

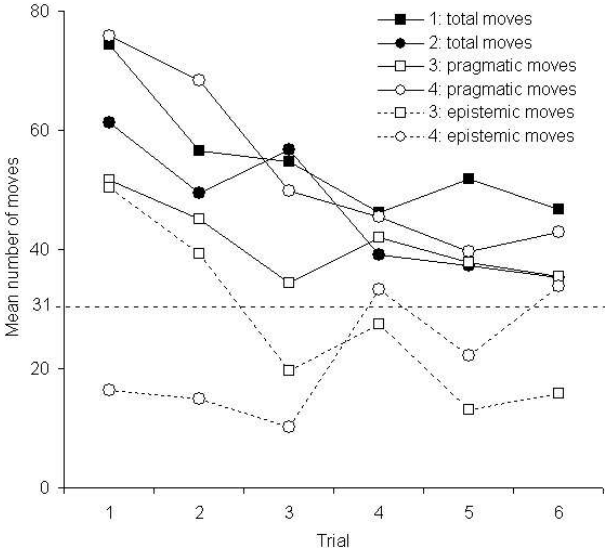


Figure 1: Mean number of moves for each of the four groups on each of six trials. For Groups 1 and 2 the number of pragmatic moves corresponds to the number of total moves, whereas Groups 3 and 4 had the option of making epistemic moves in addition to pragmatic moves. (Note: The minimum possible number of pragmatic moves to solve the task is 31.)

Whether an individual group has significantly improved over time can be assessed by conducting simple effect tests within groups by trial. These show that Groups 1, 2 and 4 significantly reduced their number of pragmatic moves over the course of the experiment. The means for Group 3 were low even at the first trials, suggesting that it did not improve significantly because it consistently performed at a high level.

On trials 4–6, in which all groups were instructed to minimize the number of pragmatic moves, both Group 2 and 3 outperformed Group 1 by taking fewer pragmatic moves ($p=.003$ and .008 respectively).

Thus, the overall results seem promising: With respect to the performance criterion Groups 1, 2 and 4 improved over time and Groups 2 and 3 managed to solve the ToH puzzle in fewer moves than Group 1 in the second half of the experiment.

Effects of Instruction (Groups 1 vs. 2) Our first specific prediction concerned Groups 1 and 2, neither of which had the epistemic device mode at their disposal, but differed in their instructions: Whereas Group 1 was instructed to minimize the number of moves in each trial, Group 2 was only asked to optimize their performance in

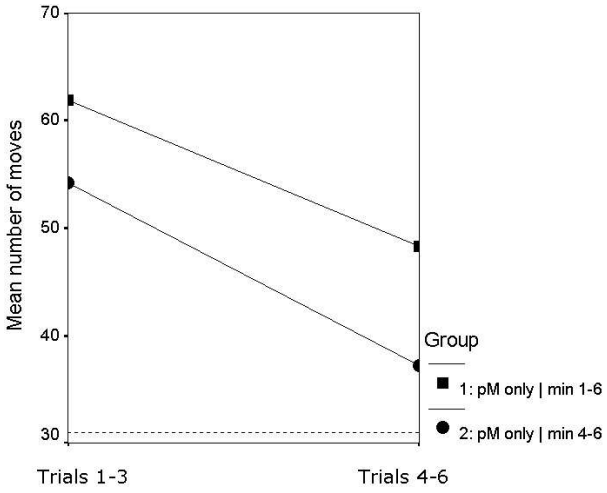


Figure 2: Mean number of moves for Groups 1 and 2 on trials 1–3 and 4–6. (Note: The minimum possible number of moves is 31.)

the last three of six trials. As the scope of this experimental manipulation juxtaposed trials 1–3 with trials 4–6 it is appropriate to collapse the data across each triple of trials by computing the respective means.

Figure 2 shows the mean number of moves for both groups on trials 1–3 and 4–6. It illustrates that there is no hint of the predicted crossover interaction [$F(1,20)=.44$, $p=.51$, $MSE=72.6$]. Instead, the predicted learning effects over both test halves [$F(1,20)=35.4$, $p=.00$, $MSE=72.6$] are combined with an unexpected main effect of group [$F(1,20)=8.1$, $p=.01$, $MSE=119.0$].

While successfully predicting that Group 2 would use fewer moves on trials 4–6 [$F(1,20)=8.9$, $p=.01$] it is immediately obvious that this advantage in performance cannot be attributed to its members using epistemic actions in the initial trials: they clearly have not invested *additional* moves on trials 1–3.

One plausible, if rather annoying explanation for this pattern of data, is that, by an accident of assignment, Group 2 might comprise more able problem solvers than Group 1. (We will briefly consider an alternative account below.)

Effects of Device Support (Groups 1 vs. 3) Groups 1 and 3 shared the same instructions (to minimize the number of moves on each trial) but differed in the options provided by the user interface (device). Specifically, members of Group 3 had the "exploratory mode" at their disposal which supported the use of epistemic moves.

In the overview of the number of pragmatic moves of all four groups we have already established that Group 3 performed better than Group 1 on trials 4–6. A mixed ANOVA of pragmatic moves by group membership and trial confirms the predicted main effect of group over all six trials [$F(1,20)=18.0$, $p=.000$, $MSE=359.9$]. Thus, Group 3 *consistently* performed better than Group 1 with respect to the criterion.

To interpret this difference in performance the number of epistemic moves of Group 3 has to be taken into account as well. (The number of epistemic moves is represented by dotted lines in Figure 1).

If we add the epistemic moves carried out by Group 3 to their pragmatic moves, Group 3 used more total moves on average than Group 1 (mean total moves=68.7 and 55.1 respectively), but this difference is not statistically significant [$F(1,20)=.11$, $p=.12$, $MSE=2165.3$].

This demonstrates that Group 3 spontaneously managed to use the device-supported option of epistemic moves to improve their performance with respect to the criterion. However, it leaves open exactly *why* and *how* members of Group 3 used the exploratory mode. We will address these issues after assessing the effects on solution latency.

Solution Times

Although participants had been told that the total experiment took a standardized length of time—hence could not assume that by being quick or slow they would alter the overall duration of their experimental session—their latencies to solve a problem can be used as an alternative indicator to assess their performance.

Overall effects As one might expect solution latencies over the course of the experiment decreased for all groups. An overall mixed ANOVA on the effects of group membership and trial on the total time required to solve each task yields a main effect of trial [$F(5,200)=16.9$, $p=.000$, $MSE=13352.7$] but no differences between groups. However, a significant interaction of the two factors [$F(15,200)=1.8$, $p=.038$, $MSE=13352.7$] drew attention to the possibility that different groups might have exploited time selectively to optimize their performance.

Subsequent simple main effect tests confirm that while the total solution times of Groups 1 and 3 significantly decreased over repeated trials, this was not the case for Groups 2 and 4. This suggests that the instructional manipulation had a selective effect on solution time, and in particular raises the possibility that the improved performance of Group 2 over Group 1 on trials 4–6 was caused in part by Group 2 exerting greater effort on these trials.

Effects of Instruction (Groups 1 vs. 2) The suggestion that Group 2 outperformed Group 1 in trials 3–6 by exerting extra effort (rather than the hypothesized investment of epistemic moves) is supported by an analysis of *move rates*, i.e. the number of moves made per second. Figure 3 shows the mean move rates of Group 1 and 2 over both test halves. A corresponding ANOVA on move rate by group and test half yields a highly significant interaction [$F(1,20)=18.3$, $p=.000$, $MSE=.002$].

If we interpret an increase in move rate (as seen in Group 1) as signalling the necessity of less effort per individual move, the *absence* of an increase in Group 2 suggests that its members invested relatively more effort in the second half of the experiment.

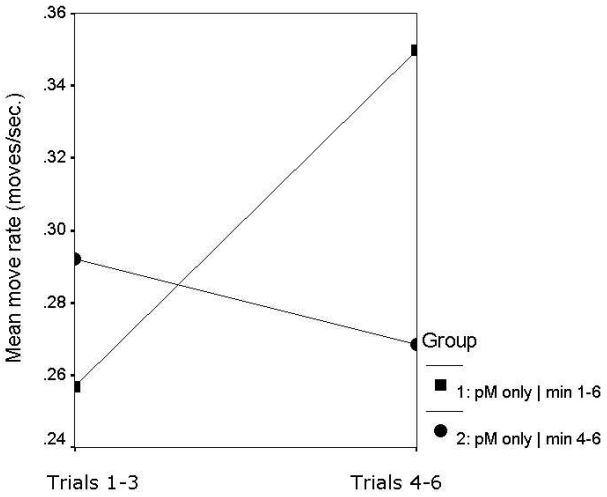


Figure 3: Mean move rates for Groups 1 and 2 on trials 1–3 and 4–6.

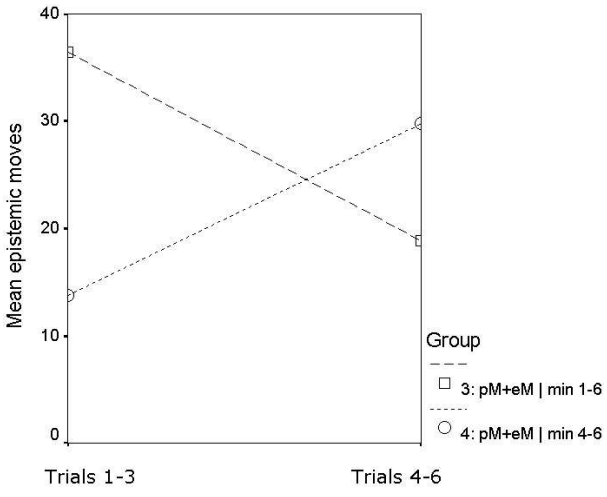


Figure 4: Mean number of epistemic moves for Groups 3 and 4 on trials 1–3 and 4–6.

Effects of Device Support (Groups 3 and 4)

One of the questions raised above was: *How* did Group 3 use epistemic moves to outperform Group 1? As the total solution times for Groups 1 and 3 did not differ [$F(1,20)=.38$, $p=.54$, $MSE=66525.2$] recourse to latency data does not resolve this issue.

Although the present experiment does not allow us to answer questions about possible causes and effects of device-supported epistemic moves conclusively, we can provide tentative evidence for some of the related issues:

- Did the use of epistemic moves actually lead to *better learning*? The fact that Group 3 *continued to use* epistemic moves until the last trials suggests that they probably did not learn more about the ToH puzzle than Group 1, but used the exploratory device mode as a tool to optimize their performance. However, our design allows for the alternative explanation that the continued use of epistemic moves might have been due to a conservative strategy.

- Did the use of epistemic moves become *more effective over time*? An index of the effectiveness of each epistemic move can be computed by dividing Group 3's mean savings of pragmatic moves (compared with Group 1) by the number of epistemic moves invested on each trial. As the six corresponding ratios (0.5, 0.3, 1.0, 0.2, 1.1, 0.7) do not show any obvious trend, we conclude that the use of exploratory mode did not become more effective over time.

- Were epistemic moves used *to learn* or *to plan*? Even without evidence for superior learning due to device-support of epistemic moves we can address the question of participants' *motivation* to use exploratory mode by comparing the usage patterns of Group 3 and 4. Figure 4 shows a clear interaction of group membership and test half on the mean number of epistemic moves [$F(1,20)=.5.4$, $p=.03$, $MSE=572.7$]. The same pattern can be observed when we consider the relative frequency of epistemic moves: Whereas the use of epistemic moves for Group 3 decreases over time, members of Group 4 increase their use of epistemic moves in the second half of the experiment. This suggests that exploratory mode was used opportunistically to meet the instructional constraints, i.e., for online-planning (acting-to-plan) on the current trial, rather than as a prospective investment into learning (acting-to-learn).

- Were epistemic moves used because they were available, i.e., was the usage of exploratory mode simply a task-demand like artifact, prompted by our device manipulation? The strategic use of epistemic moves observed by Group 4 attenuates this concern. Rather than being a mere device affordance, exploratory mode was used selectively to achieve online planning.

Discussion

Participants in the exploratory mode conditions spontaneously and effectively used the device-support to achieve a performance criterion, and in so doing they demonstrated capability for using epistemic actions to improve immediate performance.

However, both the observed unwillingness to invest additional moves in early trials and the selectivity of usage patterns suggest that participants were only willing to invest epistemic moves when they stood to gain an immedate benefit from so doing. There was no clear sign of increased learning through use of exploratory mode or willingness to use epistemic moves for learning purposes. Instead, the selective use suggests that epistemic actions were mainly serving the function of look-ahead (acting-to-plan) rather than learning prospectively about the ToH task (acting-to-learn).

Our instructional manipulation did not have the predicted effect. This may have been due to an unfortunate mismatch between the experimental groups, or it may be that our initial hypotheses about an interesting distinction between problem-solving and learning orientations are unfounded, at least for Tower of Hanoi. Perhaps more likely still is the possibility that our instructional manipulation was too subtle to invoke any change in orientation.

In Kirsh's writing on epistemic actions and related themes, which was one of the sources of inspiration for the current study, an additional concept is introduced by contrast with goal-directed behaviour, namely "complementary strategies" (Kirsh, 1995, 1996). It is not clear to us how precise a distinction Kirsh is promoting between "complementary" and "epistemic": indeed there is a hint in his writings of mere terminological evolution. Nevertheless, we suggest that there is an important distinction that might be sketched. As defined above, epistemic actions have their effect by modifying cognitive structures in the actor. By contrast, consider such example complementary strategies as moving coins in order to count them, or marking numbers in order to add them (Kirsh, 1995, 1996; Neth & Payne, 2001). Such operations work by modifying the problem so as to be more compatible with cognitive capabilities, rather than changing the cognitive state of the actor. We agree with Kirsh that complementary strategies of this kind are ubiquitous in human behaviour.

The case for ubiquity is less clear for epistemic actions. In this article we have sketched a distinction between two kinds of epistemic actions, actions-to-learn and actions-to-plan. We have found evidence for the latter, but none for the former.

One reason that acting-to-learn may be relatively less common than complementary strategies and than acting-to-plan, is the success of learning-by-doing. A second reason, ironically, is that acting may sometimes compete with learning. As shown by O'Hara and Payne (1998), and Trudel and Payne (1995), internalising problem solving activity and planning (doing more mental look-ahead or reflection) can itself increase learning in a problem solving context. For example, when exploratory learners had their interactions with a digital watch rationed, they learned more successfully how to use the watch (Trudel & Payne, 1995).

Despite these arguments, we are confident that, as defined in the introduction, actions-to-learn (i.e. actions that are *not* intended to solve the current problem but only to learn about the current problem) are indeed an important aspect of problem solving and learning. However, such actions may be less widely and spontaneously available and harder to study in conventional puzzle-solving domains.

Turning from the philosophical to the practical, one very concrete contribution of the current article is the idea of incorporating an exploratory mode, with instant undo, into the user interface. Undo functions are, of course, well-established and universally acknowledged contributors to device usability (although some thorny technical design issues are still debated). What is novel about our exploratory mode, we believe, is that it guarantees a very rapid return to particular user-chosen system states. It can accomplish this because the user makes a specific *commitment* to undoing subsequent actions. Although this might seem counter-indicated in mundane HCI contexts, we suggest related designs may be worth pursuing in any domain where people stand to benefit from "thinking by doing".

Acknowledgments

We would like to thank Will Reader for helpful comments and suggestions on an earlier draft. This research was supported by ESRC Research Studentship Award No. R00429934325 to HN.

References

Anzai, Y., & Simon, H.A. (1979). The theory of learning by doing. *Psychological Review*, *86*, 124–140.

Kirsh D. (1995). Complementary Strategies: Why we use our hands when we think. In J.D. Moore & J.F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4*, 415–452.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Klahr, D., Langley, P. & Neches, R. (1987). *Production System Models of Learning and Development*. Cambridge, MA: MIT Press.

Neth, H. & Payne, S.J. (2001). Addition as interactive problem solving. In J.D. Moore, & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 698–703). Mahwah, NJ: Lawrence Erlbaum.

O'Hara, K.P., & Payne, S.J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology*, *35*, 34–70.

Trudel, C.I., & Payne, S.J. (1995). Reflection and goal management in exploratory learning. *International Journal of Human-Computer Studies*, *42*, 307–339.

VanLehn, K. (1991). Rule acquisition events in the discovery of problem-solving strategies. *Cognitive Science*, *15*, 1–47.

Bayesian Learning at the Syntax-Semantics Interface

Sourabh Niyogi (niyogi@mit.edu)
Massachusetts Institute of Technology
Cambridge, MA USA

Abstract

Given a small number of examples of scene-utterance pairs of a novel verb, language learners can learn its syntactic and semantic features. Syntactic and semantic bootstrapping hypotheses both rely on cross-situational observation to hone in on the ambiguity present in a single observation. In this paper, we cast the distributional evidence from scenes and syntax in a unified Bayesian probablistic framework. Unlike previous approaches to modeling lexical acquisition, our framework uniquely: (1) models learning from only a small number of scene-utterance pairs (2) utilizes and integrates both syntax and semantic evidence, thus reconciling the apparent tension between syntactic and semantic bootststrapping approaches (3) robustly handles noise (4) makes prior and acquired knowledge distinctions explicit, through specification of the hypothesis space, prior and likelihood probability distributions.

Learning Word Syntax and Semantics

Given a small number of examples of scene-utterance pairs of a novel word, a child can determine both the range of syntactic constructions the novel word can appear in and inductively generalize to other scene instances likely to be covered by the concept represented (Pinker 1989). The inherent semantic, syntactic, and referential uncertainty in a single scene-utterance pair is well-established (c.f. Siskind 1996). In contrast, with multiple scene-utterance pairs, language learners can reduce the uncertainty of which semantic features and syntactic features are associated with a novel word.

Verbs exemplify the core problems of scene-utterance referential uncertainty. Verbs selectively participate in different alternation patterns, which are cues to their inherent semantic and syntactic features (Levin 1993). How are these features of words acquired, given only positive evidence of scene-utterance pairs?

The *syntactic bootstrapping* hypothesis (Gleitman 1990) is that learners exploit the distribution of "syntactic frames" to constrain possible semantic features of verbs. If a learner hears /glip/ in frames of the form /S glipped G with F/ and rarely hears /S glipped F into G/, the learner can with high confidence infer /glip/ to be in the same verb class as /fill/ and have the same sort of argument structure. A different distribution informs the learner of a different verb class. Considerable evidence has mounted in support of this hypothesis (c.f. Naigles 1990, Fisher et al 1994). In contrast, the *semantic bootstrapping* hypothesis (Pinker 1989) is that learners use what is common across scenes to constrain the possible word argument structures. If a learner sees a liquid undergoing a location change when /S glipped F/ is uttered, then /glip/ is likely to be in the same verb class as /pour/ and have the same sort of meaning.

Both hypotheses require the distribution of cross-situational observations. Prior accounts to model word learning have either ignored the essential role of syntax in word learning (Siskind 1996, Tenenbaum and Xu 2000), or require thousands of training observations (Regier et al 2001) to enable learning. In this paper we present a Bayesian model of learning the syntax and semantics of verbs that overcomes these barriers, by demonstrating how word-concept mappings can be achieved from very little evidence, where the evidence is information from both scenes and syntax.

Bayesian Learning of Features

We illustrate our approach with a Bayesian analysis of a single feature. On some accounts, verbs possess a *cause* feature which may be valued `1`, `*`, or `0` (Harley and Noyer 2000); depending on the value of the *cause* feature, the verb may appear in frame F1, F0, or both:

`1`	Externally caused - Ex: touch, load
	F1: He touched the glass.
	F0: *The glass touched.
`*`	Externally causable - Ex: break, fill
	F1: He broke the glass.
	F0: The glass broke.
`0`	Internally caused - Ex: laugh, glow
	F1: *He laughed the children.
	F0: The children laughed.

Assuming this analysis, learners who hear utterances containing a novel verb, not knowing the value of its *cause* feature, must choose between 3 distinct hypotheses H_1, H_*, and H_0. Clearly, one utterance cannot uniquely determine the value of the feature: if learners hear F1 (/S Ved O/), the feature sup-

ports H_1 or H_*; similarly, if learners hear F0 (/O Ved/), the feature may be H_0 or H_*. Two utterances cannot determine the feature uniquely either. Learners might receive both F1 and F0, supporting H_* uniquely. But they may also accidentally receive 2 utterances of the same form (F0, F0 or F1, F1), thus not resolving the ambiguity. If learners received 6 utterances of the same form F0 or F1, however, then there is overwhelming support for H_0 or H_1 respectively, and H_* seems far less likely.

A Bayesian analysis renders the above analysis precise and quantitative. Knowledge is encoded in three core components: (1) the structure of the hypothesis space $\mathcal{H}$; (2) the prior probability $p(H_i)$ on each hypothesis H_i in $\mathcal{H}$, before learners are provided any evidence; (3) the likelihood of observing evidence X given a particular H_i, $p(X|H_i)$. Given evidence $X = [x_1, \ldots, x_N]$ of N independent observations, by Bayes' rule the posterior probability of a particular hypothesis H_i is:

$$p(H_i|X) = \frac{\prod_{j=1}^{N} p(x_j|H_i)p(H_i)}{p(x_1, \ldots, x_N)} \quad (1)$$

signaling the support for a particular hypothesis H_i given evidence X.

In this case, x_j is the observation of a syntactic frame (F0 or F1), and X is a distribution of syntactic frames. One simple prior probability model $p(H_i)$ has each of the 3 hypotheses are equally likely, encoding that a verb is equally likely to be of the /touch/, /laugh/ or /break/ class:

$$p(H_1) = p(H_*) = p(H_0) = \frac{1}{3} \quad (2)$$

and a likelihood model $p(x_j|H_i)$ encoding how likely we are to observe frames F0 or F1 for the 3 different feature values of *cause*:

$$\begin{aligned} p(x_j = F1|H_1) &= .95 \quad p(x_j = F0|H_1) = .05 \\ p(x_j = F1|H_*) &= .50 \quad p(x_j = F0|H_*) = .50 \\ p(x_j = F1|H_0) &= .05 \quad p(x_j = F0|H_0) = .95 \end{aligned} \quad (3)$$

The above likelihood model says that when a verb has *cause*=1, we expect frames of the form /S Ved O/ 95% of the time; when a verb has *cause*=0, we expect /O Ved/ 95% of the time; when a verb has *cause*=*, we expect both syntactic frames.

Both the prior probability model and likelihood model are *stipulated*, encoding a learner's prior knowledge of grammar. Given these probability models, this allows for explicit computation of the support of each hypothesis. Suppose a learner receives F0. Then the support for each of the 3 hypotheses may be computed to be:

$$\begin{aligned} p(H_1|F0) &= \frac{(.05)(.33)}{(.05 + .50 + .95)(.33)} = .033 \\ p(H_*|F0) &= \frac{(.50)(.33)}{(.05 + .50 + .95)(.33)} = .333 \\ p(H_0|F0) &= \frac{(.95)(.33)}{(.05 + .50 + .95)(.33)} = \mathbf{.633} \end{aligned} \quad (4)$$

Any number of situations may be analyzed as such:

	Evidence X	$p(H_1\|X)$	$p(H_*\|X)$	$p(H_0\|X)$
1	F0	.033	.333	**.633**
2	F0, F0	.002	.216	**.781**
3	F0, F0, F0, F0, F0, F0	2e-8	.021	**.979**
4	F0, F1	.137	**.724**	.137
5	F0, F1, F0, F1, F0, F1	.007	**.986**	.007
6	F0, F1, F1, F1, F1, F1	**.712**	.288	5e-6

When only F0 is given as evidence (situation 1), while both H_0 and H_* are consistent with the observation, H_0 is nearly twice as likely. However, with 2 observations of F0 (situation 2) or 6 observations (situation 3), it is increasingly likely that H_0 is the correct hypothesis. With both F0 and F1 as evidence (situation 4), in contrast, H_* is far more likely; with more evidence (situation 5), it becomes more so. Finally, if the first frame is a "noise" frame and followed by 5 representative frames of F1 (situation 6), then H_1 is more likely instead.

Given this framework, just one or two observations is sufficient to make an informed judgement. Note that each additional observation increases certainty, and noise is handled gracefully.

Modeling Semantic Bootstrapping

In this section, we extend the single feature analysis to multiple features, where each feature represents information from scenes (from any modality, whether perceptual, mental, etc.). Setting aside verbal aspect, we may model possible verb meanings as a set of M features, where each feature represents a predicate on one or more of the arguments of the verb. For example, a set of single argument predicates might include:

moving(x), rotating(x), movingdown(x),
supported(x), liquid(x), container(x)

specifying the perceived situation about the argument of the verb (e.g. if it is moving, or moving in a particular manner, etc.) while a second set of two-argument predicates might specify the relationships between arguments, given that this is an externally caused (*cause*=1) event:

contact(x, y), support(x, y), attach(x, y)

Using these predicates, an idealized (partial) lexicon might contain the following word-concept mappings:

	cause	One arg x	Two arg x, y
/lower/	1	1*11**	11*
/raise/	1	1*01**	11*
/rise/	0	1*0***	
/fall/	0	1*1***	

specifying, in linear order, the value of each of the one and two-argument predicates above, e.g. that /lower/ has *cause*=1, moving(x)=1, rotate(x)=*, movingdown(x)=1, etc. – and thus its concept covers externally-cased motion events where an agent moves a theme downwards through supported contact. The verb /raise/ is nearly identical except it has movingdown(x)=0, while /fall/ and /rise/ involve internally-caused motion (*cause*=0) and do not

specify any two argument predicates. The values of * for the 4 rotating(x), liquid(x), container(x), and attach(x,y) predicates signal that these features are irrelevant to the verb's concept. Perception of a scene amounts to evaluating these predicates; scenes may or may not fall under the verb concept, conditioned on the values of these predicates. The presence of q of "irrelevant" features valued as * implies 2^q possible scenes consistent with the concept.

Given a hypothesis space of possible verb concepts formed by M of these sorts of predicates, the task of learning a verb's meaning given N observations $X = [x_1 \ldots x_N]$ of scenes, is to determine which of the 3^M possible concepts is the most likely. Just as before, a Bayesian model does so by computing the posterior probability distribution $p(H_i|X)$ over concepts, given a prior distribution on hypotheses $p(H_i)$ and a likelihood distribution of generating a particular x_j example given H_i:

$$p(x_j|H_i) = \begin{cases} \frac{1}{2^q} & \text{if } x_j \in H_i \\ 0 & \text{otherwise} \end{cases} ; p(H_i) = \frac{1}{3^M} \quad (5)$$

We can use Bayes' rule (Eq (1)) to compute the likelihood of any hypothesis given N independent examples. Intuitively, the above likelihood model says that out of the 2^q possible scenes that might fall under the concept H_i, all of them are equally likely; likewise, the prior probability model holds that all of the 3^M concepts are equally likely.

Consider a reduced hypothesis space where $M = 3$:

q	Concepts
0	`000, 001, 010, 011, 100, 101, 110, 111`
1	`00*, 01*, 10*, 11*, 0*0, 0*1, 1*0, 1*1, *00, *01, *10, *11`
2	`0**, *0*, **0, 1**, *1*, **1`
3	`***`

Given any distribution of scenes X, we can directly compute the posterior probability $p(H_i|X)$ of any of the 27 different concepts. Four are shown here, of increasing generality from a very specific concept (H_{000}) covering only one scene (`000`) to the most general concept H_{***} covering 2^M possible scenes:

	Observation X:	H_{000}	H_{00*}	H_{0**}	H_{***}
1	000	**.30**	.15	.07	.03
2	000, 000, 000	**.70**	.09	.01	.001
3	000, 001	.00	**.64**	.16	.04
4	000, 001, 000	.00	**.79**	.10	.01
5	000, 001, 000, 001, 000	.00	**.94**	.03	.001
6	000, 101, 010, 111, 000	.00	.00	.00	**1.0**

A single scene observation `000` is explained by all 4 hypotheses (situation 1) in a graded fashion. However, with 3 repeated observations (situation 2), most of the mass is concentrated on H_{000}. When scene observations require abstracting away irrelevant features, the more specific concepts must be discarded in favor of more general concepts (situation 3 vs 6). Each example consistent with the general concept further reduces ambiguity over the possible concepts (situation 4 vs 5).

Modeling Syntactic Bootstrapping

In this section, we demonstrate a Bayesian model of how the distribution of syntactic frames, as envisioned by Gleitman (1990), may be used to determine the semantic features of a verb. To do so, we introduce a new notion of *semantic agreement*, wherein features of a lexical head must *agree* with its complement. Consider the following idealized lexicon:

/fill/	fig: [0] con: [1]	/into/	fig: [1]
/pour/	fig: [1] liq: [1]	/with/	fig: [0]
/load/	fig: [*]	/glass/	con: [1]
		/water/	liq: [1]

A lexical head /fill/ *agrees* with a complement of /a glass with water/ but not with /water into a glass/, because the lexical head and its complement have a value 1 along the *fig* dimension. Likewise, a lexical head /pour/ agrees with a complement of /water into a glass/ but not /a glass with water/, because of the opposite value of *fig*. Finally, a lexical head such as /load/, because $*$ agrees with 0 and 1, accepts both complements. Thus, both /load the wagon with hay/ and /load hay into the wagon/ are valid derivations. A large number of verb classes can be seen to pattern into three classes along different feature dimensions in this way (Nomura et al 1994).

Any number of feature dimensions may be hypothesized, and may include selectional features, such as /fill/ requiring a container (con:[1]) or /pour/ requiring a liquid (liq:[1]) as its complement.

Suppose a learner hears /S glipped a glass with water/. The features of the novel verb /glip/ are unknown and the features of its complement /a glass with water/ are known. For the *fig* feature dimension of /glip/, there are 3 possible values, with 3 corresponding hypotheses H_0, H_1, H_*. As before, one observation is insufficient to infer H_0, as H_* is also possible. The following likelihood model for an unknown verb feature value V and the feature value of its complement C agreeing can be used for each feature dimension (fig, loc, con, etc.) to compute a probability distribution over the H_i:

$p(V, C)$	$V = 0$	$V = 1$	$V = *$
$C = 0$	.22	.01	.11
$C = 1$	.01	.22	.11
$C = -$	.11	.11	.12

Intuitively, the above says that with high probability, V and C agree, and with low probability (i.e. .01), they do not agree. The above joint distribution encodes both the prior distribution on V and the conditional distribution $p(C|V)$:

$$p(V = 0) = p(V = 1) = p(V = *) = \frac{1}{3} \quad (6)$$

$$\begin{array}{ll} p(C = V|V = 0 \text{ or } 1) = .65 & \\ p(C \neq V|V = 0 \text{ or } 1) = .03 & p(C = 0, 1|V = *) = .32 \\ p(C = *|V = 0 \text{ or } 1) = .32 & p(C = *|V = *) = .35 \end{array} \quad (7)$$

Given an assumption of perfect knowledge of the feature values of the complement, over multiple observations, the distributional evidence X in support of the 3 hypotheses can be readily evaluated. We can test how different distributions of syntactic frames correctly yield different probability distributions of a verbs syntactic and semantic features; this is thus a Bayesian model of Gleitman's (1990) "syntactic bootstrapping". Suppose a learner gets 4 syntactic frames of /glip/, all of the form /S glipped O with Z/. This is equivalent to having 4 perfect observations of fig:[0], which we annotate as $X = \texttt{0000}$. Then the likelihood $p(X|V)$ and posterior probability $p(V|X)$ of the 3 possible hypotheses can be evaluated directly via Bayes' rule:

Likelihood $p(X\|V)$	Posterior $p(V\|X)$
$p(X\|V=0) = (.65)^4$	$p(V=0\|X) = .941$
$p(X\|V=1) = (.03)^4$	$p(V=1\|X) = .000$
$p(X\|V=*) = (.32)^4$	$p(V=*\|X) = .059$

This is shown below, along with other distributions of syntactic frames:

Sit	Utterances (X)	$V=0$	$V=1$	$V=*$
1	4 /S Ved O with Z/ (0000)	**.941**	.000	.059
2	4 /S Ved O/ (****)	.292	.292	**.416**
3	2 /S Ved O with Z/, 2 /S Ved O into Z/ (0011)	.032	.032	**.936**
4	2 /S Ved O/, 2 /S Ved O with Z/ (**00)	**.769**	.000	.230
5	23 /S Ved O with Z/ 10 /S Ved O/	**1.00**	.000	.000
6	23 /S Ved O with Z/ 5 /S Ved O into Z/ 10 /S Ved O/	**.960**	.000	.040

With only 4 examples, the uncertainty of the value of the feature V is rapidly reduced (situations 1-4). As the number of examples increases (situation 4 vs 5), the evidence supports "all-or-none" or "rule-like" behavior, even with a significant number noisy frames (situation 5 vs 6).

Modeling Integrated Syntactic and Semantic Bootstrapping

We now integrate the two forms of bootstrapping described above, where given a distribution of both scenes and syntactic frames, a probability distribution over concepts consistent with both sources of evidence is determined. Consider the following possible syntactic frames:

Utterance	**u**	Attention
/Glipping!/	***	-
/S glipped water from a glass/	1**	W
/S glipped water into a glass/	1**	W
/S glipped water/	***	W
/S glipped a glass with water/	0**	G
/S glipped a glass/	***	G

and perceptually-derived semantic features of scenes:

Scene **s**	*Description/Semantic Features*
pour-fill	Person pouring water into a glass, filling it
G_{001}	Glass: Manner: None (0) State: Full (1)
W_{110}	Water: Manner: Pouring (1) State: None (0)
splash-fill	Person splashes water into a glass, filling it
G_{001}	Glass: Manner: None (0) State: Full (1)
W_{120}	Water: Manner: Splashing (2) State: None (0)
spray-fill	Person sprays water into a glass, filling it
G_{001}	Manner: None (0) State: Full (1)
W_{130}	Manner: Spraying (3) State: None (0)
pour-empty	Person pouring water out of glass, emptying it
G_{002}	Manner: None (0) State: Empty (2)
W_{110}	Manner: Pouring (1) State: None (0)
splash-empty	Person splashes water out of glass, emptying it
G_{002}	Manner: None (0) State: Empty (2)
W_{120}	Manner: Splashing (2) State: None (0)
pour-none	Person pouring some water into a glass
G_{000}	Manner: None (0) State: None (0)
W_{110}	Manner: Pouring (1) State: None (0)
spray-none	Person sprays water into a glass
G_{000}	Manner: None (0) State: None (0)
W_{130}	Manner: Spraying (3) State: None (0)

where features are ordered as:

fig, manner-of-motion, change-of-state

for each utterance **u** and scene possibility **s**. The subscripts on G and W annotate the observation of that argument for each of the 3 dimensions.

We may describe, just as before, how the cross-situational distributional evidence X of N independent scene-utterance pairs:

$$X = [(\mathbf{s}_1, \mathbf{u}_1), \ldots, (\mathbf{s}_N, \mathbf{u}_N)] \tag{8}$$

yields different word-concept mappings $p(H_i|X)$ through independent combination of the two sources of evidence:

$$p(H_i|X) = \frac{\prod_{j=1}^{N} p(\mathbf{s}_j|H_i)p(\mathbf{u}_j|H_i)p(H_i)}{p(X)} \tag{9}$$

For expository purposes, we will consider how the learner would rank each of the 6 precise hypotheses, and will assume they entertain only these:

English Verb	Hypothesis	Feature
pour	H_{pour}	11*
spray	H_{spray}	12*
splash	H_{splash}	13*
fill	H_{fill}	0*1
empty	H_{empty}	0*2
move	H_{move}	1**

The likelihood $p(\mathbf{s}_j|H_i)$ for each of the D independent dimensions ($D = 3$) is:

$$p(\mathbf{s}_j = s_1 \ldots s_D|H_i) = \prod_{k=1}^{D} p(s_k|H_i) \tag{10}$$

where our model for scene observations along the kth dimension is:

$$p(s_k|H_i) = \begin{cases} 1 - d_k\epsilon & \text{if} s_k = 0, H_i^k = * \\ \epsilon & \text{if} s_k \neq 0, H_i^k = * \\ 1 - d_k\delta & \text{if} s_k = H_i^k, H_i^k \neq * \\ \delta & \text{if} s_k \neq H_i^k, H_i^k \neq * \end{cases} \tag{11}$$

We annotate the value of the kth dimension of hypothesis H_i as H_i^k above. The first two lines model that when a feature is not valued ($H_i^k = *$), then scenes typically have 0 for the kth dimension ($d_1 = 2; d_2 = 3; d_3 = 3$), but do not match with probability ϵ. That is, observing pouring, spraying, splashing manners ($s_2 = 1, 2,$ or 3), and observing filling, emptying, or breaking change-of-states ($s_3 = 1, 2,$ or 3)

Situation	Scene **s**		Utterance **u**		H_{pour}	H_{spray}	H_{splash}	H_{fill}	H_{empty}	H_{move}
1	pour-fill	$\{G_{001}, W_{110}\}$	/S glipped water into a glass/	(1**)	**.889**	.008	.008	.000	.000	.093
2	pour-fill	$\{G_{001}, W_{110}\}$	/S glipped glass with water/	(0**)	.000	.000	.000	**.990**	.009	.000
3	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)	**.468**	.004	.004	**.468**	.004	.049
4	none		/S glipped water into a glass/	(1**)	.246	.246	.246	.004	.004	.254
5	none		/S glipped glass with water/	(0**)	.007	.007	.007	**.485**	**.485**	.007
6	none		/Glipping!/	(***)	.166	.166	.166	.166	.166	.170
7	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	pour-empty	$\{G_{002}, W_{110}\}$	/S glipped water from a glass/	(1**)	**.998**	.000	.000	.000	.000	.001
	pour-none	$\{G_{000}, W_{110}\}$	/S glipped water/	(***)						
8	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	splash-fill	$\{G_{001}, W_{120}\}$	/S glipped a glass with water/	(0**)	.000	.000	.000	**.999**	.000	.000
	spray-fill	$\{G_{001}, W_{100}\}$	/S glipped a glass/	(***)						
9	pour-fill	$\{G_{001}, W_{110}\}$	/Glipping!/	(***)						
	splash-empty	$\{G_{001}, W_{120}\}$	/S glipped water/	(***)	.064	.064	.064	.000	.000	**.808**
	spray-none	$\{G_{001}, W_{100}\}$	/S glipped water/	(***)						

Figure 1: Word-concept mapping $p(H_i|X)$, given scene-utterance evidence X of a novel verb, /glip/

is far less likely than observing no manner of motion ($s_2 = 0$) or change of state ($s_3 = 0$) at all. Since observing a different value $s_j \neq 0$ is unlikely to have occurred by accident, it may be an important feature to the concept. The second two lines of (11) model that if a feature is valued ($H_i^k \neq *$), then scenes typically match that feature in value, but do not match with probability δ. That is, for example, given hypothesis H_{pour}, then most of the scenes will contain pouring in them. In our examples, $\epsilon = .1, \delta = .01$; qualitatively, results are not sensitive to changes in these values.

The output of our model is shown in Figure 1.

Suppose, as in Situation 1, a learner is given a single scene-utterance pair (pour-fill, /S glipped water into the glass/): $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = \mathtt{1}**, W)]$, and we wish to compute $p(H_i|X)$ for all $H_i \in \mathcal{H}$. We assume the learner can attend to the argument so as to extract relevant features from the scene. Given the scene pour-fill paired with utterance /S glipped water into a glass/, our Bayesian model places high weight on H_{pour}.

In Situation 2, the scene is the same, but now the syntax /S glipped a glass with water/ provides the learner with the information to attend not to the water's manner-of-motion but to the glass' change of state. Given $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = \mathtt{0}**, G)]$ our model weights H_{fill} heavily.

In Situation 3, the scene is the same, but now the syntax /Glipping!/ gives the learner less information, since the argument in the scene that the speaker may be referring to is unknown: $X = [(\mathbf{s}_1 = \{G_{110}, W_{110}\}, \mathbf{u}_1 = ---)]$ If there are A arguments in the scene, the speaker must have had a particular argument z in mind. The learner must *condition* on all the possibilities of z:

$$p(\mathbf{s}_j|H_i) = \sum_{a=1}^{A} p(\mathbf{s}_j|H_i, z_a)p(z_a) \qquad (12)$$

If learners consider all arguments equally salient ($p(z_i) = \frac{1}{A}$) then this effectively models /Glipping!/ as equivalent to /S is glipping Z1/ with probability $p(z_1) = .5$ and /S is glipping Z2/ with probability $p(z_2) = .5$. For simplicity, we assume $A = 2$ where Z1 is water, Z2 is the glass – but further referential uncertainty can be modeled with higher A. Because of the conditioning on each of A possibilities, this yields a less certain word-concept mapping.

In situation 4 through 6, the same syntactic frames are provided as in situations 1 through 3, but without the scene information. When some syntactic information is provided by the frame (situation 4, /S is glipping water into a glass/), then the manner-of-motion locative verbs are preferred over the change-of-state locative verbs, but no differentiation is possible without the scenes. Likewise, when the frame provides the opposite cue (situation 5, /S is glipping a glass with water/), the opposite preference is achieved, again with no differentiation between possible change-of-state verb concepts. When zero syntactic information is available (situation 6, /Glipping!/), all hypotheses prove equally likely.

Whereas in situation 3 the verb-concept mapping was ambiguous, primarily between H_{pour} and H_{fill}, in situation 7 and 8, learners are provided 2 additional examples to disambiguate. Both the scenes and syntactic frames in situation 7 support H_{pour}, while in situation 8 the scenes and syntactic frames support H_{fill}.

Finally, in situation 9, 2 different scene-utterance pairs primarily support the "superordinate" concept H_{move}, and not any "subordinate" manner-of-motion concept H_{pour}, H_{splash}, or H_{spray}.

Discussion

The reason why our analysis is able to infer so much from so little evidence is because so much is embedded in the given knowledge sources:

- the structure of the hypothesis space $\mathcal{H}$. Our examples contained a small number of feature dimensions and their possible values, but these may

be specified by interfaces to perceptual, motor, memory, or other "theory" representations. If so, whether these are innate or acquired are conditional on their source.

- priors $p(H_i)$ on hypotheses in $\mathcal{H}$. We used equal priors, but updating $p(H_i)$ based on language input is natural. In the verbal domain, such phemonena are commonly observed (e.g. manner vs. path, tight/loose-fit biases).
- likelihood of scenes $\mathbf{s}$ given the word concept $p(\mathbf{s}_j|H_i)$. We stipulated static values of ϵ and δ, but this can be acquired from observation.
- perfect knowledge of the features of the complement. We made this simplifying assumption to illustrate the essential elements of our model, but learners must acquire these features in parallel.
- the likelihood of agreement, $p(C|V)$, between a feature of a novel verb V and its complement C. We speculate that there is sufficient structure in partially learned words so as to acquire the structure in the joint distribution of feature values.

This richness of knowledge is in contrast to the models employed by Regier et al (2001) and Desai (2001), who train connectionist neural networks so as to learn the word-scene associations for adjectives/ nouns and verbs respectively. The high dimensionality of their models forces the need for thousands of training trials, and the interpretation of the weights is notoriously difficult. The assumptions behind these models are not justified by these authors. In contrast, our Bayesian approach makes the hypotheses, priors, and likelihoods explicit, holding this structure to be central.

Siskind (1996) views lexical acquisition as constraint satisfaction, and offers a robust algorithm where the mapping between input and hypothesis space is accomplished by pruning hypotheses that do not occur cross-situationally. Provided an idealized tokenization of the world, the algorithm does not need a large number of examples. However, Siskind's model does not yield any form of preference between different concepts, which is especially important when two or more concepts may be equally constrained by the data. We have shown how a Bayesian analysis explicitly yields preferences between concepts in the posterior probability distribution $p(H_i|X)$.

Tenenbaum and Xu (2000) take the important step of putting word learning in the Bayesian framework that we adopt here, showing how noun learning can occur with a small number of examples in a continuous-variable input space.

Crucially, however, the above models ignore the constraining role of syntax, despite considerable evidence that children use syntax to guide their verb-concept hypothesis space (Gleitman 1990, Naigles 1990, Naigles 1994, Fisher et al 1994, Snedeker and Gleitman 2002). Qualitatively, our models' performance matches the preferences of child learners, modeling their acquisition from as little as one example.

Our use of statistics does not imply any commitment to radical empiricism. Much prior knowledge is stipulated: the structure of the hypothesis space, the priors on hypotheses, and the likelihood of scene-utterance pairs given the hypotheses. It is not specified whether these stipulations are innate or themselves learnable. Linguistics and lexical semantics provide detailed theories of a much larger syntactic and semantic hypothesis space, and little prevents their inclusion in this framework.

Acknowledgements

Many thanks to Robert C. Berwick for motivating and supporting this work. Jesse Snedeker and Josh Tenenbaum provided many stimulating discussions. This work was funded by a provost grant to Prof. Joel Moses.

References

Desai, R. (2001). Bootstrapping in Miniature Language Acquisition. In *Proceedings of the Fourth International Conference on Cognitive Modeling*, pp. 61-66. Hillsdale, NJ: Erlbaum.

Harley, H. and Noyer, R. (2000) Licensing in the non-lexicalist lexicon. In Bert Peeters, (Ed.) *The Lexicon-Encyclopedia Interface*, Amsterdam: Elsevier Press.

Fisher, C., Hall, D., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333-375.

Gleitman, L. (1990) The structural sources of verb meanings. *Language Acquisition*, 1990, 1:3-55.

Levin, B. (1993) *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, IL.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, 117:357-374.

Nomura, N., Jones, D.A. and Berwick, R.C. (1994) An Architecture for a Universal Lexicon: A Case Study on Shared Syntactic Information in Japanese, Hindi, Bengali, Greek, and English. COLING 1994, 243-249.

Pinker, S. (1989) *Learnability and Cognition.* MIT Press, Cambridge, MA.

Regier et al (2001). The Emergence of Words. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*

Siskind, J. (1996) A Computational Study of Cross-Situational Techniques for Learning Word-to-Meaning Mappings. *Cognition*, 61:39- 91

Snedeker, J. and Gleitman, L. (2002) Why it is hard to label our concepts. In G. Hall and S. Waxman (eds.), *Weaving a Lexicon*, Cambridge, MA: MIT Press.

Tenenbaum, J.B. and Xu, F. (2000) Word learning as Bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 517-522)

Objet Trouvé, Holism, and Morphogenesis in Interactive Evolution

Ron W. Noel (noelr1@rcn.com)
WCSU, Department of Psychology, 181 White Street
Danbury, CT 06810

Sylvia Acchione-Noel (sylvia.acchione-noel@corporate.ge.com)
General Electric, JF Welch Leadership Development Center, Old Albany Post Road
Ossining, NY 10562

Abstract

Evolutionary systems are conceptualized as having four transfer functions between the two state spaces of genotype and phenotype. The four transfer functions are epigenesis, survival, mate selection, and genetic recombination. The treatment of these transfer functions is uneven at best. In particular, some complain that epigenesis, the formation of an entity from the original undifferentiated cell mass into tissues, is treated in a too simplistic manner to allow for system flexibility, or creativity. This paper reports on an interactive image evolving system that mimics the morphogenesis processes in epigenesis. System description, results, and theoretical implications are discussed.

Introduction

Interactive evolutionary systems seek to interface evolutionary programming to human preference in order to create systems capable of evolving artifacts that require a human expertise that hasn't yet succumb to computation. A common area for this endeavor is the evolving of art, particularly image. The interfacing of human ability with machine computation requires resolving difficult issues in the arts, humanities, and sciences. Further, progress in the design of interactive evolutionary systems allows a glimpse into how very human abilities such as intuition, projection, and holistic perception interplay with the mechanics of machine computation. This paper reports on one such interactive evolutionary system that seeks to combine human perception with the genetic algorithm to evolve small holistic images.

Humans lack the tremendous numerical computational speed of computers; yet they can process information holistically in an automatic, rapid, and natural manner. Machines possess tremendous computational capabilities; yet no algorithm exists to perform holistic processes as well as humans do. Ideally, a good interactive system would integrate the best human cognitive qualities with machine computational capabilities enabling the resultant hybrid system to outperform either of the two components alone. Evolutionary computation as an algorithm is well suited for the creation of an interactive imaging system. However, problems exist in implementation: How can evolutionary computation best support the holistic processes of human cognition? To answer this question requires an understanding of current theory regarding human holistic processes.

Psychological Issues

A well-known area in cognitive research that studies holistic processes is the recognition of objects and, in particular, the recognition of faces. Different perceptual encoding and representational processes have traditionally differentiated theories regarding the recognition of objects as compared to faces. However, the functional separation of these processes under all conditions of object recognition remains unclear (Bruce & Humphreys, 1994). Much of basic object recognition theory has been based on the decomposition of parts and the analysis of edge features (Marr & Nishihara, 1978; Biederman, 1987; Ullman, 1989). On the other hand, face recognition theory has been based on more holistic processes which utilize surface characteristics such as texture, color, and shading (Price & Humphreys, 1989). Some research suggests that the distinctions between object and face recognition begin to fade when one examines the object recognition processes of experts, who may utilize holistic processes similar to those found in face recognition (Diamond & Carey, 1986; Rhodes & McLean, 1990).

The theory regarding holistic processing of faces can be separated into stronger and weaker stances (Bruce & Humphreys, 1994). Under the weaker stance, features may interact with each other through configural processes to form emergent properties or "second-order relational features" (Diamond & Carey, 1986). Under the stronger stance, face recognition is completely holistic; that is, its representation is non-decomposable in that no explicit description of features exists outside the context of the face (Tanaka & Farah, 1993). These stances provide two ways of approaching the development of an interactive system to support the holistic development of images: (1) A system which manipulates context-free features towards configuration, or (2) a system which develops the configuration of the image first, followed by more detailed development of features within the established context.

We sought to design a recognition-driven system of the latter type, which would support the purely holistic development of images, including faces and objects. This system would function in a feature-free space to provide a

non-decomposable representation of images. For instance, a human may perceive or project a cloud as containing the image of a face, yet a cloud contains no explicit representation of the eyes, nose, or mouth. Such features, say a nose, would only be perceived as a nose within the context of the cloud's facial image. This type of perception or projection of a natural texture is called *objet trové*, and is thought by some (Gombrich, 1960) to be paramount to the perception of art. Further, a cloud is not limited to faces; it might contain other animals, or objects. The cloud merely contains randomly distributed textures which humans can organize perceptually. Just as a cloud supports holistic recognition of images, an interactive evolutionary system could encourage recognition based upon context-dependent rather than context-free properties. We intended to provide a mechanism by which a cloud-like representation could enter into "cumulative selection", in a manner not unlike the wishful thinking of Dawkins (1987).

System Issues

The system presented in this paper can be distinguished from other work in the interactive evolution of images (Dawkins, 1987; Baker & Seltzer, 1994; Sims, 1991; Caldwell & Johnston, 1991; Johnston & Caldwell, 1997; Todd & Latham, 1992). Many interactive evolutionary systems (Dawkins, 1987; Baker & Seltzer, 1994; Sims, 1991; Todd & Latham, 1992) use aesthetic preference to determine the fitness of images that are composed of context-free features. Under conditions of aesthetic preference, the user evolves images opportunistically. These systems do not easily support evolution towards an *a priori* goal. Baker and Seltzer (1994) opportunistically evolved butterflies from randomly generated lines, but when they intentionally evolved a general "face-like" image, they could do so only with difficulty. Further, previous systems sometimes required input images to enable the evolution of faces. Sims (1991) as well as Baker and Seltzer (1994) modified facial images after providing input images of human faces, and Johnston and Caldwell (Caldwell & Johnston, 1991; Johnston & Caldwell, 1997) provided input images of features to evolve configured faces.

The Johnston and Caldwell system (Caldwell & Johnston, 1991; Johnston & Caldwell, 1997) is most similar in purpose to our system in that they used human recognition to evolve images of criminal suspects. Their “Faceprints” system allowed more goal-directed behavior within interactive evolution than previously achieved, and they developed a system that encouraged holistic processes by presenting configured faces from the start. However, their system differs from ours in that they took the weaker theoretical stance towards holistic processing by providing input images of context-free features and then placing them in a randomly generated configuration for further evolution.

The “Faceprints” system represents an approach which is common to evolutionary computation; that is, the majority of evolutionary computation is based on parameterized models which predefine features and pre-encode dimensions upon which the features can vary. A key component of evolutionary computation is the mapping between the genotype and phenotype representations. The genotype representation consists of a string of characters, usually binary, that are used as genetic codes in the reproductive process. The phenotype representation consists of a description of an organism that can be evaluated for fitness and selected for reproduction. The linkage between the genotype and phenotype representations is accomplished by a mathematical mapping that uses a parameterized model. For instance, to evolve rectangles, one could create a formula with the two parameters of height and width that would scale suitable binary numbers to a rectangle of a certain height and width. The binary numbers would form the genotype, and the resultant rectangles would form the phenotype.

There are many problems associated with approaches based on a parameterized model (Hofstadter, 1982). The main problem for creating images is that parameterized models constrain the phenotype representation. A model for rectangles can never create a circle. We might try to escape the problem by adding a selector parameter that would dictate the geometric shape to use. For instance, in a rectangle model, if we wanted to represent circles also, we could add a selector parameter that would indicate whether to implement a rectangle model or a circle model. The repair works, except that the addition leads to a discrete selector parameter function and potentially requires an infinite number of models to represent all objects. Instead, we seek to create a system that avoids the predefinition of features and the mapping of genotype to phenotype.

Image Elicitation System

Instead of using a feature-based space, we created a frequency-based space based on pixel representation. The pixel space representation affects the resolution of the images, but forces no predefined features upon the images themselves. The space is based upon atomic or molecular representation, similar to the notions of atomic or molecular decomposition by Fourier analysis or wavelets (Meyer, 1993). As championed by the pointillists, small points of just a few colors can be used to create the psychological impression of any form and any color. Atomic representation works at the sub-feature level and allows the generation of features along with their configuration. The representation is not constrained by features and encodes a dog, a tree, or a car as easily as a face. For instance, one could create a pixel space of 25-by-25 pixels (625 total units) with each pixel being any of eight colors (three bits of information.) Such a small pixel space has the informational potential to create an enormous number of images, as many as 2^{1875}. The number of possible images is so large that there exists no real constraints on the variety

of forms that may be represented; rather, the model constrains the resolution of the image. In terms of resolution, the space cannot represent objects that require more than 12.5 lines of resolution (each line requires at least one 'on' and one 'off' pixel) in the vertical or horizontal axis. However, increasing the number of pixels and decreasing the pixels' size can reduce the impact of the constraint.

System implementation required addressing additional issues in the method of reproduction and mutation function. First, usually, simple cross-over points are used as the method of reproduction, but such a linear system is inappropriate for a two-dimensional space. Instead, we increased the number of cross-over points until the reproductive system considered a cross-over point at every allele. Such a system of uniform crossovers was implemented by randomly selecting between the genes of the two parents with equal probability. Although some researchers consider uniform crossovers to be deleterious to evolutionary computation (Fogel, 1995), others have found them to be useful (Syswerda, 1989). Secondly, if one uses a mutation function that chooses among all possible genes with equal probability for an allele, the mutation function will eventually return the image to a random state. Instead, we limited the mutation function to the gene values of neighboring pixels, causing smaller changes and greater adaptability.

Each generation has a population of fifty images of which the human selects ten images for use as the parents for the next generation. The resulting image elicitation system consists of a comma plus system since the parents are available for selection in the next generation so that each generation after the first is made of parents plus their offspring (Heitköttere & Beasley, 2002). The genotype representation is an array of alleles that has the same size as the pixel representation (25x25 pixels). Each allele is a character that corresponds to one of the possible colors (or genes) for the pixel. Reproduction creates the offspring genotype by randomly and uniformly selecting between the genes of two randomly selected parents at each allele site.

Results

For purposes of this paper, and given our emphasis on holistic processes, we chose a face as the image to be elicited. The first author began with the specific goal of "elicit Abraham Lincoln" and elicited an image of Abraham Lincoln using four levels of gray (figure 1). The image represents the results of image elicitation after 245 generations. The image was originally generated on a SVGA monitor using a black background. A human's ability to recognize Abraham Lincoln is very dependent on the spatial frequency of the image. In other words, viewing figure 1 at too close or too far of a distance reduces the perceptual quality of the image. Figure 2 displays the evolution of the stochastic prototype of Abraham Lincoln. The matrix represents every fifth generation of Abraham Lincoln's image up to generation 245. The matrix should be scanned from left to right and from top to bottom. Each image is a stochastic prototype created by randomly selecting and copying a gene from one of the ten parents into the corresponding allele of the prototype until each allele of the prototype is created. The sampling function is a uniform, random distribution over the parents. As a result, the composite prototype is similar to all of its parents and evokes the average recognition of its parents.

Theoretical Impact

The process in image elicitation is best described in terms of co-evolution or holistic evolution. This description runs contrary to mainstream thought on how evolutionary computation works. There are currently two ideas on how convergence happens in evolutionary computation: the Building Block Hypothesis by Goldberg (1989) and the Schema Theorem by Holland (1992). We argue that both of the hypotheses are feature analytic and are insufficient to explain what is happening in image elicitation.

The Building Block Hypothesis suggests that the convergence process in evolutionary computation is based upon building blocks or small groups of characters whose introduction into any genotype representation will likely increase the fitness of the phenotype representation. Goldberg suggests that the genetic computation first finds as many of these building blocks as possible, and later in evolution, the building blocks are combined together to give the highest fitness. For instance, a series of 1's in the genotype might give rise to an eyebrow in the image. The presence of an eyebrow in any picture of a face should increase the image quality, and, therefore, the fitness.

The Schema Theorem suggests that the ongoing process in evolutionary computation is implicit parallelism caused by schema processing. Schemata are defined as patterns of characters in the genotype representation that may include "don't care" states. A schema can be specified by a genotype representation in which each gene contains a 1 for "on", 0 for "off", or X for "don't care". In a sense, a schema is a relaxed building block in that it relaxes how tightly clustered the group of "care" genes are. Each genotype representation can contain a large number of schemata. This leads to the implicit parallelism and speeds up search.

The basis for our criticism of the current theories lies in their assumption that one can analyze the genotype while disregarding the phenotype. It also requires one to accept that all intermediate representations (patterns in the genotype that are tried and not kept) are coincidental to the process. In such a view one need only look back from the evolved solution and trace the heritage of its genes. In both theories, the implied process is analytical.

Image elicitation challenges these theories in terms of process and representation. Image elicitation relies on multiple-gene (holistic) representation as opposed to

variable-encoded (feature-based) representation. In image elicitation, the image exists in the phenotype and in the perception of the observer, whereas, in other evolutionary computation, the image description exists in the genotype. Our system allows polygeny and pleiotropy, whereas current theory is based on a direct mapping from the genotype to the phenotype and no separate mapping backwards from the phenotype to the genotype. We can illustrate our theoretical differences through what we call "the gray argument" for holistic processes.

Consider the evolution of a medium gray. Mapping from the gray phenotype back to the genotype reveals two optimal representations as shown by *a* and *b* in figure 3. Using binary representation, where 1 equals an "on" pixel, 0 equals an "off" pixel, and X equals "don't care", one finds that 1010101... is one representation *(a)* and it's complement 0101010... is the other *(b)*. Strangely, one can breed together these complimentary representations and the offspring *c* and *d* will still appear grayish in the phenotype, regardless of genotype of the parents, or type of crossover function used. Grays *c* and *d* result from the use of simple and uniform crossover respectively. How can one describe the process of evolving grays in terms of building blocks or schemata when the phenotype being selected does not depend upon any particular gene being a 1 or 0 or X? The gray phenotype can only be described in the genotype as a fairly uniform distribution of 1's and 0's. Given that the difficulties in phenotype-to-genotype mapping of grays extend to all images, the Building Block Hypothesis and the Schemata Theorem are insufficient to describe an image's evolution. In fact, the gray argument is problematic for many current approaches used to understand, or represent phenotype-to-genotype mapping.

Conclusions

Image elicitation promotes wholeness or a lack of distinct features. If one observes the evolving of Lincoln's image one will notice that the features all co-evolve together. At no time does the process treat the nose differently from the eye. The process is so tightly interwoven that to distinguish the nose from the rest of the face constitutes a false distinction; the nose gets it description from itself and the context provided by the rest of the image.

Image elicitation affords high-order interactions. The placement, sizing, shading, and coloring of an image that bears strong resemblance to the original (e.g., a face) can only be described as highly interactive. The placement, size, etc. of the nose is dependent upon all the other features of the face, for if the nose is anywhere but the right place in relation to other features the image would have no resemblance. High-order interactions are a problem for analytical processes, but not this method. It evolves a complex stimulus within a large information space while maintaining a small population size and a reasonable number of generations.

Image elicitation appears to promote holistic rather than analytical processes. It begins with the grossest of detail and ends with the finest of detail. The elicitation process uses intermediate approximations as placeholders for features and as a means of resolution building. This process of organizing an undifferentiated representation into finer and finer regions, or features appears to mimic the process of morphogenesis found in biological reproduction. We argue that features, which by their nature are fine detail and not available until the end of convergence, cannot be used to explain the process. Such findings are problematic for the building block and schemata theories that are currently used to explain the processes in evolutionary computation.

Summary

The present image elicitation system provides a new technique for integrating the best qualities of human and machine capabilities to create images. Neither system could produce theses images alone. Machines lack the perceptual and memory skills, and humans lack the computational energy to evolve an image. The results show that current theories of evolutionary computation are insufficient to explain the convergence of the images in the absence of a feature-based parameterized space.

The technique of image elicitation allows humans to use their perceptual and cognitive systems to organize visual noise into the objects of their memories. This process of literally pulling an image out of chaos will affect our understanding of intelligent systems and future investigations across many disciplines. Image elicitation will be useful in studying machine intelligence, as well as in studying top-down processes in interactive intelligent systems. The system provides a means for humans to experience how evolutionary computation works by directly immersing themselves in the process. And, it provides cognitive researchers with a means of studying holism in human recognition.

Figures

Figure 1: A stochastic prototype of Abraham Lincoln at generation 245.

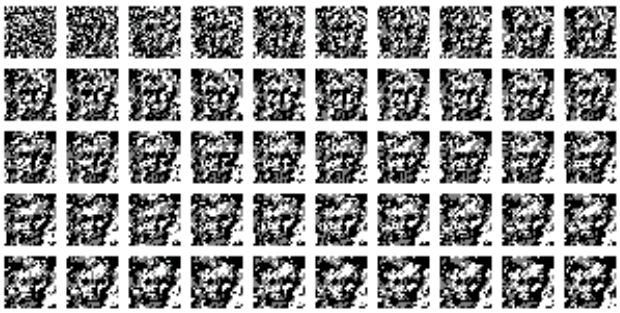


Figure 2: The evolution of an image of Abraham Lincoln, showing every fifth generation up to generation 24

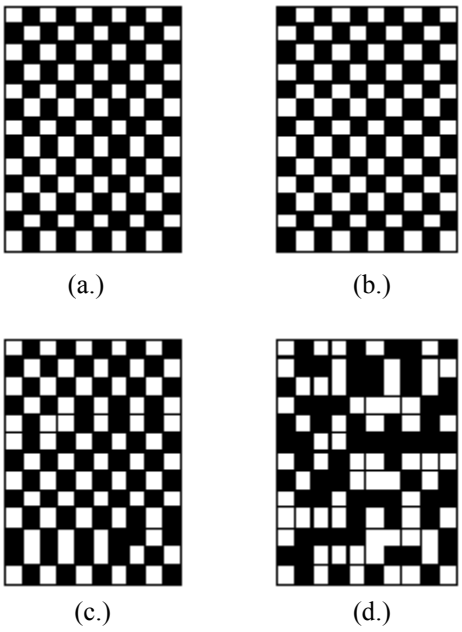


Figure 3: The gray argument for multiple-gene (holistic) representation. Grays *a* and *b* have complimentary phenotype. Grays *c* and *d* result from the use of simple or uniform crossover, respectively.

References

Baker, E. & Seltzer, M. (1994). Evolving line drawings. *Graphics Interface '94 Proceedings*.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*.

Bruce, V. & Humphreys, G. (1994) Recognizing objects and faces. *Visual Cognition*.

Caldwell, C. & Johnston, V. (1991). Tracking a Criminal Suspect through "Face-Space" with a Genetic Algorithm. *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufman.

Dawkins, R. (1987). *The blind watchmaker*. Longman Scientific and Technical, Essex, UK.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*.

Fogel, D. (1995). *Evolutionary computation: Toward a new philosophy of machine intelligence*. IEEE Press, New York.

Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley.

Gombrich, E. (1960). Art and illusion: A study in the psychology of pictorial representation. Princeton University Press, Princeton.

Heitköttere, J. & Beasley, D. (2002). *The hitch-hiker's guide to evolutionary computation*. www.cs.bham.ac.uk/Mirrors/ftp.de.uu.net/EC/clife/www/ - 7k - 04 Feb 2002 .

Hofstadter, D. (1982). Metafont, metamathematics, and metaphysics: Comments on Donald Knuth's article 'The concept of a meta-font'. *Visible Language*.

Holland, L. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge, MA: MIT Press/Bradford Books.

Marr, D. & Nishihara, N. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London*.

Meyer, Y. (1993). *Wavelets: Algorithms and applications*. (Translated and revised by R. Ryan). Society for Industrial and Applied Mathematics, Philadelphia.

Price, P. & Humphreys, G. (1989). The effects of surface detail on object categorization and naming. *Quarterly Journal of Experimental Psychology*.

Rhodes, G. & McLean, I. (1990). Distinctiveness and expertise effects with homogenous stimuli: Towards a model of configural coding. *Perception*.

Sims, K. (1991). Artificial evolution for computer graphics. *Computer Graphics*.

Syswerda, G. (1989). Uniform crossover in genetic algorithms. *Proceedings of the Third International Conference on Genetic Algorithms*. J. D. Shaffer (Eds.), Morgan Kaufmann Publishers, Los Altos, CA.

Tanaka, J. & Farah, M. (1993). Parts and wholes in face recognition. Quarterly Journal of Experimental Psychology.

Todd. S. & Latham, W. (1992). *Evolutionary art as computers*, Academic Press: Harcourt Brace Jovanovich, London.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*.

The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis?

Brent A. Olde (baolde@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Donald R. Franceschetti (dfrncsch@memphis.edu)
Department of Physics, University of Memphis, CAMPUS BOX 523390
Memphis, TN 38152 USA

Ashish Karnavat (akarnavat@chiinc.com)
CHI Systems, Inc., 716 N. Bethlehem Pike, Suite 300
Lower Gwynedd, PA 19002 USA

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

and the Tutoring Research Group

Abstract

Student responses to conceptual physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Expert evaluations of student answers to questions were correlated with LSA metrics of the similarity between student responses and ideal answers. We compared the adequacy of several text corpora in LSA performance evaluation, including the inclusion of written incorrect reasoning and tangentially relevant historical information. The results revealed that there is no benefit in meticulously eliminating the wrong or irrelevant information that normally accompanies a textbook. Results are also reported on the impact of corpus size and the addition of information that is not topic relevant.

Introduction

AutoTutor is an intelligent tutoring agent that interacts with a student using natural language dialogue (Graesser, Person, Harter, & TRG, in press; Graesser, VanLehn, Rose, Jordan, & Harter, 2001). The tutor's interactions are not limited to single-word answers or formulaic yes/no decision trees. AutoTutor attempts to tackle the problem of understanding lengthy discourse contributions of the student, which are often ungrammatical and vague. AutoTutor responds to the student with discourse moves that are pedagogically appropriate. It is this cooperative, constructive, one-on-one dialogue that is believed to produce learning gains (Graesser, Person, & Magliano, 1995). One major component in the comprehension mechanism is the knowledge representation provided by Latent Semantic Analysis (LSA). LSA is a statistical, corpus-based natural language understanding technique that computes similarity comparisons between a set of terms and texts (Kintsch, 1998; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998).

The present study focuses on the domain of conceptual physics. It should be noted that most modern physics texts (such as Hewitt, 1998) devote considerable space to the historical evolution of physical concepts, the cultural context of physics, and its social impact. Some authors also devote appreciable space to discussing discarded theories and chains of reasoning that lead to incorrect conclusions. Thus, a significant fraction of the text found in a physics text may exemplify incorrect thinking.

The Tutoring Research Group at the University of Memphis has been concerned with the best strategy for selecting a corpus of texts when constructing an LSA space. A naive approach would be to gather a number of physics texts, and combine them into one corpus. However, there are some important, unexplored issues that must be addressed about this approach. What should be done about the text that was written to illustrate incorrect reasoning? Does the inclusion of historical information or peripherally related information strengthen or dilute the accuracy with which physics concepts are represented in the LSA space? In short, how much special preparation of the corpus is needed, if it is to provide a reliable representation of the physics that students are expected to learn?

In this paper, we provide a brief overview of LSA and how it is used in our tutoring system. Then we discuss a study designed to address the matter of corpus selection by systematically testing the kind of texts

needed for a training corpus. We discuss the implications of these results for tutoring systems in general.

Latent Semantic Analysis

LSA has recently been successfully used as a statistical representation of a large body of world knowledge (Kintsch, 1998; Landauer & Dumais, 1997). LSA provides the foundation for grading essays, even essays that are not well formed grammatically or semantically. LSA-based essay graders assign grades to essays as reliably as experts in composition (Foltz, Gilliam, & Kendall, 2000). LSA has been used to evaluate the quality of student contributions in interactive dialogs between college students and AutoTutor. AutoTutor is a tutoring system in the domain of computer literacy and most recently physics (Graesser et al., in press; Graesser et al., 2001). The LSA module evaluates the quality of student answers to questions almost as reliably as graduate research assistants (Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, Harter, Person, & TRG, 2000; P. Wiemer-Hastings, K. Wiemer-Hastings, Graesser, & TRG, 1999). Having established the utility of LSA in evaluating the quality of student essays and contributions in a tutoring systems on a variety of topics, we are presently interested in exploring what qualities a useful LSA space must have.

LSA is a mathematical technique in which the information contained in the co-occurrences of words in a body of text is compressed into a set of vectors in N-dimensional space. The input to LSA is a word co-occurrence matrix M, where the individual elements Mij is the number of times that the ith word occurs in the jth document. A document is an arbitrarily defined unit, but normally is a sentence, paragraph, or section in a text; for this project we used paragraphs as our document size. The rows and columns of the matrix are then subjected to mathematical transformations that take into account the frequency of the words used in each of the documents (Berry, Dumais, & O'Brien, 1995; Landauer et al., 1998). Using the mathematical technique of singular value decomposition, the matrix is then expressed as the product of three matrices, the second of which contains the singular values on the diagonal. Changing all but the largest N singular values to zero sets the dimensionality N of the vector space representing the text. The matrices are then re-multiplied to produce a matrix of the same dimensions of the original matrix.

By removing the lowest of the singular values we are seem to be eliminating spurious co-occurrences and capturing a more accurate representation of the meaning of the text (Landauer & Dumais, 1997). The reduced number of dimensions is sufficient for evaluating the conceptual relatedness between any two bags of words. A bag is an unordered set of one or more words. The match (i.e., similarity in meaning, conceptual relatedness) between two bags of words is computed as the geometric cosine (or dot product) between the two associated vectors, with values that normally range from 0 to 1. LSA cosine values successfully predict the coherence of successive sentences in text (Foltz, Kintsch, & Landauer, 1998), the similarity between student answers and ideal answers to questions (Graesser, P. Wiemer-Hastings, et al., 2000; Wiemer-Hastings et al., 1999), and the structural distance between nodes in conceptual graph structures (Graesser, Karnavat, Pomeroy, Wiemer-Hastings, & TRG, 2000).

At this point, researchers are continuing to explore the strengths and limitations of LSA in representing world knowledge. For example, it is widely accepted that LSA is not equipped to handle syntax, word ordering constraints, Boolean expressions, negation, or other precise analytic expressions.

Overview of AutoTutor

In order to fully understand how we use LSA in AutoTutor, it is beneficial to understand the framework in which it is used. Therefore, we briefly provide a general overview of the AutoTutor architecture. A more thorough description is provided in previous publications (Graesser, Person et al., in press; Graesser et al., 1999; Wiemer-Hastings et al., 1998). AutoTutor's style of tutoring was modeled after actual human tutoring sessions (Graesser et al., 1995). The tutor starts out by asking a question or posing a problem that requires a paragraph-length answer. The tutor then works with the student to cover the essential points that the tutor deems necessary to adequately understand the answer to the question. When a question is answered, the process is repeated for a subsequent question. Since most human tutors are peers of the students, they are not what one would label as experts. Thus, they typically have a limited understanding of what the students are trying to convey, yet, they can typically determine whether a response is "in the ball park". Despite the lack of complete understanding, survey studies have shown a sizable advantage for face-to-face tutoring sessions over classroom situations (Cohen, Kulik, & Kulik, 1982).

The user interface for AutoTutor attempts to recreate this face-to-face environment. It consists of four windows: one for presenting the main question, a second for displaying animated or static graphics (simulating diagrams or drawings that a tutor might use to illustrate points), a third with an animated conversational agent, and a fourth for the student to type a reply. AutoTutor's animated agent has synthesized speech, a head, hands, and can be seen from the chest up. These features were designed to provide appropriate speech, facial reactions, and hand

gestures so the student gets both verbal and visual feedback in order to enhance and more appropriately mimic a one-on-one tutoring environment.

AutoTutor's knowledge of its tutoring domain resides in a curriculum script. This is a list of the questions or problems that the tutor is prepared to handle in a tutoring situation, along with good answers to the questions and problems (Putnam, 1987). A major portion of the script is the LSA space; it gets created from an assortment of texts collected from the domain of interest. This corpus is a set of general, non-specific information on the subject matter (e.g., a textbook on conceptual physics), plus specific information directly relevant to the curriculum script. This specific information is comprised of a relatively lengthy, complete, "ideal" answer. This complete answer is separated into a set of specific good answers which address one aspect of the ideal answer; these are sometimes called expectations or points. There are also a set of bad answers and how they would be corrected. Finally, for each expectation in the ideal answer, there are hints, prompts, and assertions that help the student construct an appropriate answer. There are a variety of other dialog moves and slots in the curriculum script that need not be addressed in the present study.

It is important to mention that the LSA corpora investigated in the present study included the general information from textbooks, but never included the question specific information. Thus, only the general physics information was trained in the LSA space. It could be argued that an LSA space should not have any trouble accounting for the content in the curriculum scripts (even if it was a small script) if the material included in the corpus was tailored specifically to the problems. Therefore, we are exploring how far we can go by exclusively focusing on the general content of physics, as manifested in a textbook on conceptual physics.

So how does AutoTutor use LSA during the tutorial interaction? Using the LSA derived cosine matches, AutoTutor evaluates the quality of the student's contributions within a conversation turn and across turns with respect to expected good answers and bad answers to the question. Based on values of these cosine matches, appropriate dialog moves are executed, such as feedback (positive, negative, neutral), pumps, prompts for specific words, hints, assertions, summaries, corrections, and follow-up questions. The smoothness of the mixed-initiative dialog in AutoTutor critically depends on the fidelity of the LSA space. This of course motivated us to test the performance of the LSA space on various tasks and measures.

Methods

Participants. Participants were 120 students from The University of Memphis and Rhodes College; 80 of the students where non-physics majors and 40 were physics majors. Each participant answered 10 problems that were randomly selected from a set of 53 physics problems. Four physics experts answered all 53 questions and graded all answers on a standard 5-point grading scale (A, B, C, D, F). The interrater reliability of the experts was r = .72. In the performance tests of LSA, we compared the expert ratings of the student answers to the LSA cosine scores. The LSA cosine scores are a match between the student answer and the ideal answer (i.e., answers created by the experts). The 4 experts had graduate degrees in physics (2 masters and 2 doctoral).

Materials. We have assembled five different physics corpora to test the effect of the content of the subject matter on the quality of the LSA solutions. The documents in the texts were classified into different rhetorical categories, such as exposition, example problems, historical material, incorrect reasoning, and so on. The fundamental research question is whether the inclusion of different texts and the resulting purity of content will have an impact on the tests of LSA performance.

All the corpora include text materials from the mechanics portion of Paul Hewitt's *Conceptual Physics* (1998). This text is widely used in conceptual physics courses at the college level. Our largest corpus, designated as "All", included chapters 2-9 of the Hewitt book plus six volumes of a comprehensive text aimed at students in technical or life science majors, two advanced texts in electromagnetism, and another two physics texts that were available electronically, a general text by Benjamin Crowell and more advances text by Frank Firk. A somewhat smaller corpus (designated as "Hewitt-Crowell (6)") was constructed from the former by deleting four of the texts; these texts were considered peripherally related to our conceptual physics domain because they were advanced texts mainly dealing with electromagnetism rather than mechanical physic. An even smaller corpus (designated "Hewitt-Crowell (2)") was created by further deleting chapters that did not cover mechanics. Next, we deleted any material from the remaining text that was identified by a physics professor as being primarily historical or involving misconceptions. This was our sanitizing procedure and resulted in the "Hewitt-Crowell (2-Sanitized)" corpus. Finally in the "Hewitt (Sanitized)" corpus, we included only those texts from Hewitt that had been sanitized. It should be noted that each of the successively refined or sanitized corpora was a proper subset of the preceding one. Table 1 summarizes the composition of the five corpora in addition to reporting

the number of paragraphs and the number of unique terms.

Table 1. List of five physics corpora via the chapters that comprise them. Columns with triangles signify sanitized corpora while squares signify unsanitized corpora.

Texts	Hewitt Sanitized	Hewitt Crowell (2-Sanitized)	Hewitt Crowell (2)	Hewitt Crowell (6)	All
Linear Motion	◢	◢	■	■	■
Nonlinear Motion	◢	◢	■	■	■
Newton's Laws of Motion	◢	◢	■	■	■
Momentum	◢	◢	■	■	■
Energy	◢	◢	■	■	■
Rotational Motion	◢	◢	■	■	■
Gravity	◢	◢	■	■	■
Satellite Motion	◢	◢	■	■	■
Newtonian Physics		◢	■	■	■
Conservation Laws		◢	■	■	■
Modern revolution in physics				■	■
Vibrations and Waves				■	■
Electricity and Magnetism				■	■
Optics				■	■
Essential Physics					■
Electromagnetic					■
Field Theory					■
Electrostatics and Circuits					■
Number of Paragraphs	416	698	2051	3445	3778
Number of Terms	1564	2183	4169	6139	6536

Measures. The performance measure was computed on the set of answers to the 53 questions. Since there were 53 questions and approximately 20 answers per question, there was a set of approximately 1000 answers. Each answer was rated by the 4 experts on a 5 point scale (1 = F and 5 = A); the final grade for the answer was the mean grade of the 4 experts. We refer to this score as the grade of the answer. Also associated with each answer was an LSA coverage score, this score compared each student answer to the set of expectations in the experts' answers to the question. More specifically, each expert answers was segregated into a set of expectations, with each expectation being one sentence. An expectation was scored as "covered" if the LSA score between any sentence in the student answer and the expectation under consideration had an LSA cosine score that was greater than or equal to some threshold *T*. The extent to which student answer *S* matched expert answer *A* was computed as the proportion of expectations in *A* that had LSA matches that met the threshold (see Graesser et al., 2000). There were 4 of these scores, one for each of the 4 experts. The maximum value of these scores was designated as the LSA coverage score for student answer *S*. Moreover, we varied the thresholds in these computations from .3 to .9 in increments of .1 (see Figure 1). The correlation between the grades of the answers and the LSA coverage scores was the critical performance measure for the LSA space. The higher the correlation, the better the performance of the LSA space.

Results and Discussion

We tested 5 different physics corpora, each having a slightly different level of specificity in the domain of conceptually based mechanical physics. Because the size of the corpus could affect the dimensionality and threshold, we tested the performance of all 5 levels of corpus size on 5 different dimensionalities (100, 200, 300, 400, and 500), and 7 critical threshold values, from 0.3 to 0.9 in 0.1 increments. For each combination of these factors, we computed the correlation between the grades and the LSA coverage scores.

Figure 1 plots performance for each level of corpus size by threshold at 300 dimensions. We used 300 dimensions for two reasons. First, the sanitized Hewitt corpus was so small that nothing higher than a 300 dimensional representation could be obtained. Second, the performance did not improve after 300 dimensions on any of the corpora. As Figure 1 shows, the LSA performance was practically identical for all corpus sizes except the smallest. Thus, it was not necessary to eliminate historical material, explanations of discarded theories, or useful demonstrations of incorrect chains of reasoning. There was no payoff in sanitizing the corpus.

The size of the corpus had a modest impact on the correlations, except for the extremely small corpus. Clearly the amount of text and the performance of LSA is not a linear relation. A relatively small amount of relevant material can produce acceptable performance with LSA.

According to the results in Figure 1, it appears that a threshold of approximately .8 provides a reasonable fit to the data. Thus, a sentence-like expectation is regarded as covered if there is a sentence in the student answer that has an LSA match score of .8 or higher.

In summary, we have developed a number of alternative physics text corpora for use in the evaluation of student answers to physics questions. Comparisons of the expert grades of the student answers and the computed LSA coverage scores suggest that the inclusion of material that is historical in nature or that exemplifies incorrect notions of physics does not hamper the performance the LSA space. It was also

surprising that the space performed as well as it did considering that there was no problem-specific information in the set of texts used for training the LSA space. Furthermore, a relatively small space in the restricted domain of physics contains enough information to mine an appropriate co-occurrence matrix and produce a properly functioning LSA space. Our current plan is to follow up this experiment by investigating how much performance is improved by adding the specific curriculum script information.

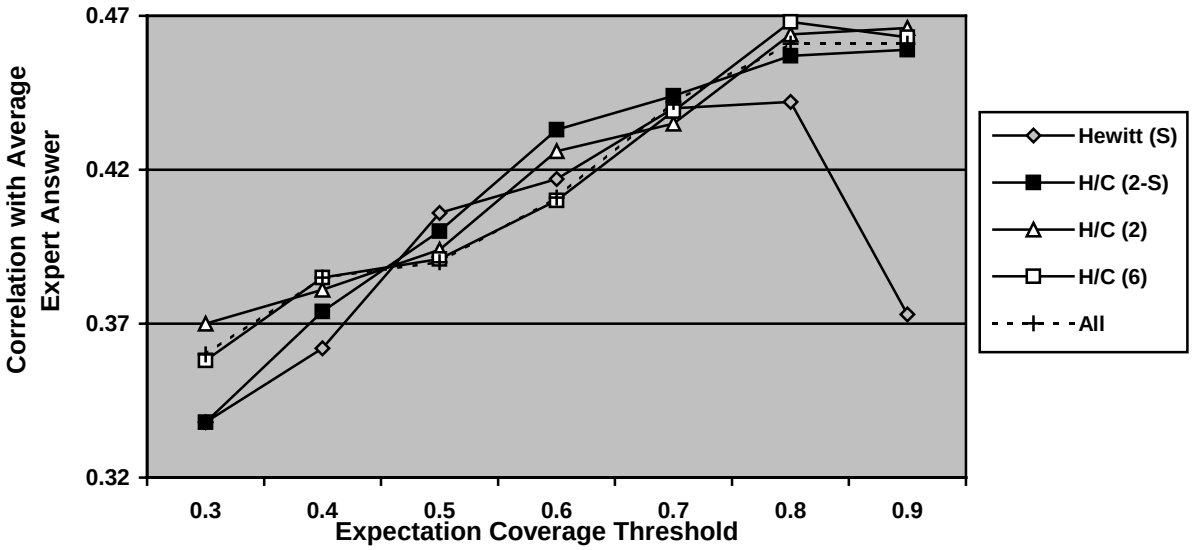


Figure 1: Correlation between the average expert grade and the student's LSA coverage score as a function of threshold and corpus of texts.

Acknowledgments

This research was directly supported by the National Science Foundation (REC 0106965) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

References

Albacete, P. L., & VanLehn, K. A. (2000), Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 25-30). Mahwah, NJ: Lawrence Erlbaum.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995), Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573-595.

Cohen, P. A., Kulik, J. A, and Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

Foltz, W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-128.

Foltz, P. W., Kintsch, W., & Landauer, T. K, (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes 25*, 285-307.

Graesser, A. C., Karnavat, A., Pomeroy, A., Wiemer-Hastings, K., & TRG (2000), Latent semantic analysis captures causal, goal-oriented, and taxonomic structures. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 184-189) Mahwah, NJ: Erlbaum.

Graesser, A.C., Person, N., Harter, D., & TRG (in press). Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 359.1-28.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., & Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*, 128-148.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the TRG (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1*, 35-51.

Hewitt, P. G. (1998) *Conceptual physics* (Ed. 8). Reading, MA: Addison Wesley Longman.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Ploetzner, R., & VanLehn, K. (1997). *Cognition & Instruction, 15*, 169-205.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24*:13-48.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review or research-based instructional strategies, *American Journal of Physics, 59*, 891-897.

Wiemer-Hastings, P., Graesser, A.C., Harter, D., and the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. *Proceedings of the 4th International Conference on Intelligent Tutoring Systems* (pp. 334-343). Berlin, Germany: Springer-Verlag.

Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A. & TRG (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 535-542). Amsterdam: IOS Press.

Experience and Pseudo-Experience: Exemplar Effects Without Feedback

Henrik Olsson (henrik.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Abstract

Many real world situations do not offer unambiguous outcome feedback on how to categorize objects. Models in the categorization literature have mostly been formulated for tasks with trial-by-trial outcome feedback. We examined if there was evidence for exemplar memory also when no external feedback is provided and the criterion is derivative of more abstract knowledge. In a "teacher-student" task, a teacher learns how to judge the toxicity of bugs from external outcome feedback and conveys this knowledge to a student that receives no outcome feedback. The results showed that the students exhibit exemplar effects even if the instructions from the teachers were in the form of rules.

Introduction

Consider listening to your very first speech by a politician. Your previous knowledge is likely to influence your attitude towards her or him. Perhaps, already from childhood your father has imprinted in you that politicians are guided by strictly egoist motives and your general conceptions thus include a belief that no politician advocates a proposal that does not lay in his or her personal interest. You hear a short speech that is neutral in content. Later you listen to another politician. How is your opinion of this second politician influenced by the first encounter? You did not receive much useful feedback from the first encounter, as you only listened to a short neutral speech. However, in your memory the first politician is stored as a person only interested in pursuing his own interest. This exemplar memory only in part derives from direct experience with a politician; in part it is derivative of more general beliefs held prior to the encounter. However, by now this belief is supported also by "concrete experience" of politicians.

This was an example of a real world situation that is different from most categorization experiments where classification models are tested in tasks with simple perceptual stimuli and trial-by-trial outcome feedback. Everyday situations often do not offer direct and unambiguous feedback and exemplar memory is thus likely to derive in part also from other sources of knowledge besides concrete experience with the objects.

There has been increasing interest in multiple representation levels (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998) and there is evidence that people can adaptively shift between different representation levels in response to the experimental demands (Jones, Juslin, Olsson, & Winman, 2000). With experience, knowledge first represented abstractly may be projected onto concrete exemplars so that in the end the beliefs are supported also by an extensive exemplar memory: a phenomenon that might be called *pseudo-experience*.

Even if some extensions of exemplar models allows for storage of exemplars as they are interpreted and not solely in terms of their physical properties (e.g., the model presented by Smith & Zarate, 1992), the argument supporting this claim is based on general observations and not linked to predictions from different models. For example, one such observation is that a reencounter of a stimulus facilitate the same reactions and processes (see the review in Smith & Zarate, 1992).

In this paper, we examine the possibility of extending the scope of *exemplar models* (Medin & Schaffer, 1978; Nosofsky, 1986) to situations where people do not receive outcome feedback, but form beliefs about the criterion from abstract knowledge of rules. In these circumstances, one possibility is that people completely abandon exemplar processes as a basis for their judgments and directly use abstract knowledge in the form of rules or prototypes.

Another possibility is that people generate the criteria from abstract knowledge and store them together with the experienced exemplars; later to rely on these stored exemplars to make their judgments. We explore these possibilities in a "teacher-student" task where a teacher learns how to judge the toxicity of bugs from outcome *feedback* and the student has to rely on *feedforward* summary information provided by the teacher. The question is if there is evidence for exemplar processing in the students judgments even if they do not receive feedback or instructions about exemplars from the teachers.

Measuring Exemplar Effects

To develop an exemplar effect index we need to consider a category structure that allows us to differentiate between predictions by the exemplar model and other plausible models, in this case a cue-abstraction model that linearly integrates cues. The results previously obtained with this task revealed large individual differences and a shift from exemplar memory to more mental cue-integration processes when the criterion is changed from classification to a continuous judgment task (Juslin, Olsson, & Olsson, 2002).

The task requires participants to use four binary cues to infer a continuous criterion. (Juslin et al., 2002). The judgments involve the toxicity of subspecies of a fictitious Death Bug. The different subspecies vary in concentration of poison from 50 ppm (harmless) to 60 ppm (lethal). The toxicity can be inferred from four visual cues of the subspecies (e.g., the length of their legs, color of their back).

The binary cues C_1, C_2, C_3, and C_4 take on values 1 or 0. The toxicity c of a subspecies is a linear, additive function of the cue values:

$$c = 50 + 4 \cdot C_1 + 3 \cdot C_2 + 2 \cdot C_3 + 1 \cdot C_4 . \qquad (1)$$

C_1 is the most important cue with *coefficient* 4 (i.e., a relative weight .4), C_2 is the second to most important with coefficient 3, and so forth. A subspecies with feature vector (0, 0, 0, 0) thus has poison concentration 50 ppm; a subspecies with feature vector (1, 1, 1, 1) has 60 ppm. The continuous criteria for all 16 subspecies (i.e., possible cue configurations) are summarized in Table 1.

Table 1
Structure of the Task

Exemplar	C_1	C_2	C_3	C_4	Criterion	Exemplartype
1	1	1	1	1	60	E
2	1	1	1	0	59	T
3	1	1	0	1	58	T
4	1	1	0	0	57	O
5	1	0	1	1	57	N
6	1	0	1	0	56	N
7	1	0	0	1	55	N
8	1	0	0	0	54	T
9	0	1	1	1	56	O
10	0	1	1	0	55	O
11	0	1	0	1	54	T
12	0	1	0	0	53	T
13	0	0	1	1	53	T
14	0	0	1	0	52	T
15	0	0	0	1	51	T
16	0	0	0	0	50	E

Note: C = Cue; E = Extrapolation; T = training exemplar; O = Old comparison exemplar in training, N = New comparison exemplar presented at test.

In *training*, the participants encounter 11 subspecies and make *continuous judgments* about the toxicity of each subspecies ("The amount of poison is 57%"). As indicated in the two right-most columns of Table 1, five subspecies are omitted in training. In a *test phase*, the participants make the same judgments as in the training phase, but for all the 16 subspecies and without feedback. The task allows perfect performance in training both by exemplar memory and induction of the task structure (i.e., by inducing the cue weights in Eq. 1).

The *exemplar model* implies that participants make judgments by retrieving similar exemplars (subspecies) from long-term memory. The *context model* of classification (Medin & Schaffer, 1978) suggests that in a task that only requires participants to judge if a bug is dangerous or not, the probability $p_E(\hat{b}=1)$ of categorization as dangerous (1) equals the ratio between the summed similarity of the judgment probe to the dangerous exemplars and the summed similarity to all exemplars:

$$p_E(\hat{b}=1) = \frac{\sum_{j=1}^{J} S(p, x_j) \cdot b_j}{\sum_{j=1}^{J} S(p, x_j)}, \qquad (4)$$

where p is the probe to be judged, x_j is stored exemplar j (j=1…J), $S(p,x_j)$ is the similarity between the probe p and exemplar x_j, and b_j is the binary criterion stored with exemplar j (b_j=1 for dangerous, b_j=1 for harmless). J depends on the size of training set of exemplars.

The similarity between probe p and exemplar x_j is computed by the multiplicative similarity rule of the context model (Medin & Schaffer, 1978):

$$S(p, x_j) = \prod_{i=1}^{4} d_i , \qquad (5)$$

where d_i is an index that takes value 1 if the cue values on cue dimension i coincide (i.e., both are 0 or both are 1), and s_i if they deviate (i.e., one is 0, the other is 1). s_i are four parameters in the interval [0, 1] that capture the impact of deviating cues (features) on the overall perceived similarity $S(p,x_j)$. s_i close to1 implies that a deviating feature on this cue dimension has no impact at all on the perceived similarity and is considered irrelevant. s_i close to 0 means that the overall similarity $S(p,x_j)$ is close to 0 if this feature is deviating, assigning crucial importance to the feature. The parameters s_i capture the similarity relations between stimuli and the attention paid to each cue dimension, where a lower s_i signifies higher attention.

The context model was developed for classification, in most cases to binary categories. To generate predictions also for judgments of a continuous criterion we relax the model by allowing the outcome index b_j to take not only binary but also continuous values. The estimate $\hat{c}_E$ of the criterion c is a weighted average of the criteria c_j stored for the exemplars, where the similarities $S(p,x_j)$ are weights (see e.g., Juslin & Persson, 2000; Smith & Zarate, 1992, for similar applications).

The *cue-abstraction model* assumes that the participants abstract explicit cue-criterion relations in training, which are mentally integrated at the time of judgment. When presented with a probe the participants retrieve rules connecting cues to the criterion from memory (e.g., "Green back goes with being poisonous"). The rules specify the sign of the relation and the importance of each cue with a cue weight. For example, after training the rule for cue C_1 may specify that C_1=1 goes with a large increase in the toxicity of a subspecies.

Cue abstraction implies that participants compute an estimate $\hat{c}_R$ of the continuous criterion c. For each cue, the appropriate rule is retrieved and the estimate of c is

adjusted according to the cue weight ω_i (i=1…4). The final estimate $\hat{c}_R$ of c is a linear additive function of the cue values C_i,

$$\hat{c}_R = k + \sum_{i=1}^{4} \omega_i \cdot C_i \,, \qquad (2)$$

where $k = 50 + .5 \cdot (10 - \sum \omega_i)$. If ω_1=4, ω_2=3, ω_3=2, and ω_4=1, Eq's 1 and 2 are identical and the model produce perfect judgments. The intercept k constrains the function relating judgments to criteria to be regressive around the midpoint (55) of the interval [50, 60] specified by the task instructions.

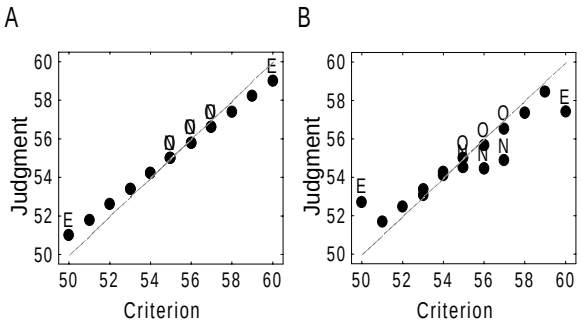


Figure 1*:* Panel A: Cue abstraction model predictions for the constrained training set. Panel B: Exemplar model predictions with similarity parameter *s*=.1 for the constrained training set. O = Old comparison exemplar in training, N = New comparison exemplar presented at test, E = Extrapolation exemplars.

The predictions from the two models are summarized in Figure 1 and shows that the models predict distinct patterns. The cue abstraction model allows accurate extrapolation beyond the distribution of criteria in the training set [51, 59]. Whenever the correct signs of the cue weights are identified, the most extreme judgments are made for exemplars 1 and 16. The exemplar model, that computes a weighted average of the criteria observed in training, can never produces a judgment outside the observed range. The most extreme judgments are made for criteria *c*=51 and 59.

Moreover, with cue abstraction there should be no systematic differences between judgments for "New" and "Old" exemplars with *c*=55, 56, and 57: the process is the same in both cases. However, with the exemplar model there is more accurate judgments for Old exemplars, because these judgments benefit from retrieval of identical exemplars with the correct criterion.

These differences in predictions allow us to define measures of the amount of exemplar processing. First, The *old-new difference index* Δ*ON* is defined as,

$$\Delta ON = \bar{d}_{Old} - \bar{d}_{New} \qquad (6)$$

where $\bar{d}_{Old}$ is the mean absolute deviation between judgment and criterion for the three old exemplars and $\bar{d}_{New}$ is the corresponding mean deviation for the three new exemplars in Table 1. When judgments for old rather than new exemplars are more accurate, the index is negative. The *extrapolation index EI* is the mean deviation from linear extrapolation,

$$EI = \frac{(x_{51} - x_{50}) + (x_{60} - x_{59}) - 2b)}{2} \qquad (7)$$

where x_{50}, x_{51}, x_{59}, and x_{60} are the judgments for exemplars with criteria 50, 51, 59 and 60, respectively. The value of *b* is determined by the difference $x_{51} - x_{50}$ (or equivalently $x_{60} - x_{59}$) predicted by a linear regression relating judgments to criteria. Perfect linear extrapolation implies an extrapolation index that is 0 (e.g., when the judgments are perfectly accurate). If the index is negative, the exemplars with extreme criteria do not receive as extreme judgments as implied by linear extrapolation. For example, the indices in Figures 1A are 0, but the indices in Figure 1B are negative. The mean of Δ*ON* and EI provides a total index of exemplar effects, *Total EE*.

Method

Participants

Forty undergraduate students participated. The participants were paid between 50 and 100 SEK depending on their performance.

Apparatus and Materials

The experiment was carried out on a PC–compatible computer. The exemplars varied in terms of four binary cues; leg length (short or long), nose length (short or long), spots or no spots on the fore back and two patterns on the back. The cues and the cue values in the abstract structure in Table 1 were randomly assigned to new concrete visual features for each new participant. Two types of presentation modes were used, one *analogue* where drawings of the bugs were presented and one *propositional* with written descriptions of the bugs.

Design and Procedure

The task was done in pairs, one participant in each pair was randomly assigned as teacher and the other as student. Half of the teacher-student pairs were randomly assigned to the analogue condition and the other half to the propositional condition.

The written instructions informed the participants that the task involved judgments of the toxicity of subspecies of a Death bug from 50 to 60 ppm and that the difference between teachers and students was that the teacher receives outcome feedback and that the student does not. The participants were also informed that they would receive a minimum payment of 50 SEK and up to 100 SEK *depending on the performance of the student*. The performance bonus was calculated by taking half the correlation between the students' judgments and the criterion values in the test phase times 100. In addition, the teacher was told that after each training block they were to write down instructions to the stu-

dent on how to assess the toxicity of the bugs as a Word file. The teachers were free to give any instructions they wanted. The word files were collected for further analysis by the experimenter. No additional contacts between teachers and students were allowed (except for strictly clarifying questions in regard to spelling errors, as mediated by the experimenter).

The training phase consisted of four blocks with 55 trials each making a total of 220 trials. Only the teacher received outcome feedback of the correct toxicity level in the learning phase. After each training block the teacher wrote down instructions to the student and the experimenter handed it over to the student that were seated in another room. In the test phase each exemplar were presented four times without feedback for both teachers and students making a total of 64 trials. The entire experiment took from one hour and fifteen minutes to two hours.

Results

Performance for teachers and students in the learning phase of the analogue and the propositional conditions measured by the absolute deviation between judgment and criterion are presented in Figure 2. The performance for teachers and students are about the same in the last part of training. Performance is better in the analogue condition than in the propositional condition.

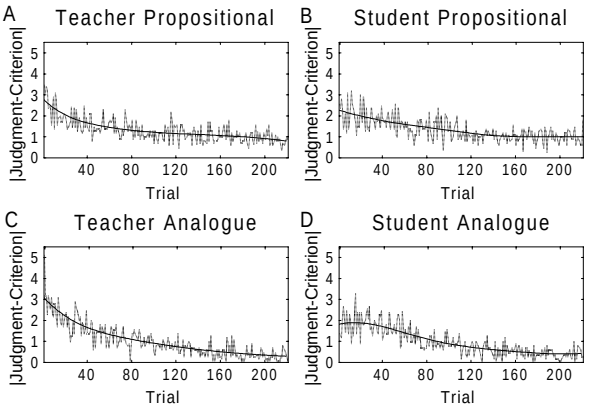


Figure 2: Panels A-D: Mean absolute deviation between judgment and criterion for teachers and students. The curves are fitted according to a negative exponentially weighted smoothing procedure.

Figure 3 shows that there appears to be some exemplar effects in all the conditions, for both teachers and students. The judgments of the new exemplars are less accurate than the old exemplars and underestimate the criteria. The figure also shows that both teachers and students have some difficulties with extrapolating beyond the distribution of criteria in the training set.

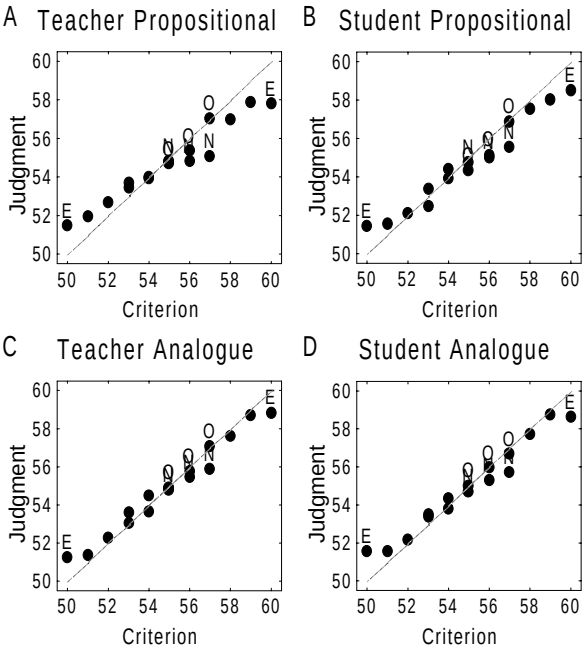


Figure 3: Panels A-D: Mean judgments for the different criteria for teachers and students.

The exemplar indexes were collapsed over the analogue and the propositional conditions as no significant differences were found between the two conditions for any of the indexes. Figure 4 shows that there are clear exemplar effects for both teachers and students, as the 95% confidence intervals does not include zero.

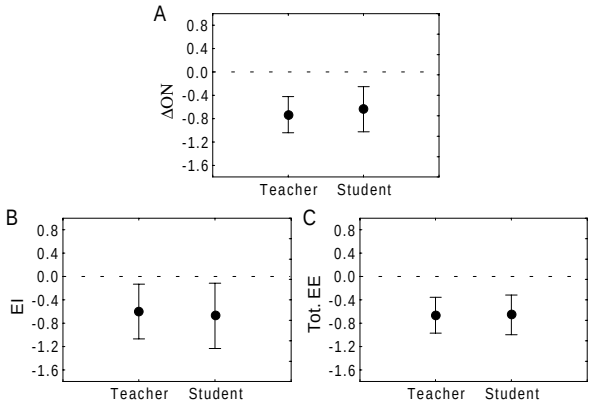


Figure 4: Panel A: Mean Old-New difference index, Δ*ON*. Panel B: Mean *extrapolation index EI*. Panel C: Mean total exemplar effects index, *EE*. The error bars are 95% confidence intervals.

We coded the instructions the teachers gave the students as containing exemplars or not by a strict coding scheme that assigned any ambiguous case as an exemplar instruction. Six teachers had instructions containing exemplar information, for example “Green body, short legs, long nose, no spots = 51%”. A typical part of

an instruction that did not contain exemplar instruction was "Begin with 50% and add: short grey legs +2…green long legs +0 [and so on for all the cues]".

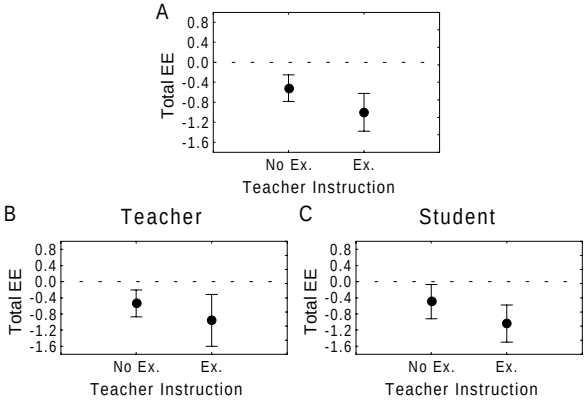


Figure 5: Panel A: Mean total exemplar effects index, *EE*, for No exemplar instructions and Exemplar instructions over all conditions. Panel B: Mean total exemplar effects index, *EE*, for No exemplar instructions and Exemplar instructions for teachers. Panel C: Mean total exemplar effects index, *EE*, for No exemplar instructions and Exemplar instructions for students. The error bars are 95% confidence intervals.

Shown in Figure 5 is the total exemplar effects index separately for the participants with no exemplars in the teachers' instructions and those with exemplars in the instruction. It can be seen that there is an effect of instruction with larger exemplar effects for exemplar instructions, $t(38) = 2.00$, $p = .026$, one-tailed. More importantly, students with no exemplar instructions from their teachers exhibit evidence for exemplar processing, as the confidence intervals do not include zero.

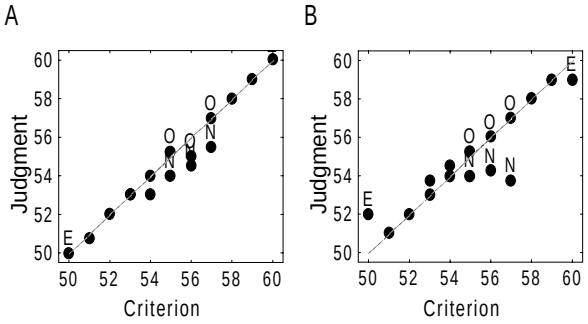


Figure 6: Mean judgments for two students that did not receive any exemplar information. Panel A: A student better described with a cue-abstraction model. Panel B: A student better described with an exemplar model.

Investigation of individual participants data reveals large individual differences. Figure 6 shows two students that did not receive any exemplar information from their teachers. One student is better described as a cue abstraction participant and the other student as an exemplar participant.

Discussion

In this paper, we have shown that people can spontaneously rely on pseudo-experience in a judgment task. Even when the information people receive does not contain information about specific exemplars, people cannot help project abstract rule knowledge onto concrete exemplars and then use these exemplars in the judgment process.

Even if there are large individual differences in data with some people operating only in accordance with the cue-abstraction model, it seems difficult for most people to totally abandon exemplar processing. Even if you initially execute a rule to determine what response you will make, the very act of executing the rule implies processing the exemplar in front of you. For example, you need to scan the object for features that fit the rule conditions. Even if you do not consciously trying to remember exemplars, the end result is incidental learning of exemplars that later influences judgments.

One caveat is that the types of categories used could affect the prevalence of exemplar based processing. For example, the results in a series of experiments by Minda, Smith and colleagues (e.g., Minda & Smith, 2001) suggest that larger categories, better structured categories, and more complex stimulus promotes prototype processing at the expense of exemplar processing.

In this, and other tasks, rule based representations provide great powers of generalization and communication. One answer why we sometimes cannot avoid storing and using exemplars may be found in the idea that our cognitive system consists of multiple levels of representation that work together or compete to determine responses in specific tasks (see e.g., Ashby et al., 1998). Our results fit into the notion that the exemplar representation has a function that separates it from other representational formats in that it acts like an automatically activated back-up system that preserves distributional and individuating information about the world.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442-481.

Jones, S., Juslin, P., Olsson, H., & Winman, A. (2000). Algorithm, heuristic or exemplar*:* Processes and representation in multiple-cue judgment. In L. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 244-249). Hillsdale, NJ: Erlbaum.

Juslin, P., Olsson, H., & Olsson, A-C. (2002). *Abstract and concrete knowledge in categorization and multiple-cue judgment*. Manuscript submitted for publication. Department of Psychology, Umeå University, Umeå, Sweden.

Juslin, P., & Persson, M. (2000). *PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge*. Manuscript submitted for publication. Department of Psychology, Umeå University, Umeå, Sweden.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 775-799.

Nosofsky, R. M (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3-21.

Simplicity: A Cure for Overgeneralizations in Language Acquisition?

Luca Onnis (l.onnis@ warwick.ac.uk)
Department of Psychology, University of Warwick
CV4 7AL Coventry, UK

Matthew Roberts (m.roberts.2@ warwick.ac.uk)
Department of Psychology, University of Warwick
CV4 7AL Coventry, UK

Nick Chater (nick.chater@warwick.ac.uk)
Department of Psychology and Institute for Applied Cognitive Science, University of Warwick
CV4 7AL Coventry, UK

Abstract

A formal model of learning as induction, the simplicity principle (e.g. Chater & Vitányi, 2001) states that the cognitive system seeks the hypothesis that provides the briefest representation of the available data– here the linguistic input to the child. Data gathered from the CHILDES database were used as an approximation of positive input the child receives from adults. We considered linguistic structures that would yield overgeneralization, according to Baker's paradox (Baker, 1979). A simplicity based simulation was run incorporating two different hypotheses about the grammar: (1) The child assumes that there are no exceptions to the grammar. This hypothesis leads to overgeneralization. (2) The child assumes that some constructions are not allowed. For small corpora of data, the first hypothesis produced a simpler representation. However, for larger corpora, the second hypothesis was preferred as it lead to a shorter input description and eliminated overgeneralization.

Introduction

Overgeneralizations are a common feature of language development. In learning the English past tense, children typically overgeneralize the '-ed' rule, producing constructions such as *we holded the baby rabbits* (Pinker, 1995). Language learners recover from these errors, in spite of the lack of negative evidence and the infinity of allowable constructions that remain unheard; it has been argued that this favours the existence of a specific language-learning device (e.g. Chomsky, 1980; Pinker, 1989). This is an aspect of the 'Poverty of the Stimulus' argument. We report on a statistical model of language acquisition, which suggests that recovery from overgeneralizations may proceed from positive evidence alone. Specifically, we show that adult linguistic competence in quasi-regular structures may stem from an interaction between a general cognitive principle, *simplicity* (Chater, 1996) and statistical properties of the input.

According to Baker's Paradox (Baker, 1979) children are exposed to linguistic structures that they subsequently overgeneralize, demonstrating that they capture some general structure of the language. However, some generalizations are grammatically incorrect and children do not receive direct negative evidence from caretakers (e.g. corrections labeling such overgeneralizations as disallowed). The paradox is that non-occurrence is not in itself evidence for the incorrectness of a construction because an infinite number of unheard sentences are still correct. The irregularities that Baker referred to can be broadly labeled *alternations* (Levin, 1993; see also Culicover, 2000). For instance, the dative alternation in English allows a class of verbs to take both the double-object construction (*He gave Mark the book*) and the prepositional construction (*He gave the book to Mark*). Hence the verb *give* alternates between two constructions. However, certain verbs seem to be constrained to one possible construction only (*He donated the book to Mark* is allowed, whereas **He donated Mark the book* is not). Such verbs are non-alternating. From empirical studies we know that children do make overgeneralization errors that involve alternations, such as **I said her no* (by analogy to *I told her no*, Bowerman, 1996; Lord 1979).

In this paper we present alternation phenomena from the CHILDES database (MacWhinney, 2000) of child-directed speech which will be used in the computer model. Secondly, we introduce the *simplicity principle* (Chater, 1996), based on the mathematical theory of Kolmogorov Complexity (Kolmogorov, 1965). Thirdly, we present an artificial language designed to model the CHILDES data, and describe simplicity-based models of language processing and the simulations of recovery from overgeneralizations. Lastly we discuss the limitations of this specific model and some implications for research on language acquisition.

Causative alternations in child-directed speech

Suppose we have a language in which verbs belong to three distinct classes (V1, V2, V3). Each class is related to two syntactic contexts (C1, C2). One class of verbs (V1) appears in both contexts. Two other classes of verbs (V2 and V3) occur in one context only. We can produce a simple table to visualize the alternation:

Table 1: Alternating and non-alternating verbs across contexts

	C1	C2
V1	1	1
V2	0	1
V3	1	0

The causative alternation in English is of this kind. Verbs like *break* behave both transitively (*I broke the vase*) and intransitively (*The vase broke*), whereas verbs like *disappear* behave only intransitively (*The rabbit disappeared* is allowed; but **I disappeared the rabbit* is not) and verbs like *cut* are found only in transitive contexts (**The bread cuts* is not allowed). An analysis of CHILDES revealed that verbs in child-directed speech fit the pattern of the above idealization: a number of verbs are exclusively transitive or intransitive (see Table 2).

Children eventually generalize the structures of the language they are exposed to. A typical generalization occurs when children say *Don't you fall me down* (Bowerman, 1982; Lord, 1979). This is an overgeneralized use of a non-causative verb as causative. In the causative construction, some verbs like *break* can be used both transitively with a semantic element of cause (*I broke the vase*) and intransitively (*the vase broke*). Verbs like *break* alternate between two constructions. However, *fall* can only be used intransitively, and *hear* only transitively. The acquisition of verbs' argument structure seems particularly complicated as the way verbs behave syntactically is largely arbitrary. Semantically similar verbs like *say* and *tell*, or *give* and *donate* allow for different constructions.

Bowerman (1982) and Lord (1979) recorded a total of 100 different cases in which two-argument verbs are used with three arguments (e.g. *You can drink me the milk*). The developmental literature suggests that when children acquire a new verb they use it productively in both constructions, without specific directional bias (Lord, 1979). It is also worth noting that alternations can be theoretically distinguished from other forms of

Table 2: Verbs in child-directed speech occurring in transitive and intransitive contexts pooled from the CHILDES English sub-corpora (MacWhinney, 2000).

	Verb	Transitive occurrences	Intransitive occurrences
Category V1	bounce	75	117
	break	1251	268
	burn	86	60
	close	855	56
	freeze	18	61
	grow	59	330
	move	966	560
	open	1590	232
	pop	104	153
	rip	139	9
	roll	405	164
	shake	147	26
	slide	65	120
	swing	38	96
	tear	167	20
	turn	2690	600
Category V2	arrive	0	41
	come	0	18437
	dance	0	370
	die	0	141
	disappear	0	73
	fall	0	2945
	go	0	65193
	rise	0	14
	run	0	1569
	stay	0	1413
Category V3	bring	3028	0
	cut	1315	0
	drop	640	0
	kill	120	0
	lift	392	0
	push	1609	0
	put	27154	0
	raise	25	0
	take	9724	0
	throw	2090	0

irregularization like the irregular past tense. In the case of *goed-went* for example, recovery from the overgeneralized form **goed* can be accounted for by directly invoking a competition strategy (MacWhinney, 1987): as the number of *went* in the input increases, it will win over the irregularized form *goed*, which has 0 frequency in the input. Alternations are interesting theoretically in that the competition model does not seem applicable for these. The overgeneralized form does not have an irregular alternative: there is simply a "hole" in the language. This argument was raised by Baker in his distinction between benign exceptions (like

the past tense) and truly problematic alternations like the ones we consider here (Baker 1979).

For the purpose of showing how such problematic irregularities can be learnt using a simplicity principle, we take the causative alternation described above as a working example. We extracted verb frequencies from the CHILDES Database. CHILDES contains a total of nearly ten million words of child-directed speech. Because we are interested in showing that the input the child receives is rich enough for recovery of overgeneralization by induction, only the adult speech in the corpus was selected and analysed.

Simplicity and Language

The simplicity principle (Chater, 1996) states that in choosing among potential models of finite data, there is a general tendency to seek simpler models over complex ones and optimize the trade-off between model complexity and accuracy of model's description (i.e. fit) to the training data. Complexity is thus defined as:

$$C = C(\text{model}) + C(\text{data}|\text{model})$$

The favoured model of any finite set of data will be that which minimizes this term.

In order to compare different grammars we need a measure of simplicity and a "common currency" for measuring both the model complexity and the error term complexity. Fortunately this is possible by viewing grammar induction as a means of *encoding* the linguistic input; the grammatical organization chosen (the "knowledge" of the language) is that which allows the simplest encoding of the input. A tradition within mathematics and computer science, Kolmogorov complexity, shows that the simplest encoding of an object can be identified with the shortest program that regenerates the object (Li & Vitanyi, 1997).

Every sentence generated from a lexicon of *n* words may be coded into a binary sequence. The length of a message refers to a binary string description of the message in an arbitrary universal programming language. The binary string can be seen as a series of binary decisions needed to specify the message; smaller lengths correspond to simpler messages. The brevity of an input A_i is associated to its probability $P(A_i)$ of occurrence. Shannon's (1948) noiseless coding theorem specifies that:

$$\text{Length} = \text{Log}_2[1/P(A_i)]$$

More probable events are therefore given shorter codes. Li & Vitanyi (1997) have shown that the length K(x) of the shortest program generating an object x is also related to its probability Q(x) by the following *coding theorem*:

$$K(x) = \log_2[1/Q(x)]$$

Finally, the *invariance theorem* (Li & Vitanyi, 1997) assures that the shortest description of any object is *invariant* (up to a constant) between different universal languages, thus granting a measure of simplicity that is independent of the data and of the programming language used to encode the data. The above formalizations allow us to replace "Complexity" with "Length" and state that "the best theory to infer from a set of data is the one which minimizes the length of the theory and the length of the data when encoded using the theory as a predictor for the data" (Quinlan and Rivest, 1989; Rissanen, 1989). It is important to note that whilst the MDL principle is well established as a machine learning tool for grammar induction, such models typically make use of parsed corpora or other psychologically implausible inputs (e.g. Osborne, 1999). This paper uses MDL as a metric to present simplicity as a specifically psychological principle.

Modelling language learning with simplicity

In any study of grammar induction, and in particular in the simplicity framework, it is crucial to see a grammar as a *hypothesis about the data*. The best hypothesis is the one that compresses the data maximally, so we can also think of a grammar as compression of the data. We can see the achievement of adult linguistic competence as a process of building different hypotheses about the language in order to achieve optimum compression. The essence of compression is to provide a shorter encoding of the data, enabling generalizations and correct predictions. Alternations are particularly informative about the possibility of a cognitive system to capture dependencies from limited data. If linguistic structures were completely regular, then generalizing from a few data would be easy. But as alternations are quasi-regular, meaning there are exceptions to their regularity, a learner must capture fine dependencies in order to generalize whilst avoiding overgeneralizations.

The issue is to choose the candidate model of the right complexity to describe the corpus data, as stated by the simplicity principle. We can compare different hypotheses (grammars) at different stages of learning and choose, for each stage, the one that minimizes the sum of the grammar-encoding-length and the data-encoding-length. In the following section we compare data compression of corpora by two similar models. The difference between them is that one posits a completely regular rule, whilst the other posits a regular rule and some exceptions to it. We can think of the second model as having 'invested' in exceptions. Each exception initially produces less compression overall,

since the exceptions cost some bits to specify. However, each exception shortens the code-length for each item in the corpus, and the second model thereby 'recoups' its investment over time.

The Models

This approach to language acquisition does not focus on how learning occurs. Rather, these simulations run several models concurrently to show that the rate of increase of code-length differs between hypotheses about language. This section describes the structure of two hypotheses (grammars); the first gives rise to overgeneralization phenomena whilst the second does not. These were designed in conjunction with a very simple artificial language, which was subsequently used to test the models. A brief outline of the language is given here to facilitate the description of the model. A more detailed consideration of how the artificial language relates to data from corpora of child-directed speech is given below.

The artificial language used consists of two syntactic categories. These can be thought of crudely as nouns and verbs. They can be combined to form two-word sentences. Sentences may be of the form NV or VN. Forms NN and VV are disallowed. In addition, a number of sentences are disallowed. Let us imagine that there are four nouns (n_1-n_4) and four verbs (v_1-v_4) in the language, and that v_4 is blocked in the sentence final position. From this it follows that four sentences are disallowed: each of the four nouns in combination with v_4 in an NV-type sentence.

Each model is comprised of 4 elements: word-level categories, sentence-level categories, exceptions, and code-length. Both models described here contain two word-level categories, comprising nouns and verbs and two sentence-level categories comprising the two sentence types (NV and VN). The exceptions category discretely specified all the disallowed sentences. In the first model this was an empty set. The code-length specified length of code, in bits, that would be needed to specify models just described and the corpus data given the model structure. The code-length for each sentence in the corpus is consequent on the model structure.

Calculating Code-Length For Each Element

The length of code necessary to specify any object, *i*, is given by:

$$\text{Bits}(i)=\text{Log}_2(1/p_i) \quad [1]$$

where p_i is the probability of object *i*. In many cases described below, p_i can be thought of as choosing one of I options. Where this is the case,

$$\text{Bits}(i)=\text{Log}_2 I \quad [2]$$

This section describes how this formula is applied to calculate the code-length for each section of the model and for the data given the model.

If a language contains r word types and n syntactic categories, then the probability of specifying one distribution of word types into categories is the inverse of the number of ways in which r word types can be distributed between n categories, assuming no empty sets. This given by:

$$\text{Distributions}(r, n)= \sum_{v=0}^{n}(-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [3]$$

The codelength for the word-level element is therefore:

$$\text{Word-level bits}(r, n)$$

$$=\text{Log}_2 \sum_{v=0}^{n}(-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [4]$$

Specifying a particular sentence-level rule (e.g. that a sentence may be of the form NV) is a function of the probability of that sentence type given the number of categories specified in the word-level element. Given that in the artificial language sentences only ever contain two words, there are four sentence types possible from two syntactic categories (NN, NV, VN, VV). The probability of any sentence type (e.g. NV) is therefore 1/4. When this has been specified, the probability any remaining sentence type (e.g. VN) is 1/3. The code-length for specifying two sentence types is therefore:

$$\text{Sentence-level bits}=\text{Log}_2(4)+\text{Log}_2(3) \quad [5]$$

Specifying the cost of an exception is the same as specifying the cost of a sentence. This is done by specifying the cost, in bits, of the first word based on the probability of its occurrence, and the cost of the second word in the same way. The probability of a word's occurrence is the inverse of the total number of possible words. The term to specify the first word in any sentence is therefore:

$$\text{Bits}(i1)=\text{Log}_2(\text{Tw}-T_{e1}) \quad [6]$$

where Bits(*i*1) is the bits required to specify word *i* in the first position, T_w is the total number of word types in the language and T_{e1}, is the total number of words blocked in the sentence initial position as listed in the exceptions category.

The first word specifies which sentence type is being used. The pool of possible words from which the second word must come is therefore reduced to the size of the sentence final category as defined by the sentence type. For example, if the first word in a sentence is a noun, the sentence type must be NV and the second word must therefore be from the category V. The term to specify the second word in a sentence is therefore:

$$\text{Bits}(j2)=\text{Log}_2\,(T_{wc}-T_{e2|1}) \qquad [7]$$

where Bit(j2) is the number of bits required to specify word j in the second position, T_{wc} is the total number of word types in category c, and $T_{e2|1}$ is the total number of words specified in the exceptions element as blocked in position two given the word in position 1. The number of bits for specifying any sentence i,j is simply:

$$\text{sentence bits}_{i,j}=\text{Bits}(i1)+\text{Bits}(j2) \qquad [8]$$

Specifying the code length for each exception is the same as specifying code length for a sentence *given the existing exceptions*. Each exception in a list of exceptions therefore requires slightly fewer bits to code than its predecessor.

It is important to note that these models code corpus data in batch mode – the order in which sentences are coded is not taken into account. A more psychologically realistic (i.e. incremental) algorithm might make use of the fact that frequently occurring words have a higher probability of occurrence and therefore cost less to code.

Simulating recovery from overgeneralization with an artificial language

The models described above were implemented in a computer program. They were then exposed to successively large corpora of sentences from an artificial language which reflected the structure of the transitive/intransitive alternation phenomena found in the CHILDES database (see Table 2, above). A model using raw CHILDES data would have been computationally impossible, but it is important to note that the artificial language closely mirrored the patterns of Table 2. The artificial language is outlined above. In these simulations the word-level categories contained 36 verbs, reflecting the number of verbs in Table 2, and 36 nouns. It was decided to keep the number of nouns equal to the number of verbs in order to avoid disparity between the code-length necessary for different sentence types. There were two sentence-types (NV and VN) reflecting the transitive and intransitive contexts of the verb constructions. Ten verbs were blocked with all 36 nouns for each sentence type (see Table 2), resulting in a total of 720 disallowed sentences.

Two of the four-element models described above were exposed to increasingly large corpora of this language. The first model contained word-level information about the 36 nouns and verbs, and sentence-level information about the NV and VN sentence types, but the exceptions element was empty: it did not contain any information about the 720 disallowed sentences. In this respect it was analogous to a learner who has acquired knowledge of word categories and sentence production rules, but has not learned that some sentences are illegal. This model would therefore be prone to overgeneralizations such as *I disappeared the rabbit*. The second model, by contrast, did contain information about the disallowed sentences. This model therefore required considerably more bits to specify initially, but the number of bits required to specify each sentence of the corpus was fewer. In addition, a language learner who had learned these exceptions would not make the same overgeneralization errors that the first model would. Table 3 shows the relative simplicity of each model for increasingly large corpora as measured by the number of bits necessary to encode the model and the corpus data.

Table 3: Code-lengths of Models 1 and 2 for successively large corpora. Code-lengths in bold show the shorter codes for the corpus size

Corpus Size (sentences)	Model 1: Codelength (bytes)	Model 2: Codelength (bytes)
0	**0.1**	7.6
4000	**45.4**	51.1
8000	**90.8**	94.7
12000	**136.2**	138.3
16000	**181.5**	181.8
20000	226.9	**225.4**
24000	272.2	**268.9**

It can be seen that for relatively small corpora (up to about 16,000 sentences), Model 1 gives a simpler encoding: less bits are required. For a learner who had heard relatively few alternation constructions, therefore, the tendency would be to code the data in these terms, resulting in overgeneralizations. For a more experienced learner, however, the simpler encoding would be that shown by Model 2, which requires less bits to encode relatively large corpora. The model does not produce any language, so there are no accuracy statistics. Rather, it is assumed that the learner produces all the sentences available in the current (shortest) hypothesis are produced, including any that are incorrect.

Conclusions and future directions

These results provide an initial confirmation that simplicity may provide a guiding principle by which some aspects of language may be learned from experience without recourse to a specific language-learning device. However, the simulations presented here are coarse-grained approximations of both the language and the language learner. Children do not process the language in batches of several thousand utterances. The models presented here were neither exposed nor sensitive to different word-type frequencies. A number of further studies which would provide considerably firmer support for the simplicity principle as a driving force for language acquisition suggest themselves.

Firstly, mathematical results show that word-type frequencies are important to the simplicity-driven learner, in that they may be the key as to when it becomes advantageous to posit exceptions to rules. Chater and Vitányi (2001) show that languages are approximately learnable given sufficiently large amounts of data. The CHILDES data in Table 2 therefore provides an indication of the order in which one would expect the learner to cease overgeneralizing words. An examination of children's speech that confirmed this order would be a major step towards providing robust support for the simplicity principle in language. Secondly, it would be useful to compare the timescale of recovery from overgeneralization in children with that of the model. This could be done by an examination of CHILDES database to determine an approximate relation between a child's age and the number of transitive/intransitive alternation constructions to which they have been exposed. It would then be possible to compare the learning rate of the child with that of the model. Again, this would be a useful source of evidence concerning the simplicity principle in language.

In this paper we have suggested that there is sufficient statistical information in the input for a learner to learn quasi-regular alternating structures. These results are achieved by choosing the model of the language which provides the simplest (shortest) description of the linguistic data that has been encountered. These results re-open the question of the viability of language learning from positive evidence under less than ideal conditions, with limited computational resources and amounts of linguistic data available. They therefore also bear, indirectly, on the arguments concerning the balance between nativism and empiricism in language acquisition. More concretely, we suggest that the working hypothesis that the search for simplicity is a guiding principle in language acquisition deserves serious attention.

References:

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533-581.

Bowerman, M. (1982). Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs. *Quaderni di semantica*, 3, 5-66.

Bowerman, M. (1996). Argument structure and learnability: Is a solution in sight? *Proceedings of the Berkeley Linguistics Society*, 22, 454-468.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566-581.

Chater, N. & Vitányi, P. (2001). A simplicity principle for language learning: re-evaluating what can be learned from positive evidence. *Manuscript submitted for publication.*

Chomsky, N. (1980). *Rules and representations*. Cambridge, MA: MIT Press.

Culicover, P. (2000). *Syntactic nuts*. Oxford: Oxford University Press.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, 1, 1-7.

Levin, B. (1993), *English verb classes and alternations*. Chicago: The University of Chicago Press.

Li, M. & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity theory and its applications* (2nd edition). Berlin: Springer.

Lord, C. (1979). Don't you fall me down: Children's generalizations regarding cause and transitivity. *Papers and Reports on Child Language Development*, 17. Stanford, CA: Stanford University Department of Linguistics.

MacWhinney, B. (1987). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2000) *The CHILDES project : tools for analyzing talk*. 3rd ed. London : Lawrence Erlbaum.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, S. (1995). *The language instinct*. Harmondsworth : Penguin.

Quinlan, J. R. & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, 80, 227-248.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry*. Singapore: World Scientific.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30, 50-64.

What's a Science Student to Do?

Tenaha O'Reilly (toreilly@odu.edu)
Danielle S. McNamara (dmcnamar@odu.edu)
and The Strategies Lab
Psychology Department, Old Dominion University
Norfolk, VA 23529 USA

Abstract

This study examined the influence of cognitive ability and student activities on high-school students' science achievement. Students (n=1651) from four high schools in three states were assessed in terms of their cognitive abilities (i.e., science knowledge, reading skill, and metacognitive reading strategies), course involvement, reading interest, and TV habits. Science achievement was measured in terms of students' course grade, comprehension of a science passage, and performance on a statewide standards of learning (SOL) test. Course involvement significantly predicted only course grade, whereas reading interest predicted SOL scores and science passage comprehension. Cognitive abilities and TV habits predicted all three of the student achievement measures. However, the effects of these cognitive variables interacted in interesting ways.

Introduction

In recent years, scientists have become increasingly interested in uncovering factors that are important for predicting educational success (e.g., Buckner, Bassuk, & Weinreb, 2001; Herman & Tucker, 2000). For example, researchers have reliably predicted academic achievement with measures of student personality (Paunonen & Ashton, 2001; Stewart, Bond, Deeds, Westrick, & Wong, 1999), parental influence (Hoge, Smit, & Crist, 1997), social economic status (Jimerson, Egeland, & Teo, 1999), and school demographics (Sutton & Soderstrom, 1999). While this line of research has certainly shed light on how student personality and social factors can impact a child's education, the utility of this information is questionable if the goal of scientific inquiry is to improve scholastic prosperity. Most personality characteristics and social factors are relatively stable; very few introverts quickly turn into extroverts, and even fewer people increase their level of social economic status overnight. In contrast, the investigation of more mutable influences such as cognitive abilities may provide a promising direction for improving academic performance. The purpose of this work was to examine the impact of three cognitive factors on students' success in their science courses: reading skill, science knowledge, and knowledge of metacognitive reading strategies.

It is generally assumed that reading skill is a critical component of academic achievement. Skilled readers are more likely to monitor their comprehension and use active reading strategies such as previewing, predicting, making inferences, drawing from background knowledge, and summarizing (Long, Oppy, & Seely, 1994; McNamara, 2001; Oakhill, 1984; Oakhill & Yuill, 1996). In addition, skilled readers tend to have more knowledge about the world – most likely from reading more often.

Readers' domain knowledge can have a dramatic impact on how well new information is acquired (Bransford & Johnson, 1972). For instance, many school texts are incomplete because they fail to make relations amongst concepts in the text explicit (Beck, McKeown & Gromoll, 1989). Accordingly, domain knowledge can facilitate comprehension by providing the reader with the necessary resources to fill in conceptual gaps (McNamara, Kintsch, Songer, & Kintsch, 1996). In addition, readers with greater prior knowledge are more likely to use effective reading strategies (Lundeberg, 1987) and convey greater interest in the reading material than low-knowledge readers (Tobias, 1994; Zhang, & Zhang, 1996). Collectively, these findings suggest that learners' prior knowledge critically determines their ability to learn and understand new information.

Metacognition refers to the ability to think about, understand and manage one's learning (Schraw & Dennison, 1994). In essence, metacognition is the capacity to monitor comprehension, and the initiative to correct misunderstanding. Recent research has revealed the significance of metacognitive awareness in learning. For instance, learners who score high on measures of metacognition are more strategic (Garner & Alexander, 1989), more likely to use problem-solving heuristics (Artzt & Armour-Thomas, 1992), better at predicting their test scores (Vadhan & Stander, 1994), and generally outperform learners who score low on metacognitive measures (Pressley & Ghatala, 1990).

More importantly, research has demonstrated the value of metacognition in predicting academic achievement. For example, greater metacognitive ability has been linked to grade point average (Everson & Tobias, 1998), math achievement (Maqsud, 1997), and reading skill (van Kraayenoord & Schneider, 1999). Moreover, McNamara and Scott (1999) demonstrated that providing metacognitive reading strategy training improved comprehension and course scores in college-level science courses.

The purpose of this investigation was to examine the influence of science knowledge, reading skill, and metacognitive reading strategies on high school students' achievement in science. While the individual effects of these factors on learning have been examined in separate studies, to the best of our knowledge, no single study has simultaneously measured the influence of all three variables

on students' comprehension and achievement in a classroom setting. Furthermore, we were interested in determining how course involvement, reading interest, and TV habits would influence science performance, and how well these variables would predict student success in comparison to the cognitive factors. Finally, we investigated whether reading skill or metacognitive reading strategies could compensate for knowledge deficits. In this study, science achievement was assessed by the student's science course grade, comprehension of a science passage, and a statewide measure of students' science achievement (Virginia's Standards of Learning, SOL). It was hypothesized that both the cognitive and student activity measures would reliably predict science achievement; but overall, it was hypothesized that the cognitive measures would better predict performance than measures of student activity.

In line with other work (Perfetti, 1989), it was hypothesized that either reading skill or metacognitive reading strategies would compensate for science knowledge. While some researchers have argued that reading skill and domain knowledge can compensate for each other (Perfetti, 1989), there is little consensus as to whether metacognitive reading strategies could make up for meager science knowledge. On the one hand, one might infer that high metacognitive reading strategies could help a learner offset a low level of science knowledge because research has shown that metacognition can compensate various cognitive abilities (Swanson, 1990). On the other hand, others have argued that metacognition has strong knowledge requirements; that is metacognition is not knowledge free (Schwartz & Bransford, 1998) and consequently, one might not expect metacognitive reading ability to compensate for low science knowledge. In any event, the issue is unclear and further investigation is required.

Method

Participants

The sample consisted of 1651 high school students from four schools. Four hundred and ninety-eight students were from an inner city high school in Norfolk, Virginia; 372 were from a rural high school in Americus, Georgia; 364 were from a rural Appalachian high school in Prestonsburg, Kentucky; and the remaining 417 were from a suburban high school in Williamsburg, Virginia. Students' grade level ranged from 9 to 12, and the average age of the students was 16.25 years.

Materials

Metacognitive reading strategy use was measured by a modified version of the Metacognitive Strategies Index (MSI) adapted for use with high-school students (Forget, 1999). The MSI is a 25-item multiple-choice questionnaire which is designed to measure six factors associated with metacognitive reading strategy use: predicting and verifying; previewing; purpose setting; self-questioning; drawing from background knowledge; and summarizing. The Cronbach's Alpha for the MSI was α=.68. Science knowledge was measured with an 18-item multiple choice test on general science information. The test consisted of questions concerning experimental methods, mathematics, and meteorology. Cronbach's Alpha for the science knowledge was α=.63. Reading skill was measured by a modified version of the Gates-MacGinitie reading skill test for grades 10-12. The test consisted of 40 multiple choice questions designed to assess student comprehension on several short text passages. The reliability of the gates-MacGinitie is typically between α=.85-.92 (Phillips, Norris, Osmond, & Maynard, 2002).

Students were given a questionnaire concerning their course involvement, reading interest and TV habits. The participants were required to rate the following statements related to their course involvement on a one to five-point scale: "How much do you enjoy learning science, or scientific concepts?"; "How much time per week do you generally spend reading and studying for this science course?" and "How much effort have you devoted to this science course?". For reading interest, the following questions were asked "How much do you enjoy reading?"; and "How many books do you read each year that are not required by your teachers?". TV habits were assessed by two questions: "How many hours of television do you watch during a school day?"; and "How many hours of television do you watch on the weekend?". The scales were designed such that higher numbers indicated larger amount of the entity in question.

Finally, participants were given an 844-word passage on meteorology (Flesch-Kincaid grade level of 6.7). The passage covered the types and origins of air masses as well as their impact on weather patterns. An accompanying set of 8 multiple choice and 12 open-ended comprehension questions were created for the passage. Cronbach's Alpha for the open ended questions was α=.72, while alpha level for the multiple choice questions was α=.57.

Design and Procedure

The students were tested during regular classroom hours in a 90-minute class period, or two 50-minute class periods, and all testing was conducted near the end of the academic year. The complete set of materials were presented in a single booklet with "stop'' pages inserted between each measure. If a student finished a particular test early, they could recheck their answers, but could not go on to the next section. The participants completed the measures in the following order and time frame: Science passage and questions (20 minutes), Gates reading test (20 minutes), prior knowledge test (10 minutes) MSI (10 minutes), and the student activity questionnaire (5 minutes). At the end of the academic year, the students' science course grade and their Standards of Learning science scores were collected.

Results

The following results were significant at the $p < .001$ level unless noted otherwise. It was verified for all analyses reported here that students' age differences did not alter the pattern of results.

What's More Important?

A factor analysis was conducted to determine whether the predictors used in this study could be grouped into smaller subset of factors (e.g., Cognitive ability, reading interest, etc.). All 10 measures of student ability and activity were entered into the analysis using the principal components method of extraction. Predictors with Eigenvalues over 1 were retained in the analysis, and the Varimax procedure was used as the method of rotation. The analysis revealed four distinct factors that accounted for 68% of the overall variance. Science knowledge, reading skill and metacognitive reading ability loaded on factor 1, *Cognitive Ability* (loadings=.800, .760, .692; Eigenvalue=2.67), and accounted for 27% of the variance. The number of books read and reading enjoyment loaded on factor 2, *Reading Interest* (loadings=.891, .849; Eigenvalue=1.90), and accounted for 19% of the variance. The amount of TV watched on a school day and the amount watched on a weekend loaded on factor 3, *TV Habits*, (loadings=.890, .890; Eigenvalue=1.20) and explained 12% of the variance. Finally course effort, time spent reading and studying the textbook, and enjoyment of learning science loaded on factor 4, *Course Involvement* (loadings=.807, .684, .638; Eigenvalue=1.07), and explained 11% of the variance. Thus, the factor analysis provided support for our initial categorical distinction of the predictors.

The four factors were regressed onto each of the measures of science achievement. For the students' **course grade**, the overall model accounted for 13% of the variance, $F(4,1295)=49.57$. Reading interest did not predict course grade, whereas cognitive ability, $t(1295)=10.15$, $\beta=.263$, and course involvement $t(1295)=9.43$, $\beta=.244$ were strong predictors. TV habits significantly predicted course grade but the relationship was small $t(1295)=-2.43$, $\beta= -.063$, $p= .015$.

For students' **SOL score**, the overall model accounted for 38% of the variance, $F(4,618)=94.09$. Course involvement did not predict SOL scores, whereas cognitive ability, $t(618)=16.24$, $\beta=.516$, TV Habits, $t(618)= -9.23$, $\beta=-.294$, and reading interest $t(618)=5.06$, $\beta=.160$, were significant predictors.

Table 1 Correlations between science achievement and student activities.

Factor	Individual Measure	Course Grade	SOL	Open Ended Comp.	Multiple Choice Comp.
Cognitive Ability	Reading Skill	.24	.58	.64	.53
	Science Knowledge	.25	.59	.55	.51
	Metacognitive Reading Strat.	.20	.15	.26	.24
Course Involvement	Enjoy Learn Science	.18	.16	.13	.14
	Time Reading & Studying	.12	N.S.	N.S.	N.S.
	Effort Given to Course	.30	N.S.	N.S.	N.S.
Reading Interest	Number of Books Read	N.S.	N.S.	N.S.	.11
	Enjoyment of Reading	.12	.16	.14	.16
TV Habits	Hrs. TV School day	-.13	-.34	-.25	-.23
	Hrs. TV Weekend	N.S.	-.27	-.23	-.23

In terms of **science passage comprehension scores**, the model accounted for 42% of the variance for open-ended questions, $F(4,1213)=219.79$, and 33% for multiple-choice questions, $F(4,1292)=158.81$. For both comprehension measures, cognitive ability ($t(1213)=27.44$, $\beta=.600$; $t(1292)=22.56$, $\beta=.514$); reading interest ($t(1213)=5.01$, $\beta=.110$; $t(1292)=5.46$, $\beta=.124$); and TV habits ($t(1213) =-9.83$, $\beta=-.215$; $t(1292)=-9.58$, $\beta=-.218$) were significant predictors, whereas course involvement was not significant.

In summary, cognitive ability and TV habits were significant predictors for all of the student achievement measures. Course involvement reliably predicted only course grade, whereas reading interest reliably predicted SOL scores and science passage comprehension.

Table 1 presents the Pearson correlations between the students' science achievement performance (i.e., course grade, SOL, open-ended and multiple choice comprehension questions) and the 10 predictors used in this study. Correlations are significant at the $p< .001$ level unless specified otherwise. Several trends emerge from the analysis. First, the correlations between science achievement and the individual measures of cognitive ability are moderate to high. In contrast, the correlations between achievement and the individual measures of student activity were generally low. However, there were two exceptions, the amount of effort given to the course was moderately correlated with course grade ($r(1472)=.298$). In fact, of the measures used in this study, effort had the highest simple correlation with course grade. Second, the amount of TV watched on a school day and the weekend (with the exception of course grade) was moderately, but negatively correlated with science achievement. The magnitude of the correlations ranged from small for course grade ($r(1493)=-.125$ to moderate for SOL ($r(693)=-.337$).

Can You Compensate for Low Knowledge?

Our second question was whether either reading skill or metacognitive strategies would compensate for science knowledge. Hence, we conducted ANOVAs for each measure including science knowledge and reading skill in the first set, and science knowledge and metacognitive reading strategies in the second set. (The three variables could not be included in one analysis because there were cells with too few participants.) Students scoring in the top and bottom thirds for each cognitive ability measure were included in the analyses. The dependent variables included course grade, SOL score, open-ended questions, and multiple-choice performance.

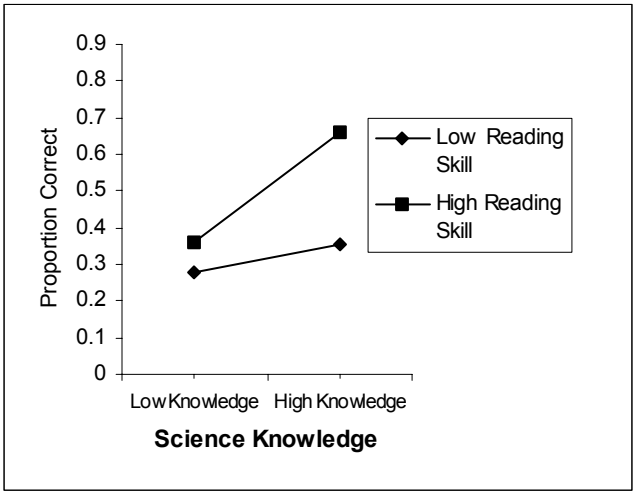


Figure 1. Proportion of multiple-choice comprehension questions correct as a function of science knowledge and reading skill.

The first set of analyses, including science knowledge and reading skill, yielded significant effects for all of the dependent measures. The results were significant at $p<.001$ unless otherwise specified. There was a significant interaction of science knowledge and reading skill only for students' performance on the multiple-choice comprehension questions, $F(4,1368)=7.85$ (see Figure 1). This interaction indicates that neither science knowledge nor reading skill had a major impact on comprehension unless the student possessed both.

The second set of analyses, including science knowledge and metacognitive reading strategies, yielded significant effects of science knowledge for all of the depended measures. Metacognitive strategies was significant for all of the measures except SOL. However, in this case, there was significant interaction of science knowledge and strategy use, $F(4,659)=2.52$, $p=.04$. As shown in Figure 2, greater metacognitive ability helped compensate for a student's low level of science knowledge. [No other interactions were significant.]

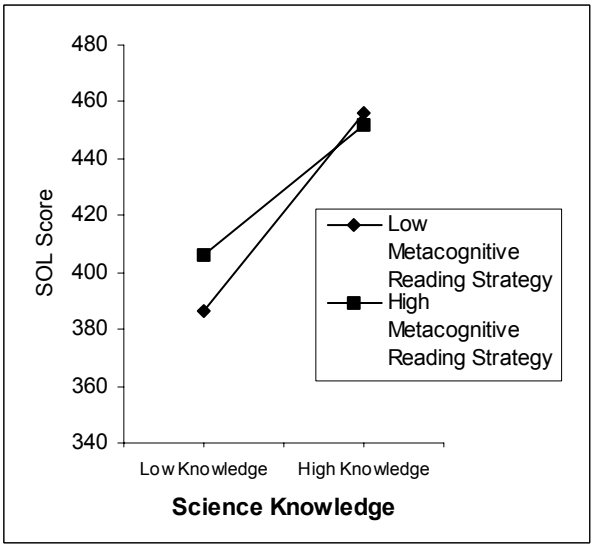


Figure 2. SOL score as a function of science knowledge and metacognitive reading strategy.

Discussion

One goal of the current investigation was to uncover some of the factors that are important in promoting high-school students' science achievement. A factor analysis of our ten measures of abilities and activities revealed that there were four distinct categories of variables: cognitive ability, TV Habits, reading interest, and course involvement. The results indicated that all four factors were important in predicting science achievement; however, some factors differentially predicted the measures of science achievement. Cognitive ability, and TV habits reliably predicted all measures of science achievement, while course involvement reliably predicted only course grade. In turn, reading interest predicted both SOL scores and passage comprehension.

A more detailed examination of the correlations between the individual components of the factors and science achievement revealed that all measures of cognitive ability and TV habits were relatively strong predictors of science achievement, while the individual measures of reading interest and course involvement were generally weak predictors. The major exception was the correlation between course effort and course grade, which proved to be the best single correlation with the students' grade.

The second goal of the study was to determine whether reading skill or metacognitive reading strategies could compensate for science knowledge (see also, Cottrell & McNamara, 2002). With multiple choice questions, science knowledge and reading skill interacted. In this case, neither science knowledge nor reading skill had a major impact on passage comprehension unless the learner had high levels of both cognitive abilities. This interaction is counter to the belief that reading skill and prior knowledge compensate for each other (e.g., Perfetti, 1989). If science knowledge and reading skill were compensatory, one would expect that a high level reading skill would make up for a low level

prior knowledge. Nevertheless, it is notable that the multiple-choice measure was the only dependent measure of science achievement for which an interaction occurred. In the other three cases (Sol score, open ended questions, and course grade), both reading skill and knowledge aided the students, and did not interact. So, in those cases, either reading skill or prior knowledge were beneficial – and thus could compensate for one another. Having both, of course, is the best scenario. Similarly, for the most part, either prior knowledge or metacognitive reading strategies were beneficial to students. In contrast, for SOL scores, metacognitive reading strategies and science knowledge interacted. High-knowledge students did not benefit from reading strategies. Yet, students with low science knowledge were presumably able to compensate for this knowledge deficit with reading strategies. The results support the notion that metacognitive reading strategies can compensate for a low level of domain knowledge.

So, what's a science student to do? The results of this study suggest several things. First, students should simply read more. Research has shown that an increase in exposure to print is associated with an increase in reading skill (see, McNamara, 2001). Accordingly, the current findings support the notion that reading skill is important for science achievement. In fact, reading skill was one of the best single correlates of student performance. Second, students should make informed decisions on the courses they take. For, example if a student is interested in taking biology courses, they should plan to take as many courses related to biology and chemistry in high school as possible. Prior knowledge is important in determining how well new information is learned. Thus the more elementary courses one has in a domain, the easier it will be to learn more advanced courses in the same domain.

However, as we well know, students will often find themselves in courses for which they are ill prepared. In that case, knowing and using metacognitive reading strategies can help the learner to partially overcome knowledge deficits. Hence, the results of this study suggest that students should increase their metacognitive awareness. Unfortunately, students do not automatically engage in such processing (Garner, 1990). Consequently, the solution is to discover and implement techniques that promote metacognitive strategy use (e.g., McNamara & Scott, 1999).

Finally, our findings suggest that parents and students should find a healthy balance between the amount of TV watched and the amount of effort the student puts into the course. Of the measures of student activity, TV habits seemed to be one of the best predictors of science performance: TV viewing was reliably related to all four of our measures of science achievement. However, the relationship between TV viewing and science achievement was negative. This result is congruent with research on TV viewing, which suggests that TV viewing can have a negative impact on reading comprehension (e.g., Koolstra, van der Voort, & van der Kamp, 1997). Conversely, our results underscore the importance of student effort on course performance; students' effort was the best single correlate of course performance. While readers often prefer the path of least resistance (McNamara et al., 1996) it is important to encourage students to expend effort into their academic endeavors.

It is important to note that these results were based on correlation, and therefore should be interpreted with caution. Despite this limitation, the conclusion we draw from this work is that both cognitive ability and student activities are important for science achievement. Moreover, it is important to develop ways to promote reading, and interest in reading, as well as ways to increase course involvement. These findings also recommend that parents should play an active role in educating children to balance their TV viewing and academic endeavors. Finally, the results suggest the need for the development and implementation of strategies to promote metacognitive awareness.

Acknowledgements

We would like to thank the members of the ODU Strategy Lab who helped to conduct this study, including Kim Cottrell, Meghan Depont, Karen Fuller, Erin McSherry, Raymond Morgan, Grant Sinclair, Danny Simmons, Karen Stockstill, and Paul Walker, This project was supported by an NSF IERI grant (REC-0089271) to the second author. Both authors can be contacted at the University of Memphis.

References

Artzt, A., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction, 9, 137-175.

Beck, I., McKeown, M., & Gromoll, E. (1989). Learning from social studies texts. Cognition and Instruction, 6, 99-158.

Bransford, J., & Johnson, M. K. (1972). Contextual prerequisites for understanding some investigations of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 11, 717-726.

Buckner, J., Bassuk, E., & Weinreb, L. (2001). Predictors of academic achievement among homeless and low-income housed children. Journal of School Psychology, 39, 45-69.

Cottrell, K. G., & McNamara, D. S. (2002). Cognitive precursors to science comprehension. Proceedings of the Twenty-fourth Annual Meeting of the Cognitive Science Society. Hillsdale, NJ: Erlbaum

Everson, H. & Tobias, S (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis. Instructional Science, 26, 65-79.

Forget, M. A. (1999). Comparative effects of three methods of staff development in content area reading instruction on urban high school teachers. *Unpublished doctoral dissertation*. Old Dominion University.

Garner, R. (1990). When children and adults do not use learning strategies: Toward a theory of settings. Review of Educational Psychology, 60, 517-529.

Garner, R., & Alexander, P. (1989). Metacognition: Answered and unanswered questions. Educational Psychologists, 24, 143-158.

Herman, K., & Tucker, C. (2000). Engagement in learning and academic success among at-risk Latino American students. Journal of Research & Development in Education, 33, 129-136.

Hoge, D., Smit, E., & Crist, J. (1997). Four Family Process Factors Predicting Academic Achievement in Sixth and Seventh Grade. Educational Research Quarterly, 21, 27-42.

Jimerson, S., Egeland, B., & Teo, A. (1999). A longitudinal study of achievement trajectories: Factors associated with change. Journal of Educational Psychology, 91, 116-126.

Koolstra, C., van der Voort, T., & van der Kamp, L. (1997). Television's impact on children's reading comprehension and decoding skills: A 3-year panel study. Reading Research Quarterly, 32, 128-152.

Long, D., Oppy, B., & Seely, M. (1994). Individual differences in the time course of inferential processing. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 1456-1470.

Lundeberg, M. (1987). Metacognitive aspects of reading comprehension: Studying understanding in legal case analysis. Reading Research Quarterly, 22, 407-432.

Maqsud, M. (1997). Effects of metacognitive skills and nonverbal ability on academic achievement of high school pupils. Educational Psychology, 17, 387-397.

McNamara, D. (2001). Book Review. Reading comprehension difficulties: Processes and Intervention. Journal of Pragmatics, 33, 943-956.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society. Hillsdale, NJ: Erlbaum. pp. 387-392

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43.

Oakhill, J. (1984). Inferential and memory skills in children's comprehension of stories. British Journal of Educational Psychology, 54, 31-39.

Oakhill, J., & Yuill, N. (1996). Higher order factors in comprehension disability: Processes and remediation. In C. Cornaldi & J. Oakhill (Eds.), Reading comprehension difficulties: Processes and Intervention. Mahwah, NJ: Erlbaum.

Paunonen, S., & Ashton, M. (2001). Big Five predictors of academic achievement. Journal of Research in Personality, 35, 78-90.

Perfetti, C. (1989). There are generalized abilities and one of them is reading. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 307-336). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Phillips, L., Norris, S., Osmond, W., & Maynard, A. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. Journal of Educational Psychology, 94(1), 3-13.

Pressley, M, & Ghatala, E. (1990). Self-regulated learning: Monitoring learning from text. Educational Psychologist, 25, 19-33.

Schraw, G. & Dennison, R. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19, 460-475.

Schwartz, D. & Bransford, J. (1998). A time for telling. Cognition and Instruction, 16, 475-522.

Stewart, S., Bond, M., Deeds, O., Westrick, J., & Wong, C. M. (1999). Predictors of high school achievement in a Hong Kong international school. International Journal of Psychology, 34, 163-174.

Sutton, A., & Soderstrom, I. (1999). Predicting elementary and secondary school achievement with school-related and demographic factors. Journal of Educational Research, 92, 330-338.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. Journal of Educational Psychology, 82(2), 306–314.

Tobias, S. (1994). Interest prior knowledge and learning. Review of Educational Research, 64, 37-54.

Vadhan, V., & Stander, P. (1994). Metacognitive ability and test performance among college students. Journal of Psychology, 128, 307-309.

van Kraayenoord, C., & Schneider, W. (1999). Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. European Journal of Psychology of Education, 14, 305-324.

Zhang, K., & Zhang, B. (1996). Impact of interest on text comprehension. Acta Psychologica Sinica, 28, 284-289.

Is There Evidence for Unconscious Reasoning Processes?

Magda Osman (M.Osman@ucl.ac.uk)
Department of Psychology, University College London,
Gower Street, London, WC1E 6BT, UK

Abstract

Current theories of reasoning propose that reasoning is governed by two systems: conscious and unconscious. Conscious processing directs analytical thinking and results in correct responding, while unconscious processing employs heuristics that often leads to poor performance in deductive reasoning tasks. The present study uses a classic propositional task to examine the properties that distinguish conscious from unconscious reasoning. Overall, the study did not find support for dissociable reasoning systems. Instead, the findings suggest that the features exclusively attributed to each system, by dual reasoning theorists, were equally applicable to both.

Dual Process Theories

At present there are three theories of reasoning that have divided the process into unconscious and conscious components: Evans and Over s (1996) Dual process theory; Slomans s (1966) Two systems theory and Stanovich and West s (2000) Two systems theory. Stanovich and West (1998) present a summary of the general attributes that distinguish conscious from unconscious reasoning processes. They propose that unconscious processes are inaccessible, automatic, inert, non-declarative, and non-verbalizable, while conscious processes are accessible, controllable, declarative, and verbalizable. The different characteristics also imply that the two reasoning systems serve different purposes, result in different responses, and encode information differently. Many studies developed to investigate the different systems originate from Wason s (1966) conditional reasoning task. One reason for this is that the general errors individuals make when solving this task have been the impetus for attributing unconscious mechanisms to reasoning. The aim of the present study is to examine the dual processes theories characterization of deductive reasoning using Wason s (1966) conditional reasoning task.

Wason (1966) developed a task (hereafter Wason s selection task) that has now become the mainstay of studies investigating deductive reasoning. It involves a conditional statement if there is a vowel on one side of the card, then there is an even number on the other side . Participants are told that they have to discover whether the statement is true by selecting cards to turn over from an array of four (e.g., E, K, 2, and 5), which are represented in logical notation as (P, ´P, Q, ´Q). The correct solution requires the selection of the E (P) and 5 (´Q) cards, because only this combination provides a means of confirming and falsifying the statement. Typically, only a small proportion of participants solve the task correctly (e.g., 5-10%), while most choose a range of alternative card combinations, the most popular of which is E (P) and 4 (Q).

The appeal of this task comes from the robust results it generates, in particular the regularity with which E and 4 cards are selected. The matching bias theory proposed by Evans (1972) and Evans and Lynch (1973) is the most accepted explanation of this phenomenon. They propose that instead of triggering reasoning processes the selection task incites participants to simply match their card choices with those named in the statement. Evans (1972) developed a paradigm to examine this by presenting participants negated versions of the statement. He found that participants still selected P and Q cards irrespective of the presence of negations in the statement, thus leading to the conclusion that the selection task is solved using simple heuristics. A more detailed account proposed by Wason and Evans (1975) explains the underlying processes that motivate matching behavior. They suggested that reasoning is comprised of two dissociated processes, one of which is unconscious and based on quick-fix strategies that are guided by particular preferences for a response (i.e., biases). The second process is conscious and rationalizes behavior, some of which the reasoner has little control over. These proposals were based on findings from protocols studies (Evans & Wason, 1976; Wason & Evans, 1975), which required participants to provide justifications for their card selections. Participants showed a lack of awareness of the actual processes involved in selecting cards, and rationalizations of their behavior were found to be independent of their actual card selections.

Following from Wason and Evans protocol studies, a variety of techniques have been developed to uncover unconscious reasoning processes e.g., transfer tasks (e.g., Berry, 1983), and attentional biases (Evans, 1996; Evans, Ball & Brooks, 1987; Roberts, 1997; Roberts & Newton, 2001).

The findings from some of these designs suggest that the characteristics attributed exclusively to one or other of the two types of reasoning process are inaccurate. For instance, Berry (1983) reported that participants possessed insight into the processes motivating their card selections, and that this knowledge contributed to transfer of correct responding across different versions of the selection task. This conflicts with the proposal that individuals lack awareness of the processes that contribute to solving the selection task, and that protocols are actually post hoc rationalizations of card choices (Evans and Wason, 1975).

The Present Study

The objective of this study is to examine three of the claims made by the dual process theories of reasoning.

Evans, Ball and Brooks (1987) measured the order in which cards were selected and rejected; they found that decisions were made earlier for card selections than rejections. They proposed that the reasons for this are the result of unconscious biases that motivate participants to focus their attention on cards that match those named in the statement. However, attentional bias has been inferred from measures of decision making/card selecting behavior. There have been a number of studies that have investigated aspects related to attentional bias (e.g., Evans, 1996; Evans, Ball & Brooks, 1987; Dominowski, 1995; Roberts, 1998; Roberts & Newton, 2001), and in general the findings are mixed. One of the objectives of this study is to examine the prediction that attention is first directed to cards that are selected first. Furthermore, there has been no direct attempt to try and separate out attentional processes from decision making processes, and the present study attempts to remedy this.

The three dual process theories characterize unconscious processing as inflexible, and this property has been used to account for the poor rate of correct responding following tutoring on conditional reasoning (e.g., Wason & Johnson-Laird, 1970; Wason & Shapiro, 1971). The present study investigates this effect by including a tutoring session in the experiment and measuring the extent to which performance is improvement in subsequent versions of the selection task.

Stanovich and West (2000) describe the emergence of individual differences within the two reasoning systems. They suggest that matched card selections are motivated by the same unconscious bias, which also implies that unconscious processes are ubiquitous and not subject to variation. By contrast, individual differences occur during conscious processing because participants have overcome the tendency to select matched cards, and have based their card choices from their own construal of the statement, which in turn results in a variety of construals and therefore card combinations. However, this is a rather circular argument, since evidence of individual differences is supported by the view that they only emerge during conscious processing, and similarly, conscious processing is identified by the selection of cards that are not matched to the rule. This study examines the occurrence of individual differences in card selecting behavior during different conditions.

Method

The present study combines a series of methods designed to examine unconscious reasoning that have not been used in combination in previous studies of the selection task. Three techniques in particular have been adapted for the purposes of this study.

Roberts and Newton (2001) developed a rapid response task (hereafter RRT) that limited conscious analytic processing in order to encourage automatic responding in the selection task. During this task participants were shown the example cards for 1 second, and asked to respond yes or no depending on whether they would select the card or not. In the present study participants were asked to study a statement, which based on the typical presentation of the conditional statement in standard versions of the selection task. Then, participants were exposed to the four cards A K 4 and 7 serially, for 90 msec. They were then asked to decide after each card presentation whether they would select the card or not, and told to respond as quickly as possible. Participants were also asked to rate how confident they were of their decision on a 1-7 point scale (1 not confident , and 7 highly confident). One problem that has pervaded this type of design is that while participants are looking at the cards they are also considering their selection, so it is hard to infer attentional biases when the measure might be contaminated by decision making processes as well (Roberts, 1997). It should be noted that the present study does not claim that the method elicits unconscious processing, only that it encourages attentional biases, and attempts to separate them from decision making processes.

This study also includes a tutoring session and uses some of the techniques developed by Green and Larking (1995). In the present study participants were asked to imagine what were the possible outcomes on the underside of each of the four cards when they turned them over. Participants were also asked to suggest what implications the outcomes would have for the conditional statement. After this, the experimenter explained the concept of material implicature and the necessity of falsification in order to solve the selection task correctly.

A generation task was used to measure the extent to which participants understood the concepts they were introduced to during the tutoring session. This design was originally used by van Duyne (1976) and later incorporated into a study by Legrenzi (1980). The general format of this task uses a conditional statement but no given premises (e.g., if ____, then _____.), and participants are required to generate their own statement, devise the examples, and then test the statement.

Participants

Forty-eight graduate and undergraduate students from Brunel University took part in the experiment. Each participant was screened for prior knowledge of the selection task. They were assigned randomly to one of the 48 possible permutations of the four cards presentations in the RRTs.

Procedure

In the present study each participant completed the six tasks, all of which were variations of the standard abstract selection task. Participants were required to solve the tasks in the same order starting with the first RRT (Task 1), 3 versions of the abstract selection task presented in a booklet (Tasks 2-4), followed by a tutoring session, a second RRT (Task 5), and finally, the generation task (Task 6). The instructions in the second RRT task were identical to the first with exception of the actual letters and numbers, and the order in which they were presented for each participant.

Results

Card Selections

Table 1 reports the frequencies of all the cards combinations selected in each of the six tasks. A log-linear analysis was favored over Pearson s chi-squared in order to determine statistical regularities in the data.

The following analyses of card selecting behavior across tasks are based on a collapsed version of Table 1. This included separate frequencies for the main card selections [P, P Q, P ~Q, Q, P Q ~Q], while the remaining figures were classed as residuals. Significant differences were found between the frequencies of card combinations selected in the six tasks, $G^2 = 180.320$ (25), $p < .0001$. On closer inspection there were no significant differences between cards selected in the three tasks presented in the booklet (Tasks 2-4), G^2 (8) = 7.960, $p > .43$, which suggests that the source of difference was based specifically on responses to the RRTs and the generation task.

Table 1: Frequencies of card selections for each of the six tasks

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
P	12	4	3	5	14	4
PQ	17	34	37	37	6	8
P~P	1	0	1	0	2	2
P~Q	0	1	0	1	13	17
~P	4	1	0	0	0	0
~P~Q	1	0	2	2	0	1
~PQ	0	0	0	0	2	0
P~PQ~Q	2	1	0	0	0	1
P~PQ	2	0	0	0	2	1
PQ~Q	2	0	0	0	6	11
P~P~Q	2	0	0	0	1	2
Q	2	7	4	3	0	0
~Q	0	0	1	0	0	0
Q~Q	3	0	0	0	2	1

Participants card selections in both RRTs were compared, this revealed a highly significant difference between the two tasks, G^2 (12) = 32.28, $p < .001$. A comparison between the frequency of cards selected in the second RRT, and the proceeding generation task revealed no significant difference, G^2 (4) = 8.257, $p > .08$. Thus, the responses to both tasks presented after tutoring were not statistically different. A further analysis between card choices for the first RRT and its proceeding task (the first booklet task) showed there was a highly significant difference, G^2 (5) = 28.315, $p < .0005$.

Analyzing correct card choices across the six task revealed that significantly more correct responses were made in tasks after the tutoring session, than preceding it, G^2 (5) = 63.013, $p < .0001$, thus suggesting that tutoring facilitated correct responding.

Decision Latencies

The mean decision times for card selections and rejections between the two RRTs were compared. This analysis showed there was no significant difference, G^2 (1) = 0.354, $p > .55$. An analysis based only on decision times of participants choosing the P and Q cards in both RRTs was conducted, a summary of the mean decision times for card selections and rejections is presented in Table 2.

Table 2: Mean decision times for cards selected (PQ) and rejected (´P´Q) in RRT 1 and RRT 2

	RRT 1	RRT 2
Card selections	10051.50 msec	8999.83 msec
Card rejections	9665.59 msec	8877.94 msec

There was no significant difference between decision times for rejected and selected cards, G^2 (1) = 1.513, p>.21. A further analysis was carried out on responses to individual cards. A 2 (response type: selection vs. rejection) x 4 (card type: P, ´P, Q, ´Q) analysis of variance (ANOVA) revealed no significant main effect for response type in the first RRT1, F (1, 46) = .003, MSE = 61776.125, p>.95, and the same was found for RRT2, F (1, 46) = 1.302, MSE = 28899003, p>.337. There was also no significant main effect for card type for RRT1, F (3, 138) = .803, MSE = 91337103, p>.504, and RRT2, F (3, 138) = .817, MSE = 7774520, p>.506. Finally, there was no significant interaction between response types and card type for either tasks, RRT1 F (3, 138) = 1.538, MSE = 14650821, p>.230, and RRT2 F (3, 138) = 1.928, MSE = 19278503, p>.196. These finding suggest that there is no difference between the decision times for cards selected and rejected and does not support Evans, Ball and Brooks (1987) claim that participants make earlier decisions for cards they select.

Along with measurements of decision times, confidence ratings for each decision made were recorded. The overall ratings were not significantly different across both tasks, G^2 (42) = 48.043, p>.24, further analyses were carried out comparing both RRTs based on ratings for individual cards. There was no significant difference between the ratings for the P card, G^2 (6) = 5.218, p>.5; ´P card, G^2 (6) = 4.065, p>.5; and the ´Q card, G^2 (6) = 3.112, p>.5. However participants responded with higher confidence ratings for the Q card in the second RRT compared to the first RRT, and this was statistically significant, G^2 (6) = 15.929, p< .01.

Analyses of confidence ratings between RRTs and within each RRT, for both rejected and selected cards, revealed significant differences only for the Q card. Participants were significantly more confident when deciding to select the Q card, G^2 (6) = 15.209, p<. 01, and to reject it, G^2 (6) = 13.055, p<.05 in the second RRT which proceeded tutoring. In the selection task literature the Q card has been thought to generate misunderstandings, which may account for the significant results found for confidence and latency measures based on this card.

Tutoring

During the tutoring sessions participants were asked to consider the possible outcomes (i.e., true, false) for the statement based on information represented on the underside of each card. This technique was used to gauge participants initial understandings of the statement and cards.

All the participants assessed the statement correctly according to the outcomes of information represented on the underside of the P card. The majority of participants reported that each outcome from turning the ´P card was irrelevant and had no consequences for the statement, which is an incorrect assumption. With the exception of one, the remainder believed that turning the Q card to reveal a P would imply that the statement was true, which is also a commonly held misconception. Approximately half correctly assumed that discovering a P on the underside of the ´Q card would suggest the statement was false.

In sum, participants have a correct understanding of the P card, and they also appreciate that the ´Q card can falsify the statement, but misunderstand its relevance, evidenced in its absence from most participants card choices prior to tutoring. The Q card was the most misunderstood, and directly related to participants difficulty in appreciating that a bi-conditional interpretation (e.g., if and only if there is a vowel, then there is an even number, which also implies that if there is an even number, then there is a vowel) could not be assumed for the conditional statement.

The data from the tutoring sessions also suggest that participants misunderstandings of the cards did not correspond to previous responses in the booklet (Tasks 2-4). To illustrate, approximately 75% of participants selected the PQ card combination in the booklet, the corresponding misconception entails assuming that turning a Q card and discovering a P would also verify the statement, and that turning the same card over to reveal a ´P would in turn suggest the statement was false. However, comparing participants prior card selections revealed that they displayed a variety of misconceptions, and there was no significant relationship between particular card choices (i.e., P and Q) and its corresponding misinterpretation.

Discussion

The present study investigated unconscious deductive processes based on the claims made by the three theories, and the findings strongly imply that the characterisation of unconscious processes is inaccurate. The findings also challenge the extent to which unconscious and conscious processes can be considered as dissociated. However, it could be argued that the present study did not adequately demonstrate unconscious reasoning processes, and this is the reason

why the claims were not supported. Certainly there is some doubt over what the methods used presently actually demonstrate, but it was thought that the most appropriate method of examining the proposals of dual process theories was to use similar types of task designs, the findings of which the theories have used to support their claims.

Card Selections in the Rapid Response Tasks

The RRT tasks were designed to separate out attentional processes from decision making processes. Thus, the brief exposure of the cards did not allow participants to think about selecting cards while viewing them. The less restricted interval for choosing enabled participants to reflect on their choices under some degree of uncertainty as to what cards they saw.

The analyses suggested that participants differed in their card selections during RRTs. If unconscious processes are inflexible, then there should be a correspondence between the cards selected in both RRTs, and tutoring should have no effect, however this was not found in the present study. Instead, the findings suggest that tutoring influenced reasoning processes employed in restricted as well as in free time tasks, implying that unconscious processes are not inflexible.

Comparisons between card choices in the first RRT and its proceeding task, which was a free time version presented in the booklet, revealed significant differences. Furthermore, 17/48 participants selected matched card selections under restricted time conditions, compared to 34/48 in the first booklet task. If matched card selections are indicative of unconscious reasoning processes then the proportion of matched card selections should be the same for both tasks. 17/48 participants selected the same card combinations in the first RRT and the first booklet task, (compared with 24/48 consistent card selections between the second RRT and the generation task), however, 13/17 participants selected matched cards in these tasks. While the later result lends some support to dual process theorists view of matched card selections, the other findings reported here provide a stronger case against their proposals.

Decision Latencies

Attentional bias has been proposed as an explanation for the longer latencies found for cards selected than rejected in inspection time studies (e.g., Evans, 1996; Evans, Ball, & Brook, 1987). The present findings suggest that participants decision times were not markedly different for different types of card selecting decisions. There were, however, differences between overall responses latencies in the RRTs. Participants made quicker decisions during the second RRT compared to the first. The tutoring participants received before the second RRT may have influenced this result, because they were better informed about the task requirements. Furthermore, there were no differences between the two RRTs based on confidence ratings. However, the only significant difference was found for ratings of the Q card, participants were more confident of their decisions during the second RRT compared to the first. One reason for this may have been the tutoring received prior to the second RRT, suggesting that an increase in understanding also results in an increase in confidence.

Tutoring

There have been many examples of unsuccessful attempts to tutor participants on conditional reasoning (e.g., Wason, 1968; Wason & Johnson-Laird, 1970; Wason & Shapiro, 1971). Wason (1968) first introduced remedial procedures or 'therapies' to invoke insight into the task. Johnson-Laird and Wason (1970) proposed that participants failed to solve the task correctly because there was a disassociation between participants selection and evaluation processes. Wason and Johnson-Laird (1970) suggested that selecting is an active process and occurs immediately before evaluation because the evaluation process is effortful and more cognitively demanding.

The lack of transfer reported in tutoring studies has been used to demonstrate dissociations between conscious and unconscious reasoning processes. Participants inability to adopt new concepts, taken together with the fact that they revert to previous card choices, typically P and Q, suggest that either the methods of tutoring are inadequate, or the processes guiding card selections in the abstract task are inflexible. The results from the present experiment challenge both views.

A reduction in the proportion of matched card choices (i.e., P Q) following tutoring, and an increase in correct card selections (i.e., P~Q) suggest that tutoring was effective, and that participants reasoning processes are not inflexible. Also, individual differences were revealed in the tutoring session, indicating that participants held a variety of misconceptions of the statement and the cards, which were corrected following tutoring. Moreover, many participants did not share the same misconceptions despite having selected the same card combinations in previous tasks. Thus, the selection of P and Q cards does not imply that participants have the same understanding of the cards, or that they are employing the same underlying reasoning process.

Conclusions

The findings from this study do not support the claims made by dual process theories of reasoning. However, it is not possible to rule out the possibility that

unconscious processes are involved in reasoning. This cautionary approach is based on problems concerning methodology. The techniques used to expose unconscious reasoning processes are not sufficiently refined to decide whether the description of the processes is inaccurate and that unconscious processes are still present, or whether there are actually no unconscious processes in reasoning.

Acknowledgments

The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. The work was part of the programme of the ESRC Research Centre for Economic Learning and Social Evolution.

References

Berry, D. (1983). Metacognitive experience and transfer of logical reasoning. *Quarterly Journal of Experimental Psychology, 35A*, 39-40.

Dominowski, R. L. (1995). Content effects in Wason's selection task. In S. E. Newstead. &. J. S. B. T.

Evans. (Eds.), *Perspectives on thinking and reasoning* (pp. 41-65). Hove, England: Lawrence Erlbaum Associates.

Evans, J. S. B. T. (1972). Interpretation and 'matching bias' in a reasoning task. *Quarterly Journal of Experimental Psychology, 24,* 193-199.

Evans, J. S. B. T. (1989). *Biases in human reasoning: Causes and consequences*. London: Lawrence Erlbaum Associates Ltd.

Evans, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87,* 223-240.

Evans, J. S., B.T., Ball, L.J., & Brooks, P.G. (1987). Attention bias and decision order in a reasoning task. *Quarterly Journal of Experimental Psychology, 51A,* 811-814.

Evans, J. S. B. T., & Lynch, J. (1973). Matching bias in the selection task. *British Journal of Psychology, 64,* 391-397.

Evans, J. S. B. T., & Wason, P.C. (1975). Dual processes in reasoning? *Cognition, 3*, 141-154.

Evans, J. S. B. T., & Wason, P.C. (1976). (1976). Rationalisation in a reasoning task. *British Journal of Psychology*, 63, 205-212.

Green, D. W., & Larking, R. (1995). The locus of facilitation in the abstract selection task. *Thinking and Reasoning, 1*, 183-199.

Johnson-Laird, P. N., & Wason, P.C. (1970). Insight into a logical relation. *Quarterly Journal of Experimental Psychology, 22A*, 49-61.

Legrenzi, P. (1980). Relations between language and reasoning about deductive rules. In G. B. Flore. D'Arcais, and Levelt, W. J. M. (Eds.), *Advances in psycholinguistics*. Amsterdam: North Holland.

Roberts, M. J. (1998). Inspection times and the selection task: Are they relevant? *Quarterly Journal of Experimental Psychology, 51*, 781-810.

Roberts, M. J., & Newton, E. J. (2001). Inspection times, the change task, and the rapid-response selection task. *Quarterly Journal of Experimental Psychology,* 54, 1031-1048.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Stanovich, K. E., & West, R. F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate. *Behavioral & Brain Sciences, 22*, 645-665.

Van Duyne, P. C. (1976). Necessity and contingency in reasoning. *Acta Psychologica, 40,* 85-101.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I* (pp. 135-151). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Wason, P. C., & Evans, J.St.B.T. (1975). Dual processes in reasoning? *Cognition, 3,* 141-154.

Wason, P. C., & Johnson-Laird, P, N. (1970). A conflict between selecting and evaluating information in an inferential task. *British Journal of Psychology, 61*, 509-515.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23A*, 63-71.

A Unified Model Of The Origins Of Phonemically Coded Syllable Systems

Pierre-yves Oudeyer
Sony Computer Science Lab, Paris
e-mail : py@csl.sony.fr

Abstract

Human sound systems are invariably phonemically coded, which means that there are parts of syllables that are re-used in other syllables. It is one of the most primitive compositional system in language. To explain this phenomenon, there existed so far three kinds of approaches : "Chomskyan"/cognitive innatism, morpho-perceptual innatism and the more recent approach of "language as a complex cultural system which adapts under the pressure of efficient communication". We proposed in (Oudeyer 2002) a new hypothesis based on a low-level model of sensory-motor interactions, characterized by the absence of functional pressure and the use of very generic neural devices. This paper presents a unified model of the origins of syllable systems which does allow a comparison of the different hypothesis on the same ground. We show that our hypothesis is the only one to be sufficient, and that all others are not necesary. Moreover, the model we present the first that shows how a population of agents can build culturally a complex sound systems without the assumption that they already share a phonemic repertoire.

What does explain phonemically coded syllable systems ?

Human sound systems have very particular properties. Perhaps the most basic is that they are phonemically coded. This means that syllables are composed of re-usable parts. These are called phonemes. Thus, syllables of a language may look rather like la, li, na, ni, bla, bli, etc ... than like la, ze, fri, won, etc .... This might seem unavoidable for us who have a phonetic writing alphabet, but in fact our vocal tract allows to produce syllable systems in which each syllable is holistically coded and has no parts which is also used in another syllable. Yet, as opposed to writing systems for which there exists both "phonetic" coding and holistic/pictographic coding (for e.g. Chinese), all human languages are invariably phonemically coded.

The question is then : Why is this so ? How did it appear ? What are the genetic, glossogenetic/cultural, and ontogenetic components of this formation process ? These questions are of particular interest and generality since phonemic coding is a form of primitive compositionality. Compositionality is thought to be the keystone of syntax, and thus understanding how it appeared might help a lot to understand syntactic languages which make humans unique. Several approaches have already been proposed in the literature.

The first one, known as the "post-structuralist" Chomskian view, defends the idea that our genome contains some sort of program which is supposed to grow a language specific neural device (the so-called Language Acquisition Device) which knows a priori all the algebraic structures of language. This concerns all aspects of language, ranging from syntax to phonetics (Chomsky and Halle, 1968). In particular this neural device is supposed to know that syllables are composed of phonemes which are made up by the combination of a few binary features like the nasality or the roundedness. Learning a particular language only amounts to the tuning of a few parameters like the on or off state of these features. It is important to note that in this approach, the innate knowledge is completely cognitive, and no reference to morpho-perceptual properties of the human articulatory and perceptual apparatuses appears. This view is becoming more and more incompatible with neuro-biological findings (which have basically failed to find a LAD), and genetics/embryology which tend to show that the genome can not contain specific and detailed information for the growth of so complex neural devices.

Another approach is that of "morpho-perceptual" innatists. They argue (Stevens 1972) that the properties of human articulatory and perceptual systems explain totally the properties of sound systems. More precisely, their theory relies on the fact that the mapping between the articulatory space and the acoustic and then perceptual spaces is highly non-linear : there are a number of "plateaus" separated by sharp boundaries. Each plateau is supposed to naturally define a category. Hence in this view, phonemic coding and phoneme inventories are direct consequences of the physical properties of the body. Yet, it seems that there are flaws to this view : first of all, it gives a poor account of the great diversity that characterize human languages. All humans have approximately the same articu-

latory/perceptual mapping, and yet different language communities use different systems of categories. One could imagine that it is because some "plateaus"/natural categories are just left unused in certain languages, but perceptual experiments (Kuhl 2000) have shown that very often there are sharp perceptual non-linearities in some part of the sound space for people speaking language L1, corresponding to boundaries in their category system, which are not perceived at all by people speaking another language L2. This means for instance that Japanese speakers cannot hear the difference between the "l" in "lead" and the "r" in "read". As a consequence, it seems that there are no natural categories. This paper will provide quantitative evidence that the morpho-perceptual innatism hypothesis is not a satisfying candidate.

A more recent approach proposes that the phenomena we are interested in come from self-organization processes under functional pressures occurring mainly at the cultural and ontogenetic scale. The basic idea is that sound systems are good solutions to the problem of finding an efficient communicative system given articulatory, perceptual and cognitive constraints. And good solutions are characterized by the regularities that we try to explain, in particular phonemic coding. This approach was initially defended by (Lindblom 1992) who showed for example that if one optimizes the energy of vowel systems as defined by a compromise between articulatory cost and perceptual distinctiveness, one finds systems which are phonemically coded which means that some targets composing syllables are re-used (note that Lindblom presupposes that syllables are sequences of targets, which we will do also in this paper). Yet, these results were obtained with very low-dimensional and discrete spaces, and it remains to be seen if they remain valid when one deals with realistic spaces.

These experiments were a breakthrough as compared to innatists theories, but provide unsatisfying hypothetical explanations : indeed, they rely on explicit optimization procedures, which never occur as such in nature. There are no little scientists in the head of humans which make calculations to find out which vowel system is cheaper. Rather, natural processes adapt and self-organize. Thus, Lindblom's model does not really provide explanations, and one has to find the processes which formed these sound systems, and can be viewed only a posteriori as optimizations. This has been done for the questions of vowel inventories regularities : indeed, in spite of the fact that our vocal tract allows us to produce thousands of different vowels, languages of the world use rarely more than 10 of them, and most often 5 of them. Moreover, among these actually used vowels, some of them appear very often (e.g. [a], [i] or [u] in 89 percent of languages) and others are very rare (e.g. [y]). Lindblom proposed a model which again optimized motor and perceptual constraints, which predicted these regularities. (de Boer 2001) developed an explanatory model in which the basic processes which do produce realistic vowel systems are imitation behaviors among humans/agents. He built a computational model which consisted of a society of agents playing culturally the so-called "imitation game". Agents were given a physical model of the vocal tract, a model of the cochlea, and a simple prototype based cognitive memory. (Oudeyer 2001b) extended this model by letting agents produce complex utterances (syllables), and showed how realistic phonotactic regularities (e.g. the sonority hierarchy principle, the high occurrence of CV syllables, etc...) could emerge without being explicitly programmed in. Yet, as far as phonemic coding is concerned, which is the focus of interest in this paper, (Oudeyer 2001b) does not provide explanations : possible phonemes that agents can use were pre-given and phonemic coding were pre-programmed. There is clearly a need to extend this model by not giving initially a limited and discrete set of possible phonemes and by not coding in phonemic coding. This is what we are going to present in this paper. Interestingly, the solution will rely on a model by (Oudeyer 2002) which was initially developed to explore the last hypothesis of phonemic coding.

Indeed, (Oudeyer 2002) is so far the only truly explanatory model for the phenomenon of phonemic coding. The hypothesis it proposes is that phonemic coding might be a non-functional consequence of sensory-motor coupling. "Non-functional" means that as opposed to models presented in last paragraph, there is no pressure of efficient communication. Phonemic coding, which is indeed useful to develop efficient communication system, yet would not have appeared for this task but may have been recruited only afterwards, being available "by chance" : this kind of phenomenon is sometimes called "exaptationism". The model is at a lower-level than others since it uses neural cortical maps, and their dynamics, coupling perception and action, provides one explanation for phonemic coding. (Oudeyer 2002) explained that this model was not incompatible with functional models such as (de Boer 2000, Oudeyer 2001b), but rather could be a possible manner to bootstrap imitation games. This is what we are going to show in this paper by integrating (Oudeyer 2001b) and (Oudeyer 2002). A problem appeared in previous research when one tried to have agents play imitation games with complex utterances : there were two levels, i.e. the level of articulatory targets/phonemes, and the level of syllables (sequences of phonemes). With simple binary feedback signal, it was difficult to know what kind of errors one agents may have done : wrong number of phonemes ? right number but one is badly imitated ? one phoneme is unknown ? or

is this the new complex sound which is unknown ? In the model presented here, these problems disappear since the low level of targets/phonemes works without supervision. Finally, the model is flexible enough to allow an implementation of all the non-cognitive innatist hypothesis' : morpho-perceptual innatism, functionalism of Lindblom, and exaptationism of Oudeyer.

We will first present an overview of the model in (Oudeyer 2002), then extend it so as to unify it with (Oudeyer 2001b), and finally show how the tuning of some parameters allows to instantiate the various hypothesis concerning the origins of phonemic coding. Then we will present results evaluating each hypothesis.

The coupled neural maps model

The model is based on topological neural maps. This type of neural network has been widely used for many models of cortical maps (Morasso et al., 1998). It relies on two neuroscientific findings (Georgopoulos 1988) : on the one hand, for each neuron/receptive field in the map there exist a stimulus vector to which it responds maximally (and the response decreases when stimuli get further from this vector) ; on the other hand, from the set of activities of all neurons at a given moment one can predict the perceived stimulus or the motor output, by computing what is termed the population vector (see Georgopoulos 1988) : it is the sum of all preferred vectors of the neurons ponderated by their activity. When there are many neurons and the preferred vectors are uniformly spread across the space, the population vector corresponds accurately to the stimulus that gave rise to the activities of neurons, while when the distribution is inhomogeneous, some imprecisions appear. (Oudeyer 2001a) showed that this imprecision allows to explain the well-known phenomenon of "perceptual magnet effect" (Kuhl 2000), which is a perceptual warping of the acoustic space. Moreover, the neural maps are recurrent, and their relaxation consists in iterating the coding/decoding with the population vector : the imprecision coupled with positive feedback loop forming neuron clusters provides well-define non-trivial attractors which can be interpreted as (phonemic) categories.

There are two neural maps : one articulatory which represents the motor space, and one acoustic which represent the perceptual space. The two maps are fully connected to each other with symmetric weights. These weights are supposed to represent the correlation of activity between neurons, and allow to perform the double direction acoustic/articulatory mapping. They are learnt with a hebbian learning rule.

The network is initially made by initializing the preferred vectors of neurons to random vectors following a uniform distribution. Part of the initial state can be visualized by plotting all the preferred vectors as in one of the upper squares of figure 1 which represents the acoustic maps of two agents (the perceptual space is 2-dimensional, and points represents the preferred vectors of neurons). One can also visualize the initial attractors of the acoustic neural maps : the lower squares of figure 1 show examples, in which each arrow has its ending point being the population coded vector after one iteration of the relaxation rule with initial activation of neurons corresponding to the population vector represented as the beginning of the arrow. What one can notice is that initially, attractors are few, trivial and random (most often there is only one).

Then there is a learning mechanism used to update the weights/preferred vectors in the two neural maps when one agent hears a sound stimulus which is represented by a temporal sequence of feature vectors, typically corresponding to the formants of the sound at a moment t (formants are the frequencies for which there is a peak in the power spectrum). For each of these feature vectors, the activation of the neurons in the acoustic map is computed, which propagates to the motor map. Then, each neuron of each map is updated so as to be a little bit more responsive to the perceived input next time it will occur (which means that their preferred vectors are shifted towards the perceived vectors).

The agents in this model produce dynamic articulations. These are generated by choosing N articulatory targets, and then using a control mechanism which drives the articulators successively to these targets. In the experiments presented here, N=3 for sake of simplicity. The choice of the articulatory targets is made by activating successively and randomly 3 neurons of the articulatory map. Their preferred vectors code for the articulatory configuration of the target. Finally, gaussian noise is introduced just before sending the target values to the control system. By default, the variance of the gaussian equals 5 percent of the extent of each articulatory dimension.

When an articulation is performed, a model of the vocal tract is used to compute the corresponding acoustic trajectory. There are two models. The first one is abstract and serves as a test model to see which properties are due to the particular shape of the articulatory/acoustic mapping and which are not. This is simply a random linear mapping between he articulatory space and the acoustic space.

The second model is realistic in the sense that it reproduces the human articulatory to perceptual mapping concerning the production of vowels. We model only vowels here for sake of computational efficiency. The three major vowel articulatory parameters are used : (Ladefoged and Maddieson, 1996) tongue height, tongue position and lip rounding. To produce the acoustic output of an articulatory configuration, a simple model of the vocal tract was used, as described in (de Boer, 2000), which generates the first and second effective formants which are

known to represent well human perception of vowels (de Boer, 2000). This model does not allow to deal with consonants, but is enough to investigate at least the phonemic coding of vowel targets.

The experiment presented consists in having a population of agents (typically 20 agents) who are going to interact through the production and perception of sounds. They are endowed with the neural system and one of the articulatory synthesizers described previously. They interact by pairs of two : at each round, one agent is chosen randomly and produces a dynamic articulation according to its articulatory neural map as described earlier. This produces a sound. Then another random agent is chosen, perceives the sound, and updates its neural map with the learning rule described earlier.

Let us describe first what we obtain when agents use the abstract articulator. Initially, as the receptive fields of neurons are randomly and uniformly distributed across the space, the different targets that compose the productions of agents are also randomly and uniformly distributed. What is very interesting, is that this initial state situation is not stable : rapidly, agents get in a situation like on figures 2 which corresponds to figures 1 after 1000 interactions in a population of 20 agents. These shows that the distribution of receptive fields is not anymore uniform but clustered. The associated point attractors are now several, very well-defined, and non-trivial. Moreover, the receptive fields distribution and attractors are approximately the same for all agents. This means that now the targets that agents use belong to one of well-defined clusters, and moreover can be classified automatically as such by the relaxation of the network. In brief, agents produce phonemically coded sounds. The code is the same for all agents at the end of a simulation, but different across simulations due to the inherent stochasticity of the process.

Now, (Oudeyer 2002) showed that when you use the realistic articulatory synthesizer, you get additionally vowel systems (defined as the set of point attractors) which do follow very well the tendencies observed in human languages. As a consequence, this model proposes and show the plausibility of the hypothesis : phonemic coding and the existence of shared categorical systems might be a result of the dynamic properties of very generic neural tissues (the same maps can be used for hand-eye coordination for instance), but which particular categories appear is due to the particular shape of the articulatory to perceptual mapping (but this alone is not necessary for phonemic coding, and we will argue here that is it also not sufficient).

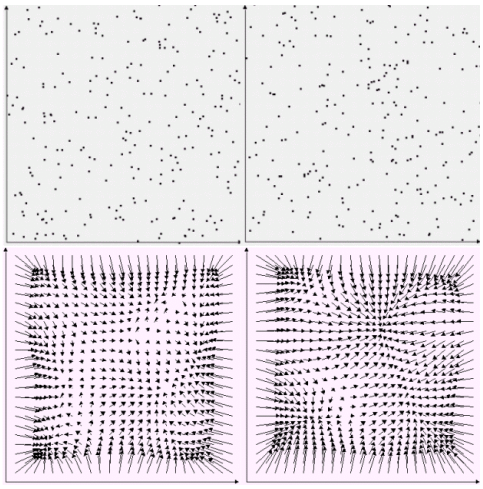


Figure 1: Acoustic 2-D neural maps at the beginning (top), and associated population vector function (bottom) for two agents

Integration within a functional model of the origins of syllable systems

In the coupled neural maps model, there was no shared repertoire of complex sound categorization which was constructed. Obviously, when one agent would hear a complex utterance, the perceptual trajectory would go through zones of the space each belonging to the basin of attraction of a category. But this does not allow to decode appropriately for instance complex utterances made of articulatory targets whose basins of attractions are not connex : the interpolation taking place during the actual articulation will lead to articulatory and perceptual trajectory who go through basins of attraction of categories which do not correspond to any of the initial targets. As a consequence, if one wants to have a model of the origins of syllable systems, it is necessary to add another mechanism. Here this mechanism will be just the one described in (Oudeyer 2001b), and as opposed to the coupled neural maps, is functional.

Basically, agents are going to play the imitation game (de Boer 2000). They possess the two neural maps presented earlier, which work in the same manner. Additionally, each of them have repertoires of syllables (here sequences of N=3 targets), and one game consists in having one agent, the speaker, choose one of its items, then utter it, and then have another random agent, the hearer, try to imitate it by producing the closest syllable in its repertoire. After the imitation, the speaker categorizes the utterance he heard and checks if it corresponds to the category of the syllable he pronounced initially. He then gives a binary (good or bad) feedback to the hearer. The items of their memories have scores (num. of times used successfully / num. of times

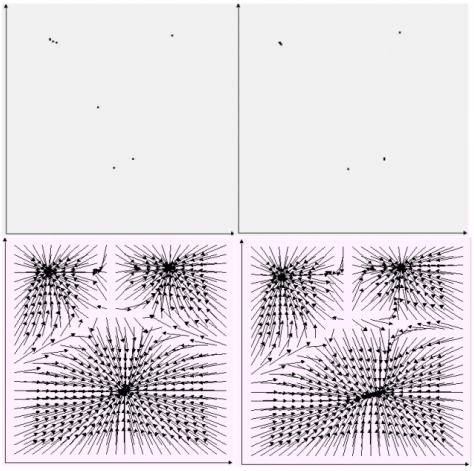


Figure 2: Neural maps and attractors after 1000 interactions, corresponding to the intial states of figure 1)

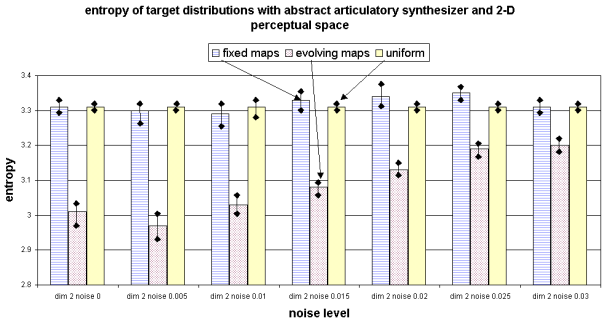


Figure 3:

used). This is used to prune the syllables which are not efficient. Initially, repertoires are empty. They grow either by imitation (agents hear a syllable that they can not imitate and yet have used a usually efficient syllable prototype), or by invention. Inventing a syllable consist in choosing randomly N=3 targets by activating three random neurons of the articulatory map. As the receptive field of these neurons are initially uniformly spread across the space, inventions at the beginning of simulations will produce syllables whose targets are uniformly spread across the space. This means exactly non phonemically coded.

This model can be tuned to show that only our non-functional hypothesis is plausible in a realistic explanatory setting : this can be done by using a realistic synthesizer or not, coupled with the ability to deactivate the learning rule of neural maps (neurons do not update their receptive fields when they perceive a stimulus). If the use of a realistic synthesizer with a deactivated learning rule does not provide phonemic coding, this does show that the

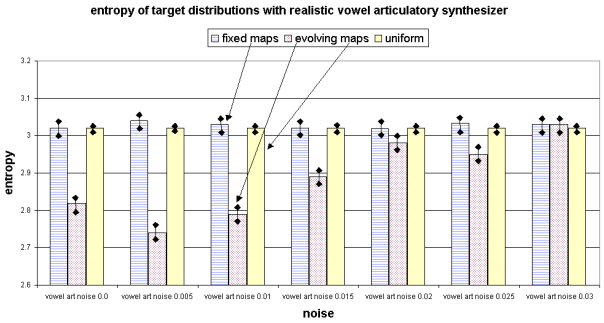


Figure 4:

morpho-innatist hypothesis is not sufficient, and as (Oudeyer 2002) showed it was not necessary either, the conclusion is that this is not a good candidate (in fact, here we would show even more : the combination of morpho-perceptual constraints and functionalism is not sufficient). If we use an abstract synthesizer with a deactivated learning rule, then if we do not get phonemic coding, Lindlom's functionalism is not sufficient either (and again, we showed it was not necessary). Then, if each time we activate the learning rule we do get phonemic coding, this will confirm the plausibility of the exaptationist hypothesis of (Oudeyer 2002).

A measure of phonemic coding of a syllable systems was developed in (Oudeyer 2002). This consists in making models of the distributions of targets, based on parzen windows : at points corresponding to a the crossings of a regular grid, one approximates the local probability density function by averaging the values of a gaussian function (centered on this point) taken at all points coded by each targets. This is very similar to making multi-dimensional histograms, and counting how many targets fall in each bin. Yet, using gaussians gives a fuzzy binning whose choice of variance is much easier than the choice of bin size in the case of histograms. This approximation is used to compute the entropy of target distributions : if they are uniformly spread, entropy is maximal, and the more they are clustered, the lower the entropy. Thus, this is a way to automatically monitor the clusteredness of targets, and so how much a system is phonemically coded.

A first series of experiment [1] was conducted with the abstract articulatory synthesizer. A first parameter that could be varied (apart from the use or not of the learning rule), was the dimensional-

[1] let us mention that the mechanism described here does allow the cultural building of a syllable systems with which agents can imitate each other successfully. Results identical to what was described in (Oudeyer 2001) were found here. This is a significant progress compared to other other models of the origins of sound systems, but we will not give details here since it is not the main topic this paper

ity of the motor and perceptual spaces. Indeed, this is interesting since the goal of agents is to position syllable prototypes in these spaces such that they are not confused. Hence, the fact that increasing linearly the number of dimensions does increase exponentially the volume of the spaces might have some consequences. Figure 3 presents the average entropies of syllables systems composed of 40 syllables, and generated by populations of agents using or not the learning rule. For each case, 50 experiments were ran and the mean entropy and standard deviation were measures. We added a "uniform" column corresponding to the entropy of syllable systems target distributions generated randomly (uniformly distributed). The values of "uniform" thus characterize syllable systems which are absolutely not phonemically coded. We do observe that except for dimension 1 (which is not very realistic), the use of the learning rule does generate phonemically coded systems while when it is not used, systems have entropies equal to the uniform case : they are not phonemically coded. A second set of experiments consisted in keeping the dimensionality of motor and perceptual spaces equal to 2 and see whether the amount of noise would change anything (noise proved to be of importance in the experiments of de Boer 2000). Figure 4 shows that even a large amount of noise (30 percent) does not push the system to be phonemically coded when it does not use the learning rule, while using it still gives phonemically coded systems when noise gets high (yet it becomes less and less phonemically coded, which is normal since when the noise is to high, this is equivalent to reshuffling permanently and completely all targets).

A second series of similar experiments were made using the realistic articulatory synthesizer. Figure 5 shows the results, when noise is again varied : we see that even if a realistic synthesize is used, no phonemic coding is obtained if the learning rule is not used. And again, as soon as it is used, and if the noise remains reasonable, we do get phonemic coding.

As a consequence, neither morpho-perceptual innatism nor functionalism is sufficient to explain phonemic coding.

Conclusion

(Oudeyer 2002) presented a new hypothesis for the origins of phonemic coding, which had the a priori advantage of being more simple and requiring less assumptions than other models. In this paper we presented a general model of the origins of syllable systems, which in addition to solve the 2-level problems faced by previous research in the origins of sound systems, allows to test all hypothesis on a common ground. We showed that clearly the morpho-perceptual innatism and Lindblom's functionalism are neither necessary nor sufficient to explain phonemic coding. We did not prove that the non-functional side effects of our coupled neural maps model are necessary (but a better hypothesis has to be invented and validated to prove they are not), but they are sufficient, and so the only existing satisfying candidate to explain phonemic coding.

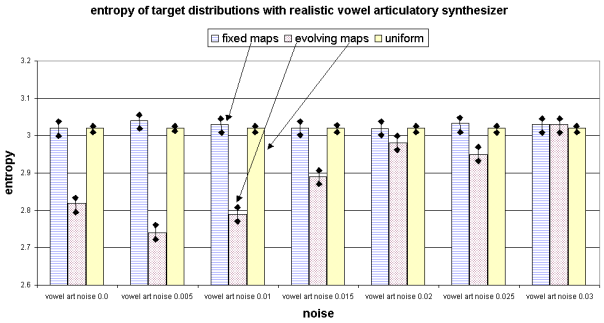


Figure 5:

References

de Boer, B. (2000) The origins of vowel systems, Oxford Linguistics, Oxford University Press.

Chomsky, N. and M. Halle (1968) The Sound Pattern of English. Harper Row, New york.

R. I. Damper (2000) Ontogenetic versus phylogenetic learning in the emergence of phonetic categories. 3rd International Workshop on the Evolution of Language, Paris, France. p.55-58.

Georgopoulos, Kettner, Schwartz (1988), Primate motor cortex and free arm movement to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. Journal of Neurosciences, 8, pp. 2928-2937.

Hurford, J., Studdert-Kennedy M., Knight C. (1998), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kirby, S. (1998), Syntax without natural selection: how compositionnality emerges from vocabulary in a population of learners, in Hurford, J., Studdert-Kennedy M., Knight C. (eds.), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Kuhl (2000) Language, mind and brain : experience alters perception, The New Cognitive Neurosciences, M. Gazzaniga (ed.), The MIT Press.

Ladefoged, P. and I. Maddison (1996) The Sounds of the World's Languages. Blackwell Publishers, Oxford.

Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development, in Ferguson, Menn, Stoel-Gammon (eds.) Phonological Development: Models, Research, Implications, York Press, Timonnium, MD, pp. 565-604.

MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.

Morasso P., Sanguinetti V., Frisone F., Perico L., (1998) Coordinate-free sensorimotor processing: computing with population codes, Neural Networks 11, 1417-1428.

Oudeyer, P-Y. (2001a) Coupled Neural Maps for the Origins of Vowel Systems. Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, Vienna, Austria, LNCS, springer verlag, Lectures Notes in Computer Science, 2001. Springer Verlag.

Oudeyer P-Y. (2001b) The Origins Of Syllable Systems : an Operational Model. in proceedings of the International Conference on Cognitive science, COGSCI'2001, Edinburgh, Scotland., 2001.

Oudeyer, P-Y. (2002) Phonemic coding might be the result of non-functional sensory-motor coupling dynamics, under review in SAB'02.

Stevens, K.N. (1972) The quantal nature of speech : evidence from articulatory-acoustic data, in David, Denes (eds.), Human Communication : a unified view, pp. 51-66, New-York:McGraw-Hill.

The Pragmatics of Number

Anna Papafragou (anna4@linc.cis.upenn.edu)
Institute for Research in Cognitive Science, 3401 Walnut Street
Philadelphia, PA 19104 USA

Julien Musolino (musolino@indiana.edu)
Department of Speech and Hearing Sciences, 200 S. Jordan Avenue
Bloomington, IN 47405 USA

Abstract

In terms of their semantic and pragmatic properties, number expressions (*one, two, three*…) have standardly been considered similar to quantifiers (*some, a few, all*). For instance, both kinds of expression form a scale: typically, an assertion containing a weaker member of the scale (*Some/Two of the dwarfs loved Snow White*) can be used to implicate that the stronger term of the scale doesn't apply (*Not all/No more than two of the dwarfs loved Snow White*). We report here results from two experiments with young speakers of Modern Greek which support the opposite conclusion: namely, that number terms and quantifiers behave differently in terms of the scalar inferences they support. We discuss implications of these findings for linguistic theories of the semantics/pragmatics of numerals, as well as for developmental theories of the acquisition of number words.

Introduction

In terms of their semantic and pragmatic properties, number expressions (*one, two, three*…) have standardly been considered as scalar expressions similar to quantifiers (*some, a few, all*). Semantically, both numerals and quantifiers have been assigned an 'at least' meaning (Horn, 1972; Grice, 1989): on this 'minimalist' analysis, *two* means *at least two* and *some* means *some (and possibly all).* Pragmatically, both numerals and quantifiers can be used to give rise to so-called *scalar implicatures.* Such implicatures arise when a speaker uses a weak member of the numerical or quantificational scale in order to implicate that the stronger term of the scale does not hold. For instance, an utterance such as (1) is typically used to implicate (2):

(1) Some/Two of the dwarfs loved Snow White.
(2) Not all/No more than two of the dwarfs loved Snow White.

The derivation of scalar implicatures is generally assumed to follow Gricean lines: for instance, if the speaker knew that the more informative statement with *all* (or a higher numeral) were true and relevant, other things being equal, s/he would have preferred to use it. The fact that s/he didn't offers grounds for assuming that such a more informative statement isn't true.

More recently, several objections have been raised to the view that the scalar semantic/pragmatic profile of numerals is identical to that of quantifiers (Carston, 1985; 1998; Horn, 1992). First, it has been observed that cardinals, but not 'inexact' quantifiers, can feature in contextually induced reversals of scale: in (3), *three* is used to communicate *at most three*:

(3) These houses are big enough for families with three kids.

But it is not possible to use *some* in a similar way (e.g. to implicate *at most some*). Second, number terms are regularly used with an 'exact' interpretation in mathematical statements (*Two plus three makes five*), a fact which is hard to reconcile with an 'at least' semantics for numerals (unless we assume that cardinals are ambiguous). Third, the scalar properties of numerals disappear under incorporation: a four-sided figure has exactly (not at least) four sides. Finally, approximation seems to work differently with numerals: *I have $300* is more likely to receive an 'at least' interpretation than its unrounded counterpart *I have $300.17*. For these and related reasons, it has been proposed that cardinals are, in fact, distinct from other scalar expressions. According to these proposals, numerals do not have an 'at least' semantics upper-bounded by a scalar implicature; rather, they are best analyzed as underspecified among an 'at least', 'exact' and 'at most' reading. Pragmatic considerations then are used to determine which reading is more appropriate in a specific context.

There is by now a vast linguistic literature which attempts to adjudicate between the 'minimalist' proposal and alternative theories for number terms (for reviews, see Carston, 1998; Levinson, 2000). The outcome of this debate is important, since theories of scalar predication are a valuable source of insights about how semantic information and contextual cues co-ordinate with each other and contribute to utterance meaning.

In this paper, we present the results from two developmental studies which compare the semantic-pragmatic properties of both cardinals and quantifiers. Our goal is to use the scalar behavior of numerically modified and quantified phrases in child language to shed light on the theoretical debate surrounding these predicates. To preview our discussion, we find that number terms and quantifiers behave differently in child language in terms of the scalar inferences they support. We take these results to be incompatible with 'minimalist' semantic accounts of numerals. Thus we show that developmental data offer an additional piece of evidence for the different status of numerals and quantifiers within semantic/pragmatic theories.

Our experiments build on previous cross-linguistic studies of the acquisition of scalar predicates (for English, see Chierchia, Crain, Guasti, Gualmini & Meroni, 2001; Gualmini, Crain, Meroni, Chierchia & Guasti, 2001; Musolino & Lidz, in press; for French, see Noveck, 2001). Even though they were not concerned with the pragmatics of number terms, these studies have shown that preschoolers have difficulty with the pragmatics of other scalar expressions such as quantifiers (even though they seem to know the semantics of such quantifiers). In our experiments, we turn to Modern Greek for further evidence. Since the scalar inferences associated with numbers and quantifiers apply universally, we should expect to see cross-linguistic similarities in the acquisition of the pragmatics of scalar predicates.

Experiment 1

Methods

Participants. Participants were a group of 20 Greek-speaking 5-year-olds between the ages of 4;11 and 5;11 (mean 5;3) and a group of 20 adult native speakers of Greek. The children who participated in this study were recruited from daycares in the Athens area. The adults were all undergraduate students at the University of Athens.

Procedure and Materials. In this experiment, we asked children (and adults) to offer pragmatic judgements on sentences containing either the numerical scale <*three, two*> or the quantifier scale <*all, some*>.[1] We used a slightly modified version of the Truth Value Judgment Task (Crain & Thornton, 1998). The TVJT typically involves two experimenters. The first experimenter acts out short stories in front of the subjects using small toys and props. The second experimenter plays the role of a puppet (in this case Minnie) who watches the stories alongside the subjects. At the end of the story, the puppet is asked to say what happened in the story. In our version, instead of asking subjects if the puppet is 'right' or 'wrong' (as in the original TVJT), we then asked whether the puppet 'answered well' (i.e., *Apantise kala;*, 'Did-(she)-answer well?'). This modification was made since we were interested in felicity, not truth. Finally, the subjects were asked to justify their answers by explaining why they thought that Minnie answered well or not.

The children were tested individually in a quiet room away from the class. Adult subjects were shown a videotaped version of the stories witnessed by the children, including the warm-up stories. They were given a score sheet and were instructed to indicate, for each story, whether Minnie had 'answered well' or not. They were also asked to provide a brief justification for their answers.

For each scale, subjects were asked to judge four statements like the ones in (4-5):

(4) Meriki apo tus dinosavrus efagan dedra.
'some of the dinosaurs ate trees'
(5) Dio apo tus dinosavrus efagan dedra.
'two of the dinosaurs ate trees'

In each case, these utterances were used to describe situations which satisfied the truth conditions of utterances containing stronger terms on the respective scales, i.e., *all, three.* The critical stories were identical for both scales. For instance, for both (4) and (5), a group of three dinosaurs went to get something to eat. After contemplating other options, all three dinosaurs ended up eating trees. In this context, assuming an 'at least' semantics for the scalar predicates, both utterances in (4) and (5) express a true but pragmatically infelicitous proposition.

Before the main phase of the experiment, each child received two 'warm-up' stories, one designed to elicit a 'Yes' answer and the other a 'No' answer. Furthermore, in addition to the critical statements, and for each scale, subjects were asked to judge four control statements like the ones in (6-7):

(6) Donald cleaned some of the cars/airplanes.
(7) Donald cleaned two of the cars/airplanes.

The purpose of these controls was to ensure that subjects, and in particular children, could accept or reject the puppet's statements when appropriate and, more importantly, that they could do so when these statements involved felicitous uses of terms like *some* and *two*. For each control statement, the experimenter had a choice between two versions: one that was a correct description of the story and would therefore elicit a 'Yes' response and one that was an incorrect

[1] For ease of exposition, we provide English glosses throughout. The Greek terms are <*tris, dio*> and <*oli, meriki*> respectively.

description of the story and would therefore elicit a 'No' response. The experimenter selected the version of the control statement (correct or incorrect description) based on the child's response of the preceding critical statement. If the child had rejected the puppet's statement on the previous critical trial, the experimenter selected the version of the control statement that would elicit a 'Yes' response, and vice-versa. This step was taken to keep a balance between 'Yes' and 'No' responses.

Subjects (5-year-olds and adults) were randomly assigned to one of two conditions, determined by scale type (i.e., <*all, some*>, <*three, two*>) which gave rise to a 2X2 design with age and scale type as between subject factors and 10 subjects per cell. The age range and mean ages for the 10 children assigned to each scale condition, i.e. <*all, some*> and <*three, two*>, are 5;0 to 5;11 (mean 5;4) and 4;11 to 5;10 (mean, 5;3) respectively. In each condition, subjects received four critical trials and four control trials administered in a pseudo-random order. Within each condition, order of presentation was counterbalanced between subjects.

Results

Beginning with test trials, we found that adult subjects overwhelmingly rejected the puppet's statements in each of the two conditions, i.e. 92.5% of the time in the <*all, some*> and 100% of the time in the <*three, two*> condition. Statistical analysis revealed no reliable difference between these rejection rates ($t(18) = 1.96$, $p = 0.06$). By contrast, we found that while 5-year-olds rejected the puppet's statements in the case of <*three, two*> 65% of the time, they did so reliably less often in the case of <*all, some*> i.e., 12.5% of the time ($t(18) = 3.47$, $p < 0.01$) .The proportions of 'No' responses were entered into an analysis of variance (ANOVA) with two factors: age (5-year-olds vs. adults) and scale type (<*all, some*> vs. <*three, two*>). The analysis revealed a main effect of age ($F(1,36) = 54.41$, $p < 0.0001$), a main effect of scale type ($F(1,36) = 14.81$, $p < 0.001$) and a reliable interaction between age and scale type ($F(1,36) = 8.33$, $p < 0.01$).

Recall that subjects in this study were also asked to provide justifications for their answers. Adults in 98% of the justifications they offered for rejecting a statement made reference to the stronger term of the scale, as expected. That is, they said that the puppet was wrong that some or two of the dinosaurs ate a tree because ALL or THREE of them had eaten a tree. Children's justifications for rejecting a numerically modified statement always invoked the pragmatically more appropriate stronger numeral. The same is true for the few cases in which a quantified statement with *some* was rejected by children.

On control trials, adults gave correct responses 100% of the time in the <*all, some*> condition and 80% of the time in the <*three, two*> condition. No reliable difference was found among these means ($t(18) = 1.92$, $p = 0.07$). On the same items, children gave correct responses 90% of the time for <*all, some*> and 95% of the time for <*three, two*>. Again, no reliable differences among the means were found ($t(18) = .77$, $p = 0.44$).

Discussion

Two main conclusions emerge from the first experiment. First, children are much less likely to compute scalar implicatures than adults. This finding comports well with previous research showing that preschoolers cross-linguistically have difficulties understanding scalar inferences, especially those associated with quantifiers (Chierchia et al., 2001; Gualmini et al. 2001; Musolino & Lidz, in press; Noveck, 2001). Second, and more crucially for present purposes, 5-year-old children are more successful in drawing scalar inferences triggered by numerals than by quantifiers. This finding is even more remarkable given that our critical trials with *some* and *two* used identical scenarios and props. This result is unexpected given standard 'minimalist' semantic accounts, since it points to a difference in status between numerals and inexact quantifiers such as *some*.

On the basis of the available evidence on children's performance with scalar inferences, previous literature has concluded that children are generally incapable of deriving scalar implicatures on-line (Chierchia et al., 2001). It might be tempting to interpret our results (at least for *some*) in a similar way. However, there are alternative hypotheses which are worth pursuing. For instance, it is possible that children's failures are not due to an inability to derive the implicatures but to a misunderstanding of the nature of the task. Perhaps children (unlike sophisticated adult communicators) treat this as a truth-value judgement task. Since no special motivation is provided for drawing the scalar inferences, children may be more willing to let the puppet score an appropriate response if she has simply given a true (albeit infelicitous) description of what happened. This inability to correctly assess the experimental demands may therefore make preschoolers in our study appear less pragmatically savvy than they really are (for similar explanations of children's 'failures', see Shipley, 1979; Gelman & Greeno, 1989). Our aim in designing Experiment 2 was to investigate whether a methodologically improved version of the same task might raise children's overall performance with scalar terms. We were also interested in testing whether such a task would yield a similar asymmetry between cardinals and quantifiers.

Experiment 2

Methods

Participants. 20 Greek-speaking children ranging in age between 5;1 and 6;3 (mean 5;6) participated in this experiment. These children were recruited at daycare centers in the same Athens area as the children used in Experiment 1.

Materials and Procedure. Experiment 2 introduces several modifications to the design of our previous study. First, we included a training phase, in which we presented children with four warm-up stories designed to enhance their awareness of the fact that they were being asked to produce pragmatic judgments. Children were told that the puppet, Minnie, sometimes said 'silly things' and that the purpose of the game was to help Minnie 'say things better'. For example, Minnie would be shown a toy dog which she would describe using the truth-conditionally accurate - but pragmatically infelicitous - statement 'This is a little animal with four legs'. The child would then be asked whether 'Minnie answered well' and whether 'we can say it better'. In case the child failed to correct the puppet and provide a better description, the experimenter eventually corrected Minnie and provided the appropriate description, i.e. 'Minnie didn't say that very well. This is a DOG'.

The second change we introduced concerns the test scenarios. As before, subjects witnessed four test stories in which they were asked to judge statements containing the scalar terms *some* and *two* in situations which satisfied the truth conditions of the stronger terms of the respective scales, i.e. *all* and *three*. However, the stories in Experiment 2 were all based on scenarios in which the main character was involved in a contest or a challenge. The main character's performance therefore became the focal point of the stories and at the end, the puppet was asked to comment on how well the character in question had done, 'How did X do?' (*Pos ta pige o X?).* In one of the stories for example, one of the characters claims that he is very good at throwing hoops around a pole and he challenges Mickey to try and do the same with three hoops. Mickey really concentrates hard and he's able to put all the hoops around the pole. At the end of the story, Minnie is asked "How did Mickey do?" and she answers by saying that "Mickey put some of the hoops around the pole". The idea behind this manipulation is to make clear the demands of the communicative situation: given that Mickey's performance is being directly evaluated, only an answer making reference to *all* the hoops would satisfy the expectations of the hearer.

Children heard four test stories and four control stories administered in a pseudo-random order. As before, order of presentation was counterbalanced between subjects within a single condition. Finally, as in Experiment 1, subjects were randomly assigned to either of the two scale conditions. The age range for the 10 children assigned to the quantifier scale condition was 5;1 to 6;2 (mean 5;6). For the number scale condition, the range was 5;4 to 6;3 (mean 5;7).

Results

As before our dependent measure was children's *Yes/No* responses to the puppet's statements. We found that the manipulations described above led children to reject the puppet's statements much more often than in Experiment 1. Nevertheless, the difference between scales persisted: children answered correctly 52.5% of the time for the <*all, some*> scale, and 90% of the time for the <*three, two*> scale ($t(18) = 2.39$, $p = 0.02$). We compared the rejection rates from Experiment 1 and Experiment 2 by entering them into a 2 (training vs. no training) by 2 (scale type*)* ANOVA. The analysis revealed a main effect of training ($F(1,36) = 8.92$, $p < 0.01$), a main effect of scale type ($F(1,36) = 17.1$, $p < 0.001$) and no reliable interaction between training and scale type ($F(1,36) = 0.47$, $p = 0.49$). On control items, children gave correct responses 85% of the time in the quantifier condition and 95% of the time in the number condition. No reliable differences between these means were found ($t(18) = 0.89$, $p = 0.38$). Finally, the justifications children offered for their rejections in the overwhelming majority of cases (93%) made explicit reference to the stronger term of the scale (just like adults' justifications in Experiment 1).

Discussion

The results from this Experiment show that children's sensitivity to scalar inferences improves dramatically if children are provided with clear contextual cues about the communicative expectations of the task.[2] This is an important and novel finding in itself (see Papafragou & Musolino, 2001, for discussion). For present purposes, a more pertinent finding is that the asymmetry found in the previous experiment persists: in child language,

[2] Children in Experiment 2 are slightly older than those in Experiment 1. However, there are strong reasons to think that the differences in performance are not due to these small age differences. First, in the <*all, some*> case, the age difference is not reliable (mean 5;4 vs. 5;6, $t(18) = 0.985$, $p = .33$) but the difference in performance persists. Second, and more importantly, previous studies have reported children's difficulties with scalar implicatures well beyond the age of 5;0 (and up to the age of 10;0).

numerically modified phrases give rise to scalar inferences much more readily than quantified phrases.

General Discussion

As we mentioned in the introduction, there are several theoretical reasons for considering cardinals as distinct from quantifiers and other scalar terms, and our experimental data seem to confirm this difference. Our studies demonstrate that, in child language, inexact quantifiers such as *some* are assigned a lower-bounded reading ('at least some', or 'some and possibly all') and the associated scalar inferences are ignored in the absence of strong contextual cues. By contrast, preschoolers typically reject lower-bounded interpretations of numerals and are very attentive to the scalar properties of number terms. Since the test environment was identical in both situations, this difference in interpretive preferences points to a difference in the semantic/pragmatic status of numbers and quantifiers. Specifically, it suggests that, while a minimalist semantics may be plausible for quantifiers and other scalar predicates, cardinals may best be analyzed in terms of either an 'exact' or an underspecified semantics.

An interesting observation which arises from our experiments is that children often used the counting routine as a means of formulating their responses. For instance, when Minnie offered *Two of the dinosaurs ate trees*, several children protested by saying *No, one, two, THREE dinosaurs ate trees* (while at the same time pointing to and counting the dinosaurs one by one). Explicit counting of this sort offers a specific and precise way of verifying statements containing number terms (by placing the referents of the corresponding NPs in a one to one correspondence with objects in the world). Counting games may also encourage an 'exact' interpretation of the numerals. Notice that neither of these steps is available for the inexact *some*.

There is additional evidence that the observed asymmetry between quantifiers such as *some* and numerals is related to the difference between discrete and non-discrete (vague) scalar predicates. Papafragou (2002) tested Greek preschoolers' understanding of the scalar modifier *half* (e.g. *The bear built half of the tower*). This modifier resembles numerals in that it is discrete (it denotes a precise partitioning of a quantity into two equal parts). By means of the same methodology as the second experiment reported above, it was found that young Greek learners rejected lower-bounded interpretations of the modifier *half* in contexts which licensed scalar implicatures. For instance, children overwhelmingly rejected the statement *The bear built half of the tower* in cases where a whole tower had been built. This pattern reinforces the conclusion that discrete quantity modifiers (e.g. *half*, numerals) have distinct properties from inexact quantifiers and other scalars (e.g. *some, a few*).

Although our results can be taken as evidence against a minimalist semantics for numerals, they leave open the question of whether number terms in natural language have an exact or an underspecified semantics. It is worth pointing out that, in the considerable developmental literature which looks at children's acquisition of number terms (Carey, 2001; Gelman & Gallistel, 1978; Wynn, 1992; Bloom & Wynn, 1997, among many others), it is usually assumed that children ultimately arrive at an 'exact' semantics for number terms (which is the correct adult meaning). Moreover, according to one influential position, children assign meaning to cardinal expressions in natural language by placing them in a one-to-one correspondence with an innate conceptual 'integer list' (Gelman & Gallistel, 1978). The results reported in this paper, even though not univocally in favor of an 'exact' over an underspecified semantics for numerals, are certainly consistent with these positions.

To conclude: Throughout this paper, we have assumed that aspects of child language can be instructive about the nature of the semantic representations in adults. This position accepts some fundamental continuity in the representational systems of children and adults - here, in the specific domain of number. It thus allows us to bring acquisition data to bear on theoretical debates about the architecture of the semantic-pragmatic system in adults. Even though several interesting questions remain unresolved about both the adult system and its acquisition by young learners, we hope to have shown that an approach which treats numbers as regular scalar predicates alongside quantifiers misses important generalizations about their developmental properties.

Acknowledgments

We wish to thank Lila Gleitman, Henry Gleitman, John Trueswell, and the members of the CHEESE seminar at the University of Pennsylvania for comments and input. Thanks go to the teachers and children at the 3rd and 5th daycares at Vrilissia (Athens), and to Profs. D. Chila-Markopoulou, D. Theofanopoulou-Kontou and S. Hoidas for the experiments conducted at the University of Athens. This research was supported by NSF-STC Grant #SBR-89-20230, and by NIH Grant #1-R01-HD37507-01L.

References

Bloom, P. & Wynn, K. (1997). Linguistic cues in the acquisition of number words. *Journal of Child Language, 24*, 511-533.

Carey, S. (2001). Cognitive foundations of arithmetic: Evolution and ontogenesis. *Mind and Language, 16,* 37-55.

Carston, R. (1990). Quantity maxims and generalized implicature. *UCL Working Papers in Linguistics* 2: 1-31. Reprinted in *Lingua, 96* (1995), 213-244.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications*. Amsterdam: Benjamins.

Chierchia, G., Crain, S., Guasti, M., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. *Proceedings from the 25th Annual BUCLD* (pp. 157-168). Somerville, MA: Cascadilla Press.

Gelman, R. & Gallistel, R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gelman, R. & Greeno, J. (1989). On the nature of competence: Principles for understanding a domain. In L. Resnick (Ed.), *Knowing and learning: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.

Grice, P. (1989). *Studies in the ways of words*. Harvard: Harvard University Press.

Gualmini, A., Crain, S., Meroni, L., Chierchia, G., & Guasti, M. (2001). At the semantics/pragmatics interface in child language. *Proceedings of SALT* 11. Ithaca, NY: Cornell University.

Horn, L. (1972). *On the semantic properties of logical operators in English.* Doctoral dissertation, Department of Linguistics, UCLA. Distributed by IULC.

Horn, L. (1992). The said and the unsaid. *Proceedings of SALT* 2 (pp. 163-191). Department of Linguistics, Ohio State University.

Levinson, S. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.

Musolino, J. & Lidz, J. (in press). Preschool logic: Truth and felicity in the acquisition of quantification. *Proceedings of BUCLD* 26. Somerville, MA: Cascadilla Press.

Noveck, I. (2001). When children are more logical than adults. *Cognition, 78,* 165-188.

Papafragou, A. (2002). Scalar implicatures in language acquisition: Some evidence from Modern Greek. To be presented at the 38th Annual Meeting of the Chicago Linguistics Society, University of Chicago, 25-27 April 2002.

Papafragou, A. & Musolino, J. (2001). *Scalar implicatures: Experiments at the Semantics-Pragmatics Interface*. IRCS Technical Report 01-14. Philadelphia, PA: University of Pennsylvania, Institute for Research in Cognitive Science.

Shipley, E. (1979). The class inclusion task: Question form and distributive comparison. *Journal of Psycholinguistic Research, 8,* 301-331.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology, 24,* 220-251.

A Computational Theory of Complex Problem Solving Using Latent Semantic Analysis

José Quesada, Walter Kintsch ([quesadaj, wkintsch]@psych.colorado.edu)
Institute of Cognitive Science, University of Colorado, Boulder
Boulder, CO 80309-0344 USA

Emilio Gomez (egomez@ugr.es)
Department of Experimental Psychology University of Granada
Campus cartuja, S/N, Granada, Spain

Abstract

Complex Problem Solving (CPS) is a hybrid between field studies and experimental studies. This paper introduces a new, abstract conceptualization of *microworlds* research based on two innovations: (1) a problem representation, which treats protocols as objects in a feature space and, (2) a similarity metric which is defined in this problem space. Latent Semantic Analysis (LSA) is used to analyze performance in CPS, using actions or states as units instead of words and trials instead of text passages. Basic examples of applications are provided, and advantages and limitations are discussed.

Introduction

Many real-word decision making and problem solving situations are (1) *dynamic*, because early actions determine the environment in which subsequent decision must be made, and features of the task environment may change independently of the solver's actions; (2) *time-dependent*, because decisions must be made at the correct moment in relation to environmental demands; and (3) *complex*, in the sense that most variables are not related to each other in one-to-one manner. In these situations, the problem requires not a single decision, but a long series of decisions which are dependent on one another. For a task that is changing continuously, the same action can be successful at moment *t1* and useless at moment *t2*. However, traditional, experimental problem solving research has focused largely on tasks such as anagrams, concept identification, puzzles, etc. that are not representative of the features described above.

In Europe, researchers led by Broadbent (e.g., Broadbent, 1977) in the UK and Dörner (e.g., Dörner, 1975) in Germany, were concerned about that fact and started working on a set of computer-based, experimental tasks that are dynamic, time-dependent, and complex, called *Microworlds*[1]. The study of microworlds is an example of Complex Problem Solving (e.g., French & Funke, 1995).

Compared to traditional Problem Solving, Complex Problem Solving (CPS) radically changed the kind of phenomena reported, the kind of explanations looked for, and even the kind of data that were generated. However, the results obtained to date are far from being integrated and consolidated. This fact led Funke to affirm that 'Despite 10 years of research in the area, there is neither a clearly formulated specific theory nor is there an agreement on how to proceed with respect to the research philosophy. Even worse, no stable phenomena have been observed' (Funke, 1992, p. 25). Almost another 10 years after Funke's argument, although more empirical research has been conducted in the area, we cannot say that the situation has changed drastically. At this moment, there is no theory able to explain even part of the specific effects that have been described or how they can be generalized.

A theory of generalization and similarity is as necessary to psychology as Newton's laws are to physics (Shepard, 1987). However, for CPS there is no common, explicit theory to explain why a complex, dynamic situation is similar to any other situation or how two slices of performance taken from a problem solving task can possibly be compared quantitatively (Klein, Orasanu, Calderwood, & Zsambok, 1993). This lack of formalized, analytical models is slowing down the development of theory in the field. At least two problems make it difficult to apply the classical problem solving approach to CPS, one theoretical and one methodological:
(1) The utility of state space representation for tasks with inner dynamics is reduced because in most CPS environments it is not possible to undo the actions. For example, imagine that two participants in *Firechief* (see below) are in an identical situation (system state) when the trial starts. One of them proceeds to make a control fire on the east side of a fire, while the other one is preparing a control fire on the north front of the fire. After these actions, the system state is no longer identical for them. Now they have to cope with rather different problems. Moreover, if the first participant wants to apply the same technique that

[1] This term sometimes has other meanings. For example, educational applications created to teach physics (Henderson, Klemes, & Eshet, 2000), simulated words in the early AI programs like the block word of SHRDLU, (Winograd, 1972) and static tasks to study decision making (Green, 2001) have been called *Microworlds*. However, we are concerned here only with tasks that fulfill the conditions described above.

the second participant used, there is absolutely no way to come back to the initial state and begin with a new strategy. This situation is not an issue in static tasks like the tower of Hanoi problem because it is always possible to undo a wrong action. Feedback delays (e.g., Brehmer, 1995; Gibson, 2000) and an upsettingly large number of possible states (e.g., Dörner, 1975; Omodei and Wearing, 1995) contribute to the reduced utility of the state space approach.
(2) Traditional methods of knowledge elicitation are not always applicable: Concurrent verbal protocols consistently interfere with performance (Dickson, Omodei & Wearing 2000); measures based on relatedness judgments like rating correlations or pathfinder distance correlations are not sensitive to context manipulations in naturalistic task like fire fighting (Calderwood, 1989).

In this paper we introduce a theory and methodology for CPS tasks based on Latent Semantic Analysis (LSA, Landauer and Dumais, 1997). The theory addresses issues concerning the induction, representation, and application of knowledge. Basically, LSA infers knowledge from the many weak constraints that are present in complex problem solving situations.

LSA does not represent all the possibilities of a system (the system's state space), but only the paths that people have actually followed when interacting with it. This offers a realistic view of how the system is understood and used by humans. LSA is a *computational* theory on how environmental constraints are learned and how they can be described. In terms of Simon's classical parable of *'the ant and the beach'* (Simon, 1981, p. 63), one could say that LSA describes and infers the shape of the beach from the thousands of tracks the ants have left on the beach. In this sense, LSA can be conceived as a computational extension to theories for describing environmental constraints, such as the abstraction hierarchy (Rasmussen, 1985; Vicente, 1999).

LSA has several interesting features that make it a suitable technique for analyzing performance on a complex, dynamic task:
(1) It does not assume independence of decisions; indeed, it uses dependencies between decisions to infer structure. Some methods employed in the past treated CPS performance in a way that assumed that decisions are independent or have short-term dependencies only.
(2) LSA reduces the dimensionality of the space. Imagine a hypothetical problem solving task that, when performed from the beginning to the end, traverses 300 states. Furthermore, let us assume that every state is described by 6 dichotomous variables (2^6 = 64 possible states). Since we have 300 states in our sample of performance, there are 64 ^ 300 = 7e541 possible paths in this task. Every sample would be represented as a matrix of 6 x 300 = 1800 values. With LSA, every sample is represented as a vector of only 100-300 values.
(3) There are no *a-priori* assumptions about ' *the beach*'. In most of the analysis performed on *microworld* data the experimenter has to impose some structure (*a-priori*, theoretically driven assumptions) on the data. However, the selection of this theoretical structure (How many strategies are possible? How many strategies are representative? Are they generalizable to different conditions?) can bias the analysis. The LSA approach is self-organizing, and does not require defining an *a priori* theoretical structure, as will be shown below.

Before we start describing what LSA is and how it can be applied to CPS, we would like to stress some abstract considerations that underlie the approach that we are about to implement. These considerations are independent of the procedure itself (other procedures could be defined using this framework), but, in our opinion, an essential step to dealing with the complexity of the tasks at hand: (1) each microworld can be conceptualized as a complex, multidimensional feature space. (2) To address the intractability problem, we usually need to create a representation or transformation of this original multidimensional feature space. To do this, we need to find a set of features that represent the characteristics that make participants different, and to delete those that are not important. (3) Last, each trial of every subject can be conceptualized as an implementation of several values in the feature space. Not only a trial, but every subpart or superpart of a participant's performance (strategies or performance patterns) can be thought of as an object in this space.

We shall illustrate how LSA can be used to analyze CPS tasks, using the *Firechief* microworld as an example.

Description of the example application task

Firechief (Omodei & Wearing, 1995) simulates a forest where a fire is spreading. Participants' task is to extinguish the fire as soon as possible. In order to do so, they can use helicopters and trucks (each one with particular characteristics) that can be controlled by mouse movements and key presses. There are three commands that are used to control the movement and functions of the appliances: (1) Drop water on the current landscape segment; (2) Start a control fire (trucks only) on the current landscape segment; (3) Move an appliance to a specified landscape segment.

Every time a participant performs an action, it is saved in a log file as a row containing action number, command (e.g. drop water or move) or event[2] (e.g., a wind change or a new fire), current performance score, appliance number, appliance type, position, and landscape type. Most of these variables are not continuous, but on a nominal scale, such as type of movement. For more information on the structure of the log files, see Omodei and Wearing (1995).

[2] Events are generated by the system, while actions are generated by the user. Events are also lines in the log file. Only 1-2% of the lines in a log file are events.

The set of trials that was used in this report (referred as *corpus*) was obtained in four experiments described in Quesada, Cañas, & Antoli (2000) and Cañas, Quesada, Antoli & Fajardo (submitted).

Description of LSA

LSA is a machine-learning model that induces representations of the meaning of words by analyzing the relation between words and passages in large bodies of text. LSA is both a method (tool) used to develop technology to improve educational applications, and a theory of knowledge representation used to model well known experimental effects in text comprehension and priming, among others (Landauer & Dumais, 1997). Latent Semantic Analysis was originally developed in the context of information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1991) as a way of overcoming problems with polysemy and synonymy. Some words appear in the same contexts (synonyms) and an important part of word usage patterns is blurred by accidental and inessential information. The method used by LSA to capture the essential semantic information is dimension reduction, selecting the most important dimensions from a co-occurrence matrix decomposed using Singular Value Decomposition (see below). As a result, LSA offers a way of assessing semantic similarity between any two samples of text in an automatic, unsupervised way.

LSA has been used in applied settings with a high degree of success in areas like automatic essay grading (Foltz, Laham, & Landauer, 1999) and automatic tutoring to improve summarization skills in children (E. Kintsch, Steinhart, Stahl, Matthews, Lamb, & the LSA Research Group, 2000). As a model, LSA's most impressive achievements have been in human language acquisition simulations (Landauer & Dumais, 1997) and in modeling of high-level comprehension phenomena like metaphor understanding, causal inferences and judgments of similarity (Kintsch, 2001).

Although LSA has been mostly used on text corpora, our basic point is that LSA can be applied to any domain of knowledge where there are a high number of weak relations between tokens, as in CPS log files. Instead of word usage statistics obtained from huge samples of text, we have used a representative amount of activity in controlling dynamic systems, and actions or states have been used to develop an objective measure of similarity in the changing, time-dependent, highly complex experimental tasks known as *microworlds*. The next sections show the basic steps involved in this analysis and presents some examples of the powerful results that can be obtained thereby.

LSA applied to Microworlds

LSA starts with the creation of a matrix of actions by trials. Note that this is not an exhaustive state space, or a mapping of all possible transitions between actions, since in most of the systems – other than small ones like Hayes & Broadbent's (1988) sugar factory and the like - this task would be excessively demanding (see Buchner, Funke & Berry, 1995). Our corpus was composed of 360,199 actions in 3441 trials. Among them, only 75,565 were different actions, which means that on average each action appears 6.25 times in the corpus. Note that we are representing *only* the information that actual people interacting with the system experienced, not all possible actions in this microworld.

Each of these 75,565 rows stands for a unique action, and each of the 3441 columns stand for a trial. Each cell contains the frequency with which the action of its row appears in the trial denoted by its column. Note that most of the cells will contain a frequency of zero, since most actions appear in only a few trials and not in the rest.

This matrix of frequencies is decomposed using Singular Value Decomposition (SVD). Any matrix can be decomposed and then recomposed perfectly using only as many factors as the smallest dimension of the original matrix. However, an interesting phenomenon occurs when the original matrix is recomposed using fewer dimensions than necessary: the reconstructed matrix is a least-squares best fit. When the actions-by-trials matrix is recomposed using a small fraction of the available dimensions (usually between 100 and 300 dimensions), the new matrix contains information that has been inferred from the dependencies between actions and the context where these actions appeared. In fact, the contexts where these actions did not appear are as important - carry as much information - as those where they did. The microworld is a new multidimensional feature space, where both actions and context (trials) are represented in a way that amplifies those characteristics that make participants different, and delete those that are not important for classifying their performance.

Some examples of possible analysis

LSA allows us to measure the functional similarity between actions in CPS tasks. Some actions can be considered as *functional synonyms*: they appear in the same contexts, and fulfill approximately the same function. The following example illustrates this idea.

Table 1: Example of how LSA captures similarity at a molecular (action) level

	Time *t1*	**Time *t2***
Example 1	move_11_9_forest	Drop_11_9_forest
Example 2	move_15_15_forest	Drop_15_15_forest
Example 3	move_10_9_forest	Drop_10_9_forest
Example 4	move_11_9_forest	control_11_9_forest

In Table 1, four different actions some actions are shown. For simplicity, some variables that are normally contained in the log files have been removed. Example1 contains a movement to the point (11, 9) in the screen, which is of type *forest*, and then, a drop water action there. Example 3 shows a very similar picture, where the movement is done to a contiguous cell (10,9) that is also of type *forest*. From a human point of view, these two examples are highly similar. For LSA they are too, as can be seen in their similarity expressed as a cosine of 0.854 in Table 2.

Table 2: Similarities between Table 1 examples (cosines).

	Example 1	**Example 2**	**Example 3**	**Example 4**
Example 1		0.124	0.854	0.662
Example 2			0.1259	0.077
Example 3				0.566
Example 4				

The second example has a rather different meaning since the cell targeted is (15,15), quite far from the cell used in examples 1 and 3. The cosines between them and example2 (.124 and .125) are, accordingly, smaller than the one between 1 and 3.

Example 4 describes an action that has been performed in the same cell as in example 1 (11,9), but this time is a control fire instead of a drop-water action. The cosine between 1 and 4 is high (0.56), expressing a certain similarity between the two actions, but not as high as in examples 1 and 3, where the objective similarity is more evident.

Tables 3 and 4 present a more complex example where wider slices of performance (8 actions) are compared. The samples labeled Example1, Example2, and Example 3 are beginnings of trials that have been selected randomly from the *corpus*. This time, all the usable information contained in the log file is displayed. Each action has six components: type of action, appliance number, appliance type, departure cell, arrival cell and type of arrival cell.

One difficulty arises. When LSA is used on text, cosines are easily understood since every reader has an intuitive experience of meaning (e.g., the sentences 'The man was driving a yellow car' and 'The man was traveling in a red car' have a cosine of .89, and our common sense tells us that these sentences convey similar information). When LSA is used on samples of performance from a *microworld*, there is no way the reader can understand the meaning of the log files without watching a replay or having an extraordinarily vivid imagination plus experience with the task. For most people, the following extracts in Table 3 are hardly understandable. For researchers familiar with *Firechief*, they should be as clear as a piece of sheet music to a musician. However, understanding the contents of these examples is not *conditio sine qua non* for understanding the advantage of LSA analysis over two other methods, namely exact matching and correlation between transition matrices. Suffice it to say that Examples 1 and 2 are very similar and Example 3 is very different from them. The attentive observer could induce this from the locations (coordinates in the Firechief map), the type of actions, and type of landscape cell. An *exact matching* method would count the number of times that the same action occurs in two examples. Then, the number of matches divided by the total number of actions in the example provides a measure of the similarity between two samples. This method would render a similarity of 1/8 between example 1 and 2, and zero in comparisons 1 vs. 3 and 2 vs. 3. This method is equivalent to keyword counting in text, which is known to be incapable of capturing similarities in meaning, because of the polysemy and synonymy effects discussed above.

A somewhat more flexible method is the use of *transitions between actions*, as proposed by Howie and Vicente (1998) and used in Quesada et al. (2000) and Cañas et al. (submitted). It is based on counting the number of times that one type of action precedes any other type. The frequencies of every transition are registered in cells in a table, and then the resulting tables for two examples are correlated. The method cannot account for all the variability in actions, because of the huge amount of zero entries that artificially

Table 3**:** First 8 movements in 3 slices randomly sampled from the *Firechief* experiments described in Quesada et al. (2000) and Cañas et al. (submitted). When an action is shared by two extracts, it is marked as a shaded cell.

Example 1	**Example 2**	**Example 3**
move_2_truck_4_11_13_3_forest	move_2_truck_4_11_12_15_forest	move_2_truck_4_11_2_2_pasture
move_1_truck_4_14_16_14_forest	move_1_truck_4_14_13_5_forest	move_1_truck_4_14_0_5_forest
move_3_copter_8_6_11_12_forest	move_4_copter_11_4_11_9_forest	move_4_copter_8_6_8_4_clearing
move_4_copter_11_4_11_9_forest	drop_water_4_copter_11_9_forest	move_3_copter_8_6_8_10_clearing
Control_fire_2_truck_13_3_forest	move_4_copter_11_9_13_8_forest	control_fire_2_truck_2_2_pasture
Control_fire_1_truck_16_14_forest	control_fire_2_truck_12_15_forest	control_fire_1_truck_0_5_forest
move_2_truck_13_3_17_7_clearing	move_2_truck_12_15_13_14_forest	move_4_copter_8_4_4_2_forest
move_1_truck_16_14_20_12_forest	control_fire_2_truck_13_14_forest	move_3_copter_8_10_2_3_clearing

increase the correlation, so only action type was considered. This analysis is shown in tables 4(*a*,*b*,*c*). Since lots of information contained in the log files has been dropped, the method does not distinguish between these examples. The correlation between table *a* and *b* is 0.971; exactly the same correlation is obtained for tables b and *c*, and the comparison between *a* and *c* is 1 since the sequence of *type of action* is exactly the same. Thus, this method is seriously flawed because it yields implausible similarity estimates.

Finally, let us look at the results of similarity estimation using LSA cosines. The vector representing the sample has been calculated as the average of the 8 action vectors. Example 1 vs. example 2 has a cosine of 0.721, a high similarity value. Even though these samples share only 1/8 of the actions, LSA has correctly inferred that the remaining actions, although different, are functionally related. Comparisons between 1-3 and 2-3 have cosines as low as 0.050 and 0.071 respectively, showing that these action sequences are different indeed.

Correlations between LSA and Human Judgment

More formal comparisons between the performance of LSA and human observers than mere plausibility judgments are also possible. The problem is that, contrary to what happens when one uses LSA to model text comprehension, it is not easy to find experts in the task at hand. Everybody is a good example of the expert reader, but few people are expert in controlling the particular dynamic system called *Firecshief*. To test our assertions about LSA, we recruited 3 persons and gave them extended practice, so they could learn the constraints of the task.

After 24 practice trials, these participants were used to assess the external validity of LSA similarities. Using *Firechief*'*s* replay option, participants had to watch 7 pairs of trials (at a pace faster than normal) and express similarity judgments about these pairs. People watched a randomly ordered series of trials, in a different order for each participant, which were selected as a function of the LSA cosines (pairs A, B, C, D, E, F, G with cosines 0.75, 0.90, 0.53, 0.60, 0.12 and 0.06 respectively). One of the pairs was presented twice to measure test-retest reliability. That is, for example, pair G was exactly the same as pair A for one participant, the same as pair F for another participant, etc. All the possible stimulus pairs were presented to each participant. Participants had to answer which pair seemed more similar. For example, LSA would say that pair B is more related than pair C, since the cosines are 0.90 and .53 respectively.

LSA cosines predicted human similarity judgments quite well. For 3 participants in this pilot study, the proportion of agreement LSA-human was 6/19, 14/19, and 13/19 respectively. Participants with strong agreement with LSA also showed more consistency in their judgments, that is they answered to the repeated item in the same way. The participant who had low agreement with LSA had performed poorly on the repeated item, which suggests that she may not have learned enough about the task or was not paying sufficient attention. Even so, the average agreement between LSA cosines and human judgments was 0.57, far superior to the agreement expected by chance, 0.5 ^ 19 = 2e-5.

Conclusions

LSA seems to be a promising new way of approaching Complex Problem Solving performance that overcomes some of the known limitations of previous methods. Apart from the features listed in the introduction, there are some additional pragmatic LSA advantages worth noting: (1) Since the basic unit of analysis is the token (action or state), even systems that are described in terms of nominal (discrete) variables can be analyzed. Both actions and states can be used as units. (2) The semantic matching mechanism permits discovery of similarities beyond simple coincidence in the log files containing actions or states. That is, participants who are using different interventions to realize the same strategy will be considered similar even if their log files share no actions (or states). (3) The level of granularity (whether we are working with individual tokens, slices of performance, whole trials, or collections of trials) need not be defined *a-priori*. Since every object, from one token to the participant's whole performance, can be represented as a vector in the high-dimensional problem space constructed by LSA, analyses can be performed at any level of detail.

There are, however, a number of limitations to the proposed method: (1) A huge sample of data is needed to construct

Table 4: Transitions between actions considering type of action only as described in Quesada et al. (Quesada et al., 2000) and Cañas et al. (submitted), for the examples 1,2 and 3. Cells contain frequencies of the transition defined by its row and its column. For instance, the number 4 in the center cell in table 4a means that in example 1 the transition move-move has appeared four times.

(a)

Example 1	drop	move	Control
Drop	0	0	0
Move	0	4	1
Control	0	1	1

(b)

Example 2	drop	move	control
Drop	0	1	0
Move	1	2	2
Control	0	1	0

(c)

Example 3	drop	move	control
drop	0	0	0
move	0	4	1
control	0	1	1

the problem space. (2) Order effects are not taken into account. This means that, for LSA, a trial where the tokens have been scrambled to a random order has exactly the same meaning as the original version This is a serious but, as we have shown, not a fatal limitation, as long as LSA is used with care in CPS tasks. (3) Though the SVD analysis is common practice and can be found in several statistical packages, a powerful computer is needed to run large analyses.

Acknowledgments

Our acknowledgements to Tom Landauer for proposing interesting issues concerning the selection of the unit of analysis in Complex Problem Solving. We are grateful to Kim Vicente and John Hajdukiewicz for sharing experimental data and insightful discussions. Many thanks to Bill Oliver, who provided passionate methodological discussions and theoretical contributions.

This research was in part supported by Grant EIA – 0121201 from the National Science Foundation.

References

Brehmer, B. (1995). Feedback delays in complex dynamic decision tasks. *In P. Frensch and J. Funke, (Eds.) Complex Problem Solving: The European Perspective. Hillsdale, NJ: Lawrence Erlbaum*.

Broadbent, D. E. (1977). Levels, hierarchies and the locus of control. *Quarterly Journal of Experimental Psychology, 32*, 109-118.

Buchner, A., Funke, J., & Berry, D. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *Quarterly Journal of Experimental Psychology, 48A*, 166-187.

Calderwood, R. (1989). The role of context in modeling domain knowledge. *Unpublished doctoral dissertation, University of New Mexico*.

Cañas, J. J., Quesada, J. F., Antolí, A., & Fajardo, I. (submitted). Cognitive flexibility and adaptability to environmental changes in dynamic complex problem solving tasks.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1991). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Dickson, J., McLennan, J, Omodei, M. M. (2000). Effects of concurrent verbalization on a time pressured dynamic decision task. *Journal of General Psychology, 127*, 217-228.

Dörner, D. (1975). Wie Menschen eine Welt verbessern wollten und sie dabei zerstörten (How people wanted to improve the world but destroyed it). *Bild der Wissenschaft, Heft 2*.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer enhanced learning On-line journal, 1(2)*.

Frensch, P., & Funke, J. (1995). Complex Problem Solving: The European Perspective. Hillsdale, NJ: Lawrence Erlbaum.

Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology, 16*, 24-43.

Gibson, F. P. (2000). Feedback delays: How can decision makers learn not to buy a new car every time the garage is empty? *Organizational Behavior and Human Decision Processes, 83*(1), 141 - 166.

Green, D. W. (2001). Understanding microworlds. *Quarterly Journal of Experimental Psychology, Section A-Human Experimental Psychology, 54*(3), 879-901.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition, 28*(3), 249-276.

Henderson, L., Klemes, J., & Eshet, Y. (2000). Just playing a game? Educational simulation software and cognitive outcomes. *Journal of Educational Computing Research, 22*(1), 105-129.

Howie, D. E., & Vicente, K. J. (1998). Measures of operator performance in complex, dynamic microworlds: Advancing the state of the art. *Ergonomics, 41*, 85-150.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., & the LSA research Group (2000). Developing summarization skills through the use of LSA-backed feedback. *Interactive Learning Environments, 8*(2), 87-109.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.

Klein, G. A., Orasanu, J., Calderwood, R., & Zsambok, C. E. (Eds.). (1993). *Decision making in action: Models and methods*. Norwood, NJ: Ablex Publishing Corporation.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Omodei, M. M., & Wearing, A. J. (1995). The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior. *Behavior Research Methods, Instruments & Computers, 27*, 303-316.

Quesada, J. F., Cañas, J. J., & Antoli, A. (2000). An explanation of human errors based on environmental changes and problem solving strategies. In P. Wright & S. Dekker & W. C.P. (Eds.), *ECCE-10: Confronting Reality*. Sweden: EACE.

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision making and system management. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15(2), 234-243.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Simon, H. A. (1981). *The sciences of the artificial*: MIT press.

Vicente, K. J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based work*. London: Lawrence Erlbaum Associates.

Winograd, T. (1972). A procedural model of language understanding. In R. Schank, Colby, K.(Ed.), *Computer models of thought and Language*. Sand Francisco: W.H. Freeman.

A Dynamical Connectionist Account of Conceptual Change

Athanassios Raftopoulos (raftop@ucy.ac.cy),
Andreas Demetriou (ademet@ucy.ac.cy)
Department of Educational Sciences, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus.

Abstract

Conceptual change can be accounted for at various levels of explanation. The cognitive level (Marr's computational level), the representational (Marr's "algorithmic"), and the implementational level. In this paper, we offer a dynamical account of types of conceptual change at the representational level. Our aim is to show that some classes of neural models can implement the types of change that we have proposed elsewhere. First we briefly describe at the cognitive level certain types of change that purport to account for some of the kinds of conceptual change. Then we lay forth the framework of dynamical connectionism; we discuss the representational level realizations of the cognitive level and claim that these can be depicted as points in the system's activational landscape. We offer, third, a dynamical account of some types change and we claim that conceptual change can be modeled as a process of modification, appearance of new and disappearance of attractors and/or basins of attraction that shape the system's landscape. Finally, we discuss the kinds of mechanisms at the representational level that could produce the types of change observed at the cognitive level, as modeled by means of dynamic connectionism.

Introduction

Conceptual change can be accounted for at various levels of explanation. Following Marr (1980), one can distinguish between three levels: the *computational,* the *algorithmic*, and the *implementational* level of explanation of cognitive systems. We prefer the term "cognitive" to "computational", and the term "representational" to "algorithmic", since there are accounts of cognition that deny the algorithmic nature of mental operations.

At the cognitive level, one can discuss cognitive operations that apply to information-processing content (such as addition and subtraction), operations that apply to structures as wholes, such as differentiation or coalescence (Carey, 1985; Chi, 1992), or, conceptual combination, generalization and abduction, and hypothesis formation (Thagard, 1992). This level addresses the issue of the functions computed by the information processing system.

At the representational level one can examine the algorithmic processes that realize conceptual change at the cognitive level by transforming representations, such as Newell and Simon's (1972) "problem behavior graph" in production systems. In the connectionist paradigm one can study the processes of the emergence of new attractors, and repositioning of points realizing representational states in high-dimensional state spaces (Horgan and Tienson, 1996), or the changes in the connection weights and network structure (Elman et al., 1996; Schultz et. al., 1995; Plunkett & Sinha, 1992).

In this paper, we will discuss a theory of different types of cognitive change and their implementation at the representation level. Our aim is to show how certain classes of neural networks could implement some of the types of change that the authors have proposed (Demetriou and Raftopoulos, 1999). First, we will summarize these types of change. In the second part we will sketch the framework for the dynamics of change, relying on the dynamical interpretation of connectionist networks to explore possible means of modeling the stipulated types. In the third part we offer a dynamical account of some types change and we claim that conceptual change can be modeled as a process of modification, appearance of new and disappearance of attractors and/or basins of attraction that shape the system's landscape. Finally, we discuss the kinds of mechanisms at the representational level that could produce the types of cognitive change.

To that end we will employ neural networks whose behavior can be viewed as falling under one or the other of our kinds of change, and describe the behavior that neural networks should exhibit if they are to implement type of change.

Types of Change

Demetriou and Raftopoulos (1999) previously published a theory of conceptual change that addresses the issue of how a learning system makes the transition from one state to another. The theory provides a detailed analysis of the types of change that are observed both in cognitive development and during learning. The types of change are summarized in Figure 1.

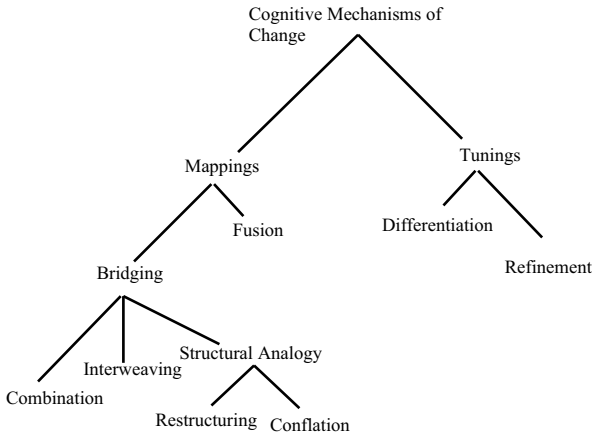


Fig. 1. The types of change

We will briefly present here *combination*, and *fusion. Bridging* is a class of types of change, whose unifying feature is that (a) two or more existing structures are brought together to bear on each other and form a more complex structure, and (b) after bridging the constituent structures retain their functional autonomy, even though they may have been modified. The blended structures may remain unaltered and the resulting structure(s) may retain the characteristics of the constituent structures (as when "striped" and "apple" are combined to produce "striped apple". In this case, the type of change is *combination*.

Fusion differs from bridging in that the mapped structures do not retain their relative autonomy after the mapping; instead, they fuse to one of the existing structures, or form a new structure. An example would be the fusion of retrieval and counting strategies, which are involved in simple operations of addition and subtraction performed by children aged 4-6. After fusion, around the age 6-7, the predominant strategy is retrieval by rote memory (Siegler, 1996).

The Representational Level

We discuss here the way representations can be modeled as properties of cognitive systems. At this level one examines the mathematical implementation of the cognitive level. In other words, we examine the way cognitive states are represented and how they are transformed and processed by means of operations performed on data structures.

These transformations can be either algorithmic (determined by a set of rules that apply to discrete static symbols that are the representations of the system) or dynamical (determined by mathematical relations that apply to continuous variables and specify their interrelations and evolution in time). This is why we call these transformative processes mathematical-state transitions; they describe the way the system moves between points in its state-space. We will address the issue of change from the perspective of connectionist theory interpreted in a dynamical way. Thus, will assume that a cognitive system is associated with a dynamical system physically realized by a neural network.

Neural Networks as Dynamical Systems

Recurrent neural networks (Elman, 1990) with distributed representations and continuous activation levels can naturally be construed in a dynamical way. They can be described by means of the evolution of the activation values of their units over time. To be able to model growth and avoid problems of lifelong (mainly catastrophic interference), one needs to consider a special class of networks, namely adaptive or generative networks. These networks can modify their structure during learning by adding or deleting nodes and can change their learning rates.

The number of units of the network determines the number of dimensions of the state-space associated with the system. Their activation values constitute the actual position in the state-space of the system. Adding a time-dependent parameter yields the phase-space of the system. Both in state- and phase-space, one can represent all the possible states that a system can take in time. Hence, in the connectionist account, the states of a cognitive system are depicted by the sets of activation values of the units that distributively encode these states.

These activation values are the variables of the dynamical system and their temporal variation constitutes the internal dynamics of the system. In addition to the state-space of a system, an external control space is also defined. The external space contains the real-value control parameters that control the behavior of the system, i.e., the connection weights, biases, thresholds, and, in networks in whose structural properties are implemented as real-value parameters (Raijmakers et. al., 1996), the structure of the system. In dynamical systems the fast internal dynamics is often accompanied by a slow external dynamics. The external dynamics consist of the temporal paths in the external control space. The external dynamics consist of the network's learning dynamics (the various learning rules) and the dynamics that determine structural changes, such as the rules for inserting nodes in cascade correlation and growing radial basis function networks.

When the network receives input, activation spreads from the input units to the rest of the network. Each pattern of activation values defines a vector or a point, within the activational space of the system whose coordinates are the activation values of the pattern. The activation rules determine the state transitions that specify the internal dynamics of the system, i.e., the functions of the evolution of the system in time. Thus, the behavior of such a system is depicted as a trajectory between points in the activational state space.

The activation rules, the number of units, the pattern of their connectivity, and the learning rate(s) of the network determine the architecture of the system. These factors are determined by its long-term history of experiences, since the class of networks discussed here may modify either their patterns of connectivity, as they learn, by adding nodes, deleting nodes, and sharpening their connections, or their biases and learning rates. The activation vectors and the behavior of the system evolve as a result of the synergies among the architecture of the network, the input it receives, and the previous activity of the network, under the control of the external dynamics.

The behavior of the system is a collective effect of cooperation and competition, (Kelso, 1995). The competition is due to the effort of the system to retain its current state in the face of incoming information. If this information cannot be assimilated by the system, then the weights of a network change and the network may alter its structure to accommodate the new input.

The activation states, in which a network may settle into after it is provided with an input signal, are the attractors of the system. These are the regions in state-space toward which the system evolves in time. The points in state-space from which the system evolves toward a certain attractor lie within the basin of attraction of this particular attractor. Thus, the inputs that land within the basin of attraction of an attractor will be transformed by the connectivity of the network so that they end up at this attractor where the system will settle.

Networks in which the outputs change over time until the pattern of activation of the system settles into one of several states, depending upon the input, are called attractor networks. The sets of possible states into which the system can settle are the attractors. If the network is used to model cognitive behavior, then the attractors can be construed as realizing cognitive states to which the system moves from other cognitive states that lie within the attractor's basin of attraction.

The signal of the input is transformed as it moves through the hidden units into an attractor pattern as follows: a given input moves the system into an initial state realized by an initial point. This input feeds the system with an activation that spreads causing the units of the system to change their states. The processing may take several steps, as the signal is recycled through the recurrent connections in the network. Since any pattern of activity of the units corresponds to a point in activation space, these changes correspond to a movement of the initial point in this state space. When the network settles, this point arrives at the attractor that lies at the bottom of the basin in which the initial point had landed. In this sense, the inputs fed into the system are the initial conditions of the dynamic system. Similarly to a dynamical system that settles into a mode depending on its initial conditions, a neural network settles into the attractor state in whose basin of attraction the input falls.

For instance, in a semantic network meanings of words are represented as patterns of activity over a large number of semantic features. However, only some of the combinations of semantic features are features of objects. The patterns that correspond to these combinations are the attractors of the network, which are points in the state space corresponding to the semantic features of the prototype of the object signified by the word. These attractors are the meanings of words.

The concepts "attractor" and "basin of attraction" suggest a way of simulating the classical notion of symbol. The attractor basins that emerge as the network interacts with specific inputs might be construed to have symbolic-like properties, in that inputs with small variations that fall within the same attractor basin are pulled toward the same attractor (or cognitive state) of the system. Thus, various inputs (tokens) give rise to the same stable point of attraction, the attractor (type), which in this sense offers a dynamical analog of the classical symbol (Elman, 1995).

The dynamical "symbols", unlike the symbols of classical cognitivism, are dynamic and fluid rather than static and context independent. The dynamic properties result from the dynamical nature of the activations of associative patterns of units. As the network learns and develops, the connection strengths continuously change. The same happens when new units emerge and old units "die" and the system reconfigures to maintain its knowledge and skills. All these cause changes in the original pattern in which an attractor/symbol was created in the first place, and as a result, subsequent activations differ. The same effects are caused by the different contexts in which the "symbol" may be activated. This happens because connections from the differing contextual features bias the activation of the units of the original pattern in different ways emphasizing some feature of the pattern or other. Thus, the attractor/symbol is almost never instantiated with the same activation values of the units that realize it.

The activational state-space of a network is a high-dimensional mathematical landscape. The state transitions in such a system are trajectories from one point on to another. Attractors correspond to cognitive states and the activation pattern that realizes each state is a vector, or a point. Thus, cognitive states are realized by points on this landscape. Since the distributed encoding of a cognitive state does not involve all units of the system, there will be points on the activational landscape that will realize more than one cognitive states (the set of coordinates of a point may satisfy the partial coordinates given by several activational vectors).

During the phase of activation-value changes the system passes through various possible outputs. All these outputs can be viewed as lying on an energy surface. When the system passes through a certain output-state whose energy is not lower than the energies of the neighboring states, it goes through another phase of activation-value changes in order to reduce the energy of the output state. When it reaches a point at which all the neighboring states have higher energies, it settles.

These states of minimum local energy are the attractors and can be construed as valley bottoms on energy surfaces. Thus, attractors should be distinguished from the networks' outputs in general. Not all outputs are settling points. Attractors form a subset of the set of outputs of a network, in that they are those outputs at which the system can settle. When the input of the system is such that the activation state of the system lies within the walls of the valley, the system will settle at the attractor at its bottom. Hence, the valley is the basin of attraction that leads to the specific attractor-state of minimum energy. Since the network has many attractors and basins of attraction, their relative position shapes the relief of the activational landscape of the system.

Modeling the Dynamics of Cognitive Change

In this theoretical framework, cognitive change results from the molding of the activational landscape, as a result of changes in the weights and the architecture of the network,

as the network attempts to accommodate new input signals. The molding may result either in the emergence of new, and/or disappearance of old, attractors, or in the reshaping of the basins of attraction. This process corresponds to a trajectory on the activational landscape. The idea that change is to be modeled by means of transitions in the state space of a dynamic system is at the heart of dynamical theories of cognition. Transitions in the state space of a dynamic system substitute for the algorithmic syntactically governed transitions of cognitivism.

The relief of the landscape determines the trajectories that are allowed, and the possible transformations among cognitive states. Cognitive change, thus, depends on the activational landscape of the system that learns. When information enters, the system tends to assimilate it within the existing framework of knowledge, which, in neural networks, is determined by the connection weights and the architecture of the network, which, in their turn, distribute the points that realize cognitive states on the network's landscape. We have posited certain types of cognitive change. In what follows we will sketch their dynamic realization at the representational level.

Combination

This type of change involves the combination of structures in such a way that the existing attractors and the landscape's relief (their basins of attraction) of the system are not affected. The new structure is superimposed, as it were, on the constituent structures. Consider the networks that simulate learning to pronounce words and non-words (Plaut et. al. 1996). These networks learn the pronunciation of both regular and irregular words, by building the appropriate attractors. The attractors of regular words consist of componential attractors, in which case the basin of attraction is the intersection of the sub-basins of attraction of the componential attractors. The exception words have their own attractors with a lesser degree of componentiality. Combination explains the ability of the network to learn the pronunciation of words and non-words, in that this knowledge is the result of the combination of the sub-knowledge encoded by the componential attractors, as is shown in Figure 2.

In this figure only two componential attractors are depicted, for onset and the vowel in the reduced two-dimensional activation space of the phonemic units of the network. The basins of attraction for the word "by" and the non-word "dy" are the intersections of the sub-basins for pronunciation of *b*, *d*, and *y*, that is, the regions in the state space in which these sub-basins overlap. The black circle is the attractor for the word *by*, and the striped circle is the attractor for the non-word *dy*. The trained network learns to pronounce words by applying its knowledge regarding the pronunciation of the parts of the words (and of the role of context in pronunciation when it comes to exception words). The reduced componentiality of the exception words is depicted by means of a deformation of the intersection of the salient attractors for the onset *d* and the vowel *o*. The componential attractors and their basins of attraction remain unaltered.

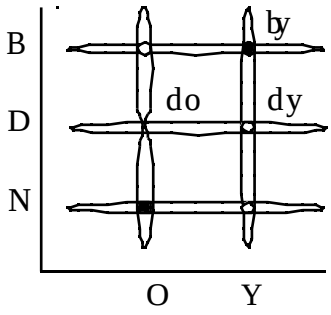


Figure 2. Componential attractors

After the new pronunciation is learned, the two basins change their relative positions so that they intersect. Their intersection (i.e., the pattern corresponding to both sets of features) forms a new basin, which is the area in which the two basins overlap. The appearance of a new basin of attraction represents the learning of a new concept. The new basin of attraction is superimposed onto the two intersecting basins. The basins of attraction (sub-basins) and the attractors do not change. Whatever input was falling within one of the two basins before learning, still does so after the network has learned the new concept. The only change after learning is that some inputs fall within both the new basin and the old basins of attraction, This is a result of the superimposition of the new basin of attraction onto the two sub-basins.

Fusion

Stable structures within the neural net can be thought of as attractor states. Thus, the activation pattern of the structure attracts all other activation patterns that are similar enough with it (that is, all activation patterns that fall within the basin of attraction of the attractor). As a network learns, a new attractor state may emerge, which swallows the attractors that existed before. This is what happens in fusion. The two initial basins of attraction are also swallowed by the new one, so that all patterns that were falling within the one or the other now fall within the new basin of attraction. The system undergoes a phase transition that can be described as a reverse Hopf bifurcation (Figure 3), in which two stable states (bistability) are fused and disappear, and one stable state emerges (unistability).

Figure 4 displays the phase transitions associated with the fusion of "counting from one" and "memory retrieval" strategies (used by 4-6 year old children in simple arithmetic tasks) to the "memory retrieval" strategy that becomes predominant between 6 and 7 years of age.

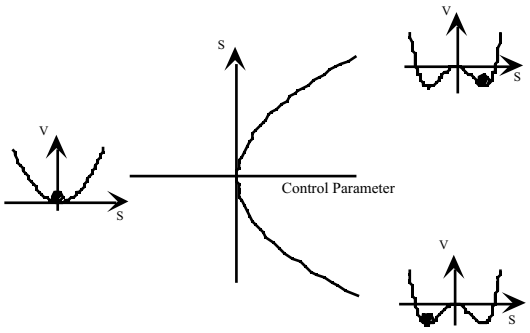


Figure 3. Fusion as an inverse Hopf bifurcation

The generative networks designed by Schultz et. al., (1995) to model a series of cognitive tasks simulate the variability of the strategies available to children. Networks at some stage of their training in the balance-beam tasks may "employ" two different strategies to solve the same problem and, as training continues, progress to using reliably the more advanced strategy. These networks implement "fusion", by moving from bistability to unistability.

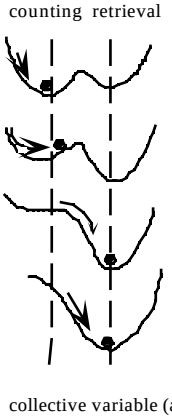


Figure 4. Fusion of two counting strategies

Strong and Weak Cognitive Change

In this context of dynamical connectionism, cognitive change consists in changes in connection weights, the structure, and the learning rates of the network. In connectionist networks an individual's state of knowledge is determined by the weights of the hidden units. Cognitive representational change is regarded as the individual's actual path through the space of possible synaptic configurations, that is, as the transformations of the weight vector in an n-dimensional weight space, where *n* is the number of the weights.

The appearance of novel cognitive states, and thus, the appearance of new attractors and the disappearance of old ones, implies a change of relief in the network's landscape (molding). Since the relief depends on the structure of the network, that is, the number of nodes, their connectivity, and the activation functions, its molding is the result of changes in the structure of the network. Networks evolve as a result of the system's effort to adapt to a new environment, by superimposing new representations to old ones. Thus, the system modifies the "knowing assumptions" that do not fit in.

The account of cognitive change at the representational level allows us to recast the discussion regarding strong and weak representational change in terms of dynamic systems theory. Whether a cognitive change is weak or strong depends on whether the new structure increases the representational resources of the system. Since representations are points in the state space of the system, the expressive contents of the system correspond to such points. If the relief of the landscape is such that the system cannot settle at a content realizing point, that is, if this point is not a possible attractor state, the content that is realized by this point is not within the expressive capabilities of the system. When changes in the relief render this point an attractor, the change is strong; it results in an increase in representational power.

But the mere appearance of an attractor does not necessarily imply that a radical change has taken place, that is, that this is a novel attractor state. This is so if the content realizing point that appears as a new attractor was in fact expressible within the system; that is, if the system could have settled at that point, even if it had not done so, up to that time. When the structures "striped" and "apple" are combined an attractor state "appears and the system acquires the new concept of "striped apple". This "new" attractor is a region in the state space, which realizes the content "striped apple" and is superimposed on the attractors of "apple" and "striped". But this is not a novel attractor, because this content was already within the expressive power of the system, since the relief of the landscape was such that the system could have settled if fed with the appropriate input at this point. In other words, the "new" attractor was situated at a local energy minimum in which the system could have settled if it had been fed with the appropriate input (the experience of a striped apple). The attractor appears without the landscape being molded and this attractor is just the sum of information expressed by the other attractors, which remain intact. In this case, the ensuing change is weak.

Weak change refers to changes in the semantic content of representations, which broaden their field of application but do not increase the expressive capabilities of the system. Attractors are merely repositioned in the landscape, which means that the activation patterns that define them do change. Reposition of any content-realizing point is accompanied by changes in the activation values that constitute the point's activation pattern, and changes in its spatial relations with other content realizing points. Since semantic information in dynamic systems is captured by the relative positions of content realizing points, repositioning is accompanied by semantic change.

This scenario does not apply to the case of fusion. No mere intersection of existing basins of attraction or any simple repositioning, could accommodate the salient input. A reshaping of attractor basins is required, as well as the

disappearance of an older attractor and the emergence of a novel one. These changes mould the landscape.

Thus, when new information is learned with repositioning of attractors and basins of attraction, and attractors are preserved (though the slope of the basins may change, with some becoming steeper and others becoming less steep), the resulting change is weak. Updating the connection weights seems to suffice for this. If the change in weights does not suffice for learning, the landscape is molded by changes in the network's structure (Horgan and Tienson, 1996). This may induce the appearance of new attractors; since the attractors are points on the landscape, the appearance of new cognitive states realizing points on this landscape, and the disappearance of old constitute strong changes, since the content-expressive power of the system increases. This process may require structural, i.e., qualitative, change.

Mechanisms of Change

At the cognitive level, the main Piagetian mechanisms driving conceptual changes are assimilation, accommodation, and equilibration. It is time now to consider the mechanisms driving change at the representational level. In each of the types of change discussed previously the processes that lead to the change are the same, always reducing to quantitative and qualitative changes in connection weights and the architectural structure of the network. These processes cause the repositioning of existing attractors, the disappearance of old ones, the appearance of new ones, and changes in the basins of attraction that shape the relief of the landscape. It could hardly be otherwise. In connectionism the computational mechanisms are domain general, statistical learning mechanisms, based on brain-style computation, that is, (a) on the spreading of the activation of each unit to other units, (b) on the modification of the connection weights, and (c) on the modification of the network structure.

McClelland (1989) argued that Piagetian "assimilation" corresponds to the activation spread in a network when a signal is presented to the input units and propagates through the network causing the activation of its units. The alteration of the weights, as a result of the network's learning, models Piaget's "accommodation", that is, the change that the network undergoes trying to fit in new experiences. Shultz, et al., (1995), and others, have proposed networks that adapt their structure as they learn by increasing their hidden units to accommodate the demands of the task. They offer a variation of McClelland's account that is suited better for networks that can modify their structure. The quantitative phase of error reduction and weight change may correspond to Piaget's "assimilation" of information in a pre-existing structure, whereas the qualitative structural change corresponds to Piaget's "accommodation" of the system. Quantitative change renders possible knowledge acquisition within a fixed representational framework, whereas qualitative change allows an increase in representational power.

References

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: The MIT Press.

Chi, M. T. H. (1992). Conceptual Change within and across Ontological Categories. In R. Giere (Ed.), *Cognitive models of science*. Minnesota University Press, 112-136.

Demetriou, A., & Raftopoulos, A. (1999). Modeling the developing mind: From structure to change. *Developmental Review, 19*, 319-368.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L. (1995). Language as a dynamic system. In R. F. Port & T. Van Gelder (Eds.), *Mind As motion: exploration in the dynamics of cognition.* Cambridge, MA: The MIT Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development.* Cambridge, MA: The MIT Press.

Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology.* Cambridge, MA: The MIT Press.

Kelso, S. (1995). *Dynamic patterns: the self organization of brain and behavior.* Cambridge MA: The MIT Press.

Marr, D. (1982). *Vision: A computational investigation into human representation and processing of visual information.* San Francisco, CA: Freeman.

McClelland, J. L. (1989). Parallel distributed processing: implications for cognition and development. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology.* Oxford: Oxford University Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving.* Englewood Cliffs, NJ: Prentice Hall.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology, 10*, 209-254.

Raijmakers, M. E. J., van der Maas, H. L. J., & Molenaar, P. C. M. (1996a). Numerical bifurcation analysis of distance-dependent on-center off-surround shunting neural networks. *Biological Cybernetics, 75*, 495-507.

Shultz, T. R, Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: new approaches to process modeling*. Hillsdale, NJ: Erlbaum.

Siegler, R. S. (1996). *Emerging minds*. Oxford: Oxford University Press.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Thagard, P. (1992). *Conceptual Revolutions.* Princeton, NJ: The Princeton University Press.

Deictic Codes, Demonstratives, and Reference: A Step Toward Solving the Grounding Problem

Athanassios Raftopoulos (raftop@ucy.ac.cy),
Department of Educational Sciences, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus

Vincent C. Müller (vmueller@act.edu)
Department of Philosophy & Social Sciences, American College of Thessaloniki
P. O. Box 21021, 55510 Pylea, Greece

Abstract

In this paper we address the issue of grounding for experiential concepts. Given that perceptual demonstratives are a basic form of such concepts, we examine ways of fixing the referents of such demonstratives. To avoid 'encodingism', that is, relating representations to representations, we postulate that the process of reference fixing must be bottom-up and non-conceptual, so that it can break the circle of conceptual content and touch the world. For that purpose, an appropriate causal relation between representations and the world is needed. We claim that this relation is provided by spatial and object-centered attention that leads to the formation of object-files through the function of deictic acts. This entire causal process takes place at a pre-conceptual level, meeting the requirement for a solution to the grounding problem. Finally we claim that our account captures fundamental insights in Putnam's and Kripke's work on "new" reference.

Introduction

John Campbell (1997) claims that the problem of the reference of concepts is the problem of relating concepts with imagistic content. His 'imagistic content' is the content of our experiences as we consciously access it and use it to see things as 'being such and such'. The most basic form of reference is when one perceives a thing and refers to it on the basis of one's perception by using a demonstrative, such as, "that" or "this". The reference of such a perceptual demonstrative is determined by spatial attention.

Though Campbell's thesis regarding the importance of spatial attention and perceptual demonstratives in establishing the reference of concepts to spatiotemporal objects goes in the right direction, his account has some serious problems. The main issue we wish to take on here is his view that the matter of the reference of concepts is exhausted by relating propositional with pictorial content. The problem is that both propositional and experiential content are representations, while the issue of reference is supposed to be a matter of grounding representations to the world, not of relating representations of different kinds to each other. This classical threat of infinite regress is now known under the label of 'encodingism', the view that representations are connected with the represented entities via some kind of correspondence between the two (see the critique in Bickhard, 1993 and Christiansen and Chater, 1993). Solutions that propose some such correspondence that, in turn, stands in need of an 'interpretation' cannot answer the symbol grounding problem, since they fail to account for how "symbol meaning [is] to be grounded in something other than just more meaningless symbols." (Harnad, 1990) The encoding of symbols in further symbols cannot be the solution.

We know that the problem has a solution since (some of) our concepts do have reference. A solution that does not fall into the trap of encodingism could be provided if we could single out a non-symbolic connection between our representations and the world. It seems plausible that such a connection is a causal one and that it would take place without the involvement of conceptual (i. e. symbolic) means. This is where Campbell has pointed in the right direction.

So, our task is to give a contribution to the problem how the concepts in the mind of a particular speaker can refer to objects in the world. How is it possible that I can use the concept "house" to refer successfully to certain objects? We need to single out a causal process, that could be a grounding for such reference (and that would not presuppose concepts). Once such a grounding is laid, the speaker can expand the initial grounding from experientially accessible kinds of objects to objects to which he has no such access (for contingent or principled reasons, as in the case of abstract objects). Each use of a concept would depend on such a grounding through a causal chain, starting from the initial grounding(s). How the grounding of non-experiential concepts takes place is not discussed in this paper.

In this paper we argue that spatial attention and/or object-centered attention establish the referents of certain kinds of concepts, namely perceptual demonstratives. Demonstratives are a promising start because they might rely on bodily movements in a context, not on conceptual entities that would require an interpretation. We first discuss these forms of attention and the way they individuate objects, arguing that spatiotemporal information individuates the referents and that this can be done in a bottom-up, non-conceptual way.

Then we employ Garcia-Carpintero's (2000) and Devitt's (1996) theory of demonstratives to show how the senses of demonstratives individuate their referents (demonstrata). We claim that the senses of demonstratives use the spatio-temporal information contained in the object files to fix reference. The non-conceptual use of this information provides the causal relation that grounds representations in the world. To explain how individuation takes place we employ Ballard et. al.'s (1997) theory of deictic codes and Kahneman and Treisman's (1984) theory of object-files.

Finally, in the third part of the paper, our thesis regarding reference is discussed in the context of Putnam's (1975, 1983; 1991) and Kripke's (1980) "new" theory of reference. We claim that the notion of an object-file containing predominantly spatio-temporal information provides the causal connection with the world that Putnam and Kripke sought to establish. We discuss the grounding of concepts whose referents are the basis of one's perception when one uses a demonstrative. In this sense the solution provided here, even if successful, is only the first step toward solving the problem of concept grounding in general.

Individuating Objects

Campbell (1997) argues that object individuation takes place by means of selective spatial attention that picks out objects features, forms feature maps, and integrates those that are found at the same location into forming objects in the way described by Treisman's *Feature Integration Theory* (*FIT*). In vision, information from different feature maps is bound together by extracting the location encoded implicitly in any feature information. Spatial attention makes the implicit location explicit. Information localized at the same location is bound together and thought to pertain to a certain object that occupies that space.

FIT belongs to a family of theories that hold that when one attends to an object then one automatically encodes all of its features in visual working memory. Against this, there is evidence for the existence of object-based attention which overrides featural information (other than spatiotemporal information) and which on certain occasions may pick out objects without any regard even for spatial information (Scholl & Leslie, 2000; Scholl, 2001). The role of object centered attention is primarily the parsing of a scene into discrete persisting objects, and the selection of some among these objects. The same evidence suggests that selection based on spatio-temporal information occurs very early in information processing (though segmentations of a scene into various discrete objects probably occurs at all levels of vision); in the case of vision it takes place in mid-level vision. Mid-level vision is bottom-up and cognitively impenetrable (Pylyshyn 1999; Raftopoulos, 2001), i. e. not accessible to conscious cognition, so it is not conceptual. In other words, some form of the selection of objects and the parsing of a scene is a bottom-up, cognitively impenetrable process (Carey and Xu, 2001; Scholl, 2001). Such a process would be a good candidate for the causal process we are looking for.

Given the pre-conscious processes in mid-level vision, we need to distinguish two steps: 1) *object individuation*, the processes that selects objects as discrete entities that persist in time, and 2) *object identification*, the processes that lead to the representation of objects under a certain description. The latter involves a strong semantical component, in the sense that the object represented has been identified as being such and such (e. g. a house). The former involves a much weaker level of representation. It purports to convey the sense that an object file has been opened for that specific object, that is, that the object has been "catalogued" or "indexed" as *something* that exists and persists separately of other objects with its own continuous spatio-temporal history – not as something that has certain properties (such as that of being a house). Object-files are allocated and maintained primarily on the basis of spatiotemporal information. Objects can be parsed and tracked without being identified. This representation allows access to the object but it does not describe the object. Object individuation does not require the existence of a concept associated with that object. (Of course, in theory successful object identification could also be used for individuation, as in definite descriptions like "the large house with the porch", but this presupposes grounded symbols, so we are interested in the inverse: individuation without identification.)

As an example for object individuation, think of two identical red squares that are situated in different locations. Since they are identical with regard to their features, the only way they could be treated as two distinct objects is by considering their spatiotemporal history. This presupposes that there is an object-centered attentional mechanism that is sensitive only to spatiotemporal information and not to feature information, which can pick up these objects by opening object-files. Precisely this conclusion is reached in the *MOT* (Multiple Object Tracking) experiment (Pylyshyn & Storm, 1988). In these experiments subjects must track a number of independently moving identical objects, that are initially tagged by attentional cues, among identical distractors. The success in *MOT* presupposes that the subjects attend to spatiotemporal information (relative location and direction of motion) and not to features, such as color and shape, or even the actual location of the objects. One could say that the attentional cues individuate/index in parallel the targets by assigning them tags that the subject can follow afterwards through motion. Thus, this mechanism individuates these objects and allows the subject to follow their paths and transformations while maintaining their identity as distinct objects.

Carey and Xu (2001) argue that a mechanism tracking the spatiotemporal history of objects precedes feature tracking mechanisms, and that this mechanism may override conflicting featural information. In other words, object individuation precedes feature identification – as we said above. The cognizer does not "know" or "believe" that an object moves in continuous paths, that it persists in time, though it uses this information to index and follow the object. She does not encode the object's features in some concept. She may not even have acquired the concept "object" (in this sense, you can see a house without having the concept of "house"). Object individuation may eventually result in the belief that an object is here or there, but this indexing itself does not appeal to some stored concepts regarding objects. Hence, if object individuation establishes reference, then the reference of demonstratives is not determined by a set of descriptions of features.

The discussion of object based attention shows that object files are opened and maintained on the basis of spatiotemporal information by means of cognitively impenetrable mechanisms. Petitot (1995) talks of the "positional (local) content-structure" of the scene. This positional structure is nonconceptual, and conveys information about nonvisual properties, such as causal relations (e.g., *X* "transfers" something to *Y*). In this latter category one can include the

functional properties of objects, referred to as 'affordances' of objects. Suppose that one witnesses a scene in which *X* gives *Z* to *Y*. The semantics of the scene consists of two parts: (i) the semantic content of *X*, *Z*, *Y* and "give" as a specific action, and (ii) the purely positional local content. The latter is in fact the image scheme of the "transfer" type. *X*, *Y*, and *Z* occupy a specific location in the space occupied by the scene. In the image scheme, *X*, *Y*, *Z* are thus reduced to featureless objects that occupy specific relative locations, and in that sense can be viewed as pure abstract places. More specifically, *X*, *Y*, and *Z*, which in a linguistic description of the scene are the semantic roles, "are reduced to pure abstract places, locations that must be filled by 'true' participants." These places are related by means of an action, of a "transfer" type.

Petitot's "places" do not refer to the actual locations that are occupied by the objects in a scene. What Petitot seems to allude to using the spatial metaphor is the notion of an object devoid of all features (including actual location), except that it persists in time, and occupies some space – similar to the talk of "two-place predicate". It is this individuation of an object that Petitot seeks to describe by saying that the objects' only property is that they occupy their own space and this is what the notion of objects as "pure abstract places" purports to convey. In this framework, the concepts that are used in the linguistic descriptions of a scene are locational configurations, that is, spatial structures. Petitot describes the routines and algorithms of early vision that might retrieve from the morphology of a scene in a bottom-up manner the global positional information contained in it.

Several theories of mechanisms of object indexing (what we call "individuation") have been proposed. They include the *FINST* theory of visual indexing (Pylyshyn, 2001), the object-indexing theory (Scholl & Leslie, 1999), the object-files theory of Kahneman & Treisman (1984) and more recently Ballard's et. al., (1997) theory of deictic codes. The common thread of these theories is the claim that there exists a level of (visual) processing in which objects present in a scene are parsed and tracked as distinct individual objects without being recognized as particular objects that are such and such. Thus, they stress the point that object individuation precedes object identification and that there is a level of object representation that does not encode features and does not presuppose concepts; a preconceptual level of object representation. We have already discussed «object-files». We are now going to describe a plausible mechanism that allows object individuation and tracking.

Deictic Codes, Object Individuation and Tracking

A recent theory of deictic pointers has been developed by Ballard, et. al., (1997). They claim that at time scales of approximately 1/3 of a second, orienting movements of the body play a crucial role in cognition and form a useful computational level. At this "embodiment level," the constraints of the physical system determine the nature of cognitive operations. "The key synergy is that at time scales of about 1/3 of a second, the natural sequentiality of body movements can be matched to the natural computational economies of sequential decision systems through a system of implicit reference called deictic in which pointing movements are used to bind objects in the world to cognitive programs."

Our discussion revolves around the issue of relating internal representational states with the world. We have emphasized the potential role played by non-conceptual processes in mediating the relation between representations and the world. Now, as we mentioned, the shortest time at which bodily actions and movements, such as eye movements, hand movements, or spoken words, can be observed is the 1/3-second-time scale, the embodiment level. Thus, the mechanisms that relate conceptual content with the world through action must be sought at this level. Suppose that one looks at a scene and selects a part of it through eye focusing. The brain's representation is about, or refers to, that specific part of the scene. Acts such as the eye focusing are called "deictic strategies", from the Greek word 'deixis' (pointing at), since they are analogous to pointing with one's hand. When one's internal representation refers to an object through such a deictic representation, this is a "deictic reference."

Eye fixation exemplifies the role of deictic mechanisms, or pointers, as grounding devices, that is, as devices that ground internal representations and cognitive programs to objects in the world, through deictic reference. This binding is implemented by two functional routines in the visual system. When a scene is perceived, the eye movements perform two main functions; they extract properties of pointer locations (object identification) and they point to aspects of the environment (object localization). The second task is that of our object individuation.

Perceptual Demonstratives and Reference

Let us start by having a brief look at the three most influential accounts of demonstratives. The standard Fregean analysis of demonstratives considers them similar to definite descriptions and assigns them a reference (Bedeutung) and a sense (the mode of presentation of the referred object). Frege's senses are descriptive, in that they provide descriptions in terms of features of the singular term. However, demonstratives do not function quite like definite descriptions do, since demonstratives and indexicals in general are rigid designators (Kripke, 1972) whereas definite descriptions are not. A token of "that house" refers to the salient house, while "the largest house in town" may refer to one house today and to another next year. This has been used to argue that the senses of demonstratives, if any, are not descriptive. Below, we will argue that the 'senses' of demonstratives consist in causal chains that ground them in the world (a thesis similar to that of Devitt, 1996, Kripke, 1980 and Putnam, 1975; 1991). The causal chains start with a direct perceptual encounter with an object, an encounter that grounds the demonstrative in the world. Devitt (1996, 164) calls such an encounter a "grounding" – and we shall use this for the "grounding problem".

We join Garcia-Carpintero (2000), and Devitt (1996) in the view that the difference between definite descriptions and demonstratives does not discredit the role of senses of demonstratives in determining their content. Our thesis is that the content or meaning of a demonstrative consists both of its referent and its sense. This is the second main account

of demonstratives. We will not argue in favor of this view, though, because it is really not crucial to the main argument developed in this paper, namely that reference construed as object individuation can be fixed by means of bottom-up perceptual processes that involve non conceptual content. What is important to this argument is the existence of such a process; this claim is independent of whether the referent is part of the meaning of demonstratives. The argument, however, as we shall see, essentially involves the role of the mode of presentation of a demonstrative in individuating objects. Thus, our claims go against the third important construal of demonstratives.

This is the direct reference theory, according to which the only content of a demonstrative is its denotation, or in other words, that the only linguistic function of a demonstrative is that it refers to its demonstratum, its referent (Kaplan, 1989). It does not have a sense. Paraphrasing Kaplan's (1989) account of the theory, one could say that a demonstrative does not describe its referent as possessing any identifying properties, it only refers to it. Though we agree that demonstratives do not provide identifying descriptions of their referent, we argue that they allow the individuation of the referent as a singular persisting object, by means of object-centered attention and spatial attention. These two provide the causal chains that ground the demonstrative. Thus, the mode of presentation of a demonstrative is not descriptive but causal. They can do that because they have a mode of presentation of the referent. But what is the mode of presentation when one says for example "that" pointing to a house?

Campbell (1997) thinks that the problem of the sense of a perceptual demonstrative is a problem about selective attention, in so far as he considers the mode of presentation to provide imagistic information related to the referent. It is the role of selective spatial attention to isolate that information in a scene that pertains to the referent. Thus, Campbell takes the mode of presentation of a demonstrative to include information that could individuate the referent on the basis of its observable features and, in an essential manner, on the basis of its spatial location. In fact, difference of location only suffice to establish difference in the mode of presentation of the same object by two different demonstratives.

Garcia-Carpintero (2000) and Devitt (1996) offer a thorough account of the senses of demonstratives, which is similar in some respects to that of Campbell's. The sense, according to Garcia-Carpintero, is an ingredient of presuppositions of acquaintance with the referent; "presuppositions" meaning "propositions that are taken for granted" when a statement is uttered. In this fashion, senses are individuating properties that allow the individuation of the referent.

Suppose one perceives something as being a house and utters the statement "that is *f*" pointing at a certain object (the house) and assigning it the property *f* (e. g. "beautiful"). The term "that" is a singular term associated with the description "the *f* house". According to Garcia-Carpintero, when one uses the singular term "that" one takes oneself to be acquainted with an object by having a 'dossier' for "the *f* house", which picks it out. The object fulfills the conditions specified in the dossier, in our case the proposition "there is a unique house most salient when the token *t* of 'that' is produced and *t* refers to that house." Now, the phrase "most salient when *t* occurs" is equivalent to the expression "house in such and such a location with such and such visual features." The "in such and such a place with such and such visual features" is the mode of presentation of the token *t* of the demonstrative "that". This mode of presentation individuates the object to which the demonstrative refers.

The dossier of the object that acquaints one with the object can be updated by new information, by adding contents or by revising its content. One notes a distinction between an object being singled out as the referent of a demonstrative and its acquaintance dossier (file). The latter ontologically presupposes the former; one needs an object to create its dossier. One also needs to ensure that the object with such and such features at time t_1 is the same object with such and such features at time t_2. Perception must provide for a mechanism that establishes the existence of an object as a distinct entity and opens a dynamic file on it. One needs, in other words, a mechanism that individuates the demonstrata of perceptual demonstratives.

Object Individuation and Reference

Let us see where we stand with regard to the issue of the reference of perceptual demonstratives related to object individuation. When one uses a demonstrative one opens a file for the object being demonstrated. According to the psychological evidence, the first thing that this file does is to individuate the object based on spatiotemporal information. This ensures the existence of a distinct object whose paths in space and time can be tracked. The object file thus allows acquaintance with the referent of the demonstrative, and in this sense, it constitutes its mode of presentation. Kahneman's "object-file" becomes a truncated version of Garcia-Carpintero's "dossier", a dossier that contains only spatiotemporal information. As the object moves in space-time, feature detection mechanisms infuse the file with feature information allowing feature *identification* (the full "dossier".)

Let us investigate the power of the account sketched so far with Brian Loar's (1976) example, also used by Garcia-Carpintero (2000) to argue that descriptive senses fix the referents of the terms with which they are associated: Suppose that Smith and Jones see a man on the train every morning. One evening they watch a man being interviewed on a television show, they are unaware that this man is the same man they meet on the train every morning, and it so happens that during the show they have just been talking about the man on the train. Suppose now that Smith switches his attention to the man on the television and says, "he is a stockbroker", referring to the man on the television. Jones, unaware of Smith's attention switch, takes Smith to refer to the man on the train about whom they have been talking. Though Jones has correctly identified the referent, since the man on the train is the same as the man on the television, one feels that Jones has failed to understand Smith's utterance. This shows that the manner of presentation of singular terms is important even on referential uses for grasping the meaning of what is being communicated.

The upshot of Loar's example is that although Jones' belief to the effect that the man on the train is a stockbroker has the same truth conditions as Smith's belief that the man on the television is a stockbroker (since the referent in both beliefs is the same person), Smith is justified in holding his belief, whereas Jones' is not.

Jones missed the information that would have justified his belief, because he does not know that the man on the television and the man on the train are the same person. So, for Jones, and thus information pertaining to the former does not apply to the latter. To use the terminology of this paper, Jones has two different object-files; one for the person on the television and one for the person whom he meets on the train. The role of the mode of presentation of a singular term is to clarify this point, namely whether the object under consideration has been individuated the appropriate way. Spatiotemporal information purports to do exactly that: had Jones followed the spatiotemporal path of the person on the train, he would have known that it is the same person that appears on the television and he would have used all relevant information to update that person's object-file; so his belief would have been just as justified as Smith's.

It seems thus, that object individuation (the mode of presentation) is indispensable to fixing the referent of a perceptual demonstrative. The individuation is accomplished by opening an object-file fixing the object to which the demonstrative refers and allowing its tracking. In the course of tracking, additional information, e. g. on shape and color, may be added to the "dossier" to allow tracking in difficult circumstances (as when one thing is inside some other thing). It is essential for the success of concept grounding and the escape from the regress of encodingism that this individuation process is not cognitively penetrable. No conceptual content, no existing representations can be used in the individuation process, so it has to be inaccessible to conscious content-laden processing. Also, individuation should not be seen as establishing a concept – this is what happens in the step of identification. Individuation just grounds the concept, fixing it onto an object so that the concept can be "filled" with information.

New Theories of Reference

If object individuation can fix reference and if object individuation can be carried out without conceptual involvement, then reference can be fixed in a nonconceptual manner. Of course, this goes against the standard descriptive theories of reference, according to which a sign is associated with a concept in the mind, a "sense", which constitutes its meaning and determines what the sign refers to. It allows one to pick out the objects in the environment that are 'fall under' the concept. The reference of a word is fixed by certain of the descriptions associated with the word: that thing over there counts as a "house", given that it is a building which could be used as a human dwelling.

A problem with these kinds of theories has been expressed in terms that remind strongly of the symbol grounding problem: Devitt (1996, 159) argues that descriptive theories of reference are incomplete because by explaining references by descriptions, they appeal to the application of descriptions of other words; thus, they explain reference by appealing to the reference of other words. To escape the lurking infinite regress, there must be some words whose reference does not depend on that of other words, that is words that are founded directly in the world.

Kripke (1972) and Putnam (1975; 1983; 1991) have argued that the standard conception of reference fails for certain kinds of words, namely demonstratives, proper names and natural kind terms. It is interesting to see whether our notion of reference is compatible with Putnam's (1975; 1983) direct-reference theory ("direct" in that it avoids the mediation of conceptual content in establishing reference). According to this theory, descriptions ascribing properties would identify the wrong referents of the terms. Once a causal contact between concept and object is established, the world itself has a say on the fixing, what Putnam (1991) will later call the "contribution of environment".

Putnam (1991) argues that there is an indexical (deictic) component that participates in reference fixing. When one takes a liquid sample to be water, one does so because one thinks that this liquid sample has a property, namely, "the property of behaving like any other sample of pure water from our environment" (Putnam 1991, 33). This property is not a purely qualitative property (meaning that membership is not determined by a set of criteria); its description involves a particular example of water, one given by pointing or focusing (hence, the term '"indexical"). The stuff out there, to which the act of pointing is an essential part of fixing reference of the natural kind term "water", is the contribution of the environment. Putnam says that "… the extension of certain kinds of terms … is not fixed by a set of 'criteria' laid down in advance, but is, in part, *fixed by the world*. There are *objective laws* obeyed by multiple sclerosis, by gold by horses, by electricity; and what it is rational to include in these classes will depend on what those laws turn out to be." (Putnam 1983, 71). This brings into mind the notion of causal chains by means of which demonstratives refer, causal chains that are established through object based and spatial attention.

Kripke (1980) refers to this assigning of names as "initial baptisms". Suppose, that one points to a star and says, "that is to be Alpha Centauri" (Kripke, 1980, 95). By this one commits himself to the following: "By 'Alpha Centauri' I shall mean that star over there with such and such coordinates." Kripke (1980, 135) claims that the reference of general natural kind terms is similarly fixed: "the reference (of singular terms) can be fixed in various ways. In an initial baptism ostentation or a description typically fixes it. … The same observation holds for such a general term as 'gold'."

These are telling examples, because they point out the role of spatial information and of object based attention in fixing the reference of singular terms. The causal chain that grounds the term starts with spatio-temporal non-descriptive information that opens an object-file for some object. This way of fixing the referents of singular and natural kind terms captures adequately Kripke's intuition that: "Don't ask: how can I identify this table in another possible world, except by its properties? I have the table in my hands, I can point to it, and when I ask whether *it* might have been in another room, I am talking, by definition, about *it*. I don't have

to identify it after seeing it through a telescope. If I am talking about it, I am talking about *it*." (1980, 52-53). Though Kripke speaks of proper names, his analysis easily transfers to all singular terms, and thus, to perceptual demonstratives (Garcia-Carpintero, 2000). Names and indexicals, are associated with something extralinguistic, their referents. Some existentially given thing is essential in fixing these referents.

We have claimed that the mode of presentation of the referent by its demonstrative is essential in reference fixing, and we have argued that the mode of presentation fixes reference by opening an object-file for the referent of the demonstrative. This object-file includes spatiotemporal information and its function is to individuate the referent, that is, to establish the existence of a distinct body that perseveres through space and time. It establishes the causal continuity with the thing originally "pointed at" by the perceptual demonstrative, satisfying Putnam's criterion for reference fixing. It also provides the causal relation between the representation and the world that grounds the former in the latter. The object-file provides the indexical component that participates in reference fixing. The content of the object-file being retrieved in a bottom-up manner warrants that this object file is the 'contribution of the environment' and not the contribution of conceptual content.

Conclusion

We argue that perceptual demonstratives capture the essential way in which one refers to objects in one's experience. The sense of "demonstrative reference" involved, however, departs from the notion of the referent as an object that is individuated by some description. The representation of the referent in the sense intended here does not encode any featural properties, is pre-conceptual, and the process that leads to its formation is cognitively impenetrable. The only property that the individuated referent has is that it is being tagged as a discrete object that persists in time and occupies some space, and thus, is being rendered accessible to the viewer. We claim that this process of reference fixing provides the causal relation required to solve the grounding problem.

References

Ballard, H., Hayhoe, M., Pook, P., & Rajesh, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20,* 723-767.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5,* 285-333.

Carey, S., & Xu, F. (2001). Infant's knowledge of objects: beyond object files and object tracking. *Cognition, 80*, 179-213.

Campbell, J. (1997). Sense, reference and selective attention. *Proceedings of the Aristotelian Society,* 55-74.

Cristiansen, M. H., & Chater, N. (1993). Symbol grounding - the emperor's new theory of meaning. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Devitt, M. (1996). *Coming to our senses: a naturalistic program for semantic localism*. Cambridge: Cambridge University Press.

Garcia-Carpintero, M. (2000). A presuppositional account of reference fixing. *Journal of Philosophy, XCVII*(3), 109-147.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42,* 335-346.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of the object files: object-specific integration of information. *Cognitive Psychology, 24*, 174-219.

Kaplan, D. (1989). Demonstratives. In Joseph Almog, John Perry, & Howard Wettstein (Eds.), *Themes from Kaplan*. New York: Oxford University Press.

Kripke, S. A. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Loar, B. (1976). The semantics of singular terms. *Philosophical Studies, 30*, 353-377.

Petitot, J. (1995). Morphodynamics and attractor syntax: constituency in visual perception and cognitive grammar. In R. F. Port & T. Van Gelder (Eds.), *Mind as motion: explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.

Putnam, H. (1975). The meaning of meaning. In *Mind, language and reality: Philosophical papers. Vol. 2.* Cambridge: Cambridge University Press.

Putnam, H. (1983). Reference and truth. In *Realism and reason: Philosophical papers, Vol. 3*. Cambridge: Cambridge University Press.

Putnam, H. (1991). *Representation and reality.* Cambridge, MA: MIT Press.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22*, 341-423.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80,* 127-158.

Pylyshyn, Z. & Storm, R. W., (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3*, 178-197.

Raftopoulos, A. (2001). Is Perception informationally encapsulated? The issue of the theory-ladenness of perception. *Cognitive Science, 25,* 423-451.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition, 80*, 1-46.

Scholl, B. J., & Leslie, A. M. (1999). Explaining the infant's object concept: beyond the perception/cognition dichotomy. In E. Lepore & Z. Pylyshyn (Ed.), *What is cognitive science?* Malden, MA: Blackwell.

Treisman, A. (1993). The perception of features and objects. In A. Baddeley and L. Weiskrantz (Eds.), *Attention: selection, awareness and control*. Oxford: Oxford University Press.

When the Fly Flied and When the Fly Flew: The Effects of Semantics on the Comprehension of Past Tense Inflections

Michael Ramscar (michael@cogsci.ed.ac.uk)
Division of Informatics
University of Edinburgh, 2 Buccleuch Place
Edinburgh, EH8 9LW, Scotland.

Abstract

Although previous theories of past-tense verb inflection have considered phonological and grammatical information to be the only relevant factors in the inflection process (e.g. Bybee & Moder, 1983; Rumelhart & McClelland, 1986; Kim, Pinker, Prince & Prasada, 1991), Ramscar (in press) demonstrated the importance of semantics in processing inflectional morphology. This paper presents an experiment that demonstrates the on-line effects of semantics on inflection. These findings indicate that regular and irregular inflections are determined by semantic and phonological similarities in memory, and furthermore that people are not responsive to the kind of grammatical distinctions amongst verb roots that default rule theories of inflection (Pinker, 1999) presuppose.

Introduction

In most theories -- and studies -- of past-tense verb inflection, phonological and grammatical information have been considered to be the two relevant factors in the inflection process (e.g. Bybee & Moder, 1983; Rumelhart & McClelland, 1986; Kim, Pinker, Prince & Prasada, 1991; Pinker, 1991; 1999). However, in some models of inflectional processing (MacWhinney & Leinbach, 1991; Joanisse & Seidenberg, 1999), semantic processes have been included to help explain the processing of homophone verbs (e.g. *brake*/*break*). Since *brake* and *break* both sound the same, phonology alone cannot distinguish which of *broke* or *braked* is to be the correct past tense form for the input breɪk.

Although using semantic information to guide this process appears intuitively plausible, this suggestion has been fiercely criticised by Pinker and colleagues (Kim et al, 1991; Pinker, 1999), who put forward an alternative, nativist account of homophone inflection (Pinker, 1991; 1999). This predicts that the regularisation of irregular sounding verb stems is driven by innate grammatical sensitivity: verbs that are instinctively perceived to be denominal will be automatically regularised. This account was supported by results reported by Kim et al (1991) which indicate that grammatical factors correlate better than semantic factors with people's ratings of the acceptability of past tense forms in context, although these results did not rule out any semantic role in inflection.

However, a recent series of experiments, Ramscar (in press) has demonstrated that the assumption that inflection is driven purely by grammar and phonology is flawed. A series of elicited inflection tasks showed that the semantic context in which a novel verb occurred influenced the forms participants later produced to mark the past tense of that verb. If participants first encountered the novel verb *sprink* in a context that involved consuming large quantities of fish and vodka (semantically similar to *drink*), they were likely to produce an irregular past tense form for it (*sprank*). But if they first encountered *sprink* in a context which presents as a verb to describe symptoms associated with a disease that involve rapid movements of the eyelid (semantically similar to *blink*), they were likely to go on and produce a regular past tense form (*sprinked*). Further, a comparison of the forms participants produced for the nonce verbs *sprink* and *frink* in a sparse, 'neutral' context (70% irregular) to those produced in the context involving rapid movements of the eyelid (70% regular) showed that the production of regular past tense forms increased when the semantic similarity between *sprink* and *frink* and the regular verbs *blink* and *wink* was increased. From these results it appears that not only irregular forms can be produced by analogy, but regular forms as well.

Semantics versus grammar in homophone inflection

Evidence that semantics affects inflection offers a solution to the homophone problem: different forms of homophone verbs are distinguished and computed by reference to their different meaning. Further, Ramscar (in press) contrasted the semantic account of homophone inflection with a nativist attempt to solve this problem put forward by Pinker and colleagues (Kim et al., 1991; Pinker, 1991, 1999, 2001) which predicts that the regularization of irregular sounding verb stems is driven by innate grammatical sensitivity: that any verb that is *perceived* to be denominal will be automatically regularized, resulting in different inflection patterns for denominal verbs that are phonologically identical to irregular deverbal verbs. Ramscar (in press) found that participants' sense of the semantic similarities between verb forms correlated strongly with participants preference for a regular or

irregular past tense form of a homophone verb in context (after partialling out the effects of grammar, r=.723), whereas participants perception of the grammatical origins of verbs correlated poorly with their references for irregular versus regular past tense forms (after partialling out semantics, r=.066). Further experiments showed that on both nonce and existing verbs, if the semantics of the verb were similar to those of an existing phonologically similar irregular, participants would favor irregular inflections even when they perceived the verbs to be denominal. Ramscar (in press) concluded that in fact, the grammatical origins of verbs had no effect on inflection, which was entirely governed by phonology and semantics.

One or two routes to inflection?

A further implication of these findings is that they undermine the one in principle objection to modeling past tense inflection using a single mechanism (Ramscar, in press). Pinker and colleagues (e.g. Pinker & Prince, 1988; Pinker, 1991, 1999, 2001) have claimed that the systematic regularization of verbs based on nouns would require two mechanisms for determining inflections, one method using phonological analogy (to explain cluster effects in inflection, resulting in forms such as *spling/splang*), and another method using grammatical information (i.e. a rule) to explain how verbs based on nouns are automatically regularized. The findings that semantics is used to distinguish homophone verbs and that the grammatical origins of verbs do not determine their past tense forms (Ramscar in press; see also examples such as *shoe/shod* versus *shoo/shooed* where the denominal verb is the irregular) obviate any *requirement* for models to account for this second, grammatically determined method of inflection.

Since it appears that single-route models *may* be entirely capable of modeling inflection patterns based on phonological and semantic properties (see e.g. MacWhinney & Leinbach, 1991; Joanisse, & Seidenberg, 1999) it appears that Rumelhart and McClelland's (1986) claim that single-route accounts provide "a distinct alternative to the view that children learn the rules of English past tense acquisition in any explicit sense…" merits further investigation. As Pinker (1991, 1999, 2001) has argued, the peculiarities of the irregular past tense system are best explained by an associative system based on analogy to stored forms, and not by rules: but if regular and irregular past tense forms *are* produced by the same mechanism – based on semantic and phonological analogy – then it may well be that learning the English past tense really does not involve acquiring a rule in any explicit sense.

The experiment described in this paper was designed to further probe this question. It was designed to examine the way in which semantics affects the comprehension of existing past tense forms. The dual-route model of past-tense inflection claims that regular inflection is unaffected by meaning or associative factors in memory (Pinker, 1991, 1999, 2001). In this experiment the meanings of existing verbs were manipulated to examine the effects of this on both their regular and irregular forms.

Experiment 1

This experiment was designed to test whether meaning has an effect on the comprehension of past tense verb forms by measuring the reading-times of regular and irregular forms of existing verbs in different semantic contexts. The dual-route model of inflectional morphology claims that the processing of regular past-tense inflection is unaffected by meaning or associative factors in memory:

> "[Regular inflection] is modular, independent of real-world meaning, non-associative (unaffected by frequency and similarity) sensitive to abstract formal abstractions (for example, root versus derived, noun versus verb), more sophisticated than the kinds of "rules" that are explicitly taught, developing on a schedule not timed by environmental input, organized by principles that could not have been learned, possibly with a distinct neural substrate and genetic basis." (Pinker, 1991, p. 534; see also Pinker 1999, 2001)

Accordingly, the dual-route predicts that semantic factors can only affect the comprehension of irregular forms. In line with the findings of Ramscar (in press, Experiments 2, 3 and 4), where semantics appeared to affect regular production, it was expected instead that semantics would affect the comprehension of all simple past tense forms. The contrasting single-route prediction tested here was that a regular past-tense form should be easier to read in a context that is semantically dissimilar to the ordinary usage of a phonologically identical irregular verb and an irregular past-tense form should be easier to read in a context that is semantically similar to the ordinary usage of a phonologically identical irregular verb.

Participants

Participants were 36 undergraduate students from Edinburgh University. All participants took part voluntarily in the study.

Materials

Four sets of materials examined four existing verb forms (*sink*, *fly*, *drink* and *food-drive*).

Each verb was presented in one of two contexts. In each context, the verb examined was introduced as a noun (to distinguish its meaning from ordinary uses of the corresponding irregular verb), and then later used as a verb. The contexts in which the verbs were presented were identical apart from a single semantic contextualizing sentence (shown in italics in Table 1)

that was varied across the contexts to manipulate the degree of semantic similarity between the verb and the ordinary irregular verb from which it was derived.

Table 1 - Example Context (The denominal verb is highlighted).

To promote business, the pesticide shop always stands a man in a giant fly costume at the entrance of their shop, to greet customers. This is especially fun for children. Whenever a child enters the shop, the greeter performs "the fly". *The greeter tells the children jokes and gives out prizes.* In the shop, the term to describe how the greeter greets children in this way is "to fly them". One hot day in June, sweating in his fly costume, I saw the greeter ***fly*** 40 children in a single afternoon. The look of tiredness on his face was really something.

Alternate context sentence
The child sits between the wings on the greeter's back, and they buzz up and down the aisles, ducking and swooping.

The two contextualizing sentences are italicized in table 1. The first context described an action that had some semantic similarity to *flying* simpliciter. The second context was semantically dissimilar to *flying* simpliciter. In order to obtain independent confirmation of the predicted semantic similarities, three naive raters were presented with the contexts on cards in randomized order and asked to order the contexts in each set according to how much the actions described in them matched the action they normally associated with the appropriate irregular verb (*fly*, *drink*, *sink*, and *drive*). The raters concurred with the ordering assigned to the contexts in the experiment, and inter-rater agreement was 100%.

Procedure

Participants told they were taking part in a memory study. Passages were presented on-screen and participants were instructed to memorize them. After memorizing a particular passage, participants were asked to indicate whether five sentences relating to the context passage were "True or False" by pressing the appropriate button on a computer keyboard as quickly as they could whilst concentrating on accuracy. The correct answer to three of these questions was "False" (e.g., in relation to the example in Table 1 participants were asked to state whether "The greeter was dressed as a pig" was true or false). The other two questions checked that participants remembered the noun use of the verb in question (e.g. "The greeter performs 'the Fly'") and also that they had remembered the semantic reinforcement sentences in the context. The correct answers to these questions were always "True." The presentation order of these five preliminary questions was randomized.

A final, sixth sentence presented to participants was also true, but it presented a fact stated in the initial context in a passive voice as an active past tense. This tense took either a regular or irregular form, e.g. in relation to "One hot day in June, sweating in his fly costume, I saw the greeter fly 40 children…" the fact was presented in an actively voiced manner, e.g.: "The greeter flew 40 children." or "The greeter flied 40 children."

The delay in milliseconds between the presentation of this sentence on-screen, and the onset of participants' responses was recorded.

Each participant was presented with one training item, followed by one context from each of the four sets of stimuli. Each participant completed one from each of the four conditions of the experiment (e.g. a context describing an action that was semantically similar to that implied by an existing irregular verb, with the verb inflected regularly in the target sentence (e.g. *fly – flied*), similar context / irregularly inflected verb, dissimilar context / irregularly inflected verb and dissimilar context / regularly inflected verb).

The experimental task was embedded in a series of unrelated tasks that participants also completed.

Table 2 - Mean reading times in milliseconds for the target sentences in Experiment 1.

	Semantically similar to irregular	Semantically dissimilar to irregular
drank	1490	2084
drinked	2759	1642
food-drove	1781	2166
food-drived	2435	1577
flew	2483	3051
flied	2776	1686
sank	1342	2890
sinked	1873	1582

Results

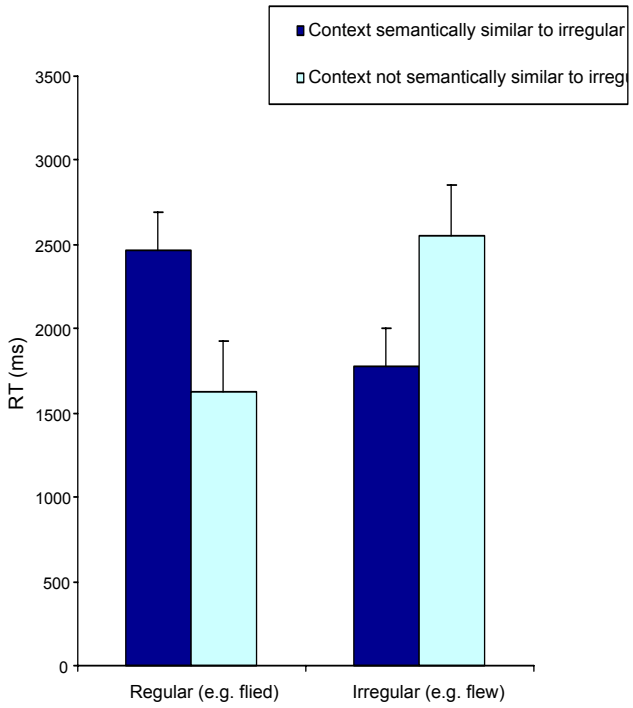
The mean reading time for each item is given in Table 2. Two unrelated t-tests showed that as predicted by single-route models (and in contrast to dual-route predictions) the target sentences containing the regular past tense forms of the verbs were processed faster in the dissimilar context (mean sentence reading time = 1622 ms) than in the similar contexts (2461 ms); $t(70)=3.282$, $p<0.001$. Irregular past-tense forms were processed more easily (1774 ms) when they had first been presented in an uninflected form in a context that was semantically similar to their ordinary usage as

opposed to a dissimilar context (2548 ms); t(70)=2.178, p<0.02.

Two-way repeated measures analyses of variance (ANOVAs) were conducted on the reaction time data, treating both subjects (F_1) and items (F_2) as random effects. There were no reliable main effects of either Meaning, $F_1(1,35)=0.23$, $p>0.87$; $F_2(1,3)=0.22$, $p>0.89$, or Grammaticality (Regular versus Irregular verb types) $F_1(1,35)=.235$, $p=<.63$; $F_2(1,3)=.309$, $p>.6$. The lack of a Meaning main effect indicates that, collapsing over the paragraph contexts in which the verbs were embedded, meaning did not produce a processing bias for the verbs. The lack of a main effect of Grammaticality indicates an analogous absence of bias for regular or irregular verbs forms.

There was a significant Meaning × Grammaticality interaction: $F_1(1,35)=12.911$, $p<0.001$; $F_2(1,3)=156.978$, $p<0.001$. As indicated by Figure 1, the interaction was due to Meaning effects at each level of Grammaticality (Regular versus Irregular verb types).

Figure 1. Overall reaction times in Experiment 1.



There were no significant increases in the error rate (participants answering "false" to statements that were assumed to be true) across all of the test sentences. For the true test sentences that were common to each context (the denominal and semantic reinforcement sentences) it was 10.4%. When the semantic context was consistent with the predicted verb tense the error rate for the target sentences was 12.5% and the inconsistent error rate = 9.7%. The error rates for particular tenses of the target verbs were 12.5% for irregulars and 9.7% for regular). Further ANOVAS were calculated considering only the "True" responses to the tests sentences containing the target verbs, which again showed no main effects of Meaning, $F_1(1,35)=0.138$, $p>0.71$; $F_2(1,3)=0.000$, or Grammaticality $F_1(1,3)=1.131$, $p=0.3$; $F_2(1,3)=.519$, $p>.5$, but did show a significant Meaning × Grammaticality interaction: $F_1(1,35)=10.635$, $p<0.005$; $F_2(1,3)=99.047$, $p<0.005$.

Discussion

Consistent with findings in ratings and elicitation tasks (Ramscar, in press), it appears from the results of this experiment that semantics affect the on-line comprehension of both regular and irregular past tense forms. Strikingly, the on-line processing of regular forms was significantly affected by semantics: if participants had to read "the greeter *flied* 40 children" in a context where to "do the fly" involved something like ordinary *flying* while dressed in an insect costume, it took longer to process than when "doing the fly" involved telling jokes and giving out prizes clad in the self-same fly outfit. This was despite the fact that the participants behavioral responses were identical in either instance: participants agreed in each case that it was true that "the greeter *flied* 40 children."

These findings are difficult to reconcile with the claim that the processing of regularly inflected forms is entirely "independent of real-world meaning" (Pinker, 1991). Further, the interaction between meaning and past tense form (i.e. whether a verb takes a regular or irregular inflection) in this experiment is hardly suggestive of a model in which two independent mechanisms are separately responsible for regular and irregular past tense processing, with one element – the regular – encapsulated and insensitive to the semantic factors that affect the other. Rather, it appears that both regular and irregular past tense comprehension relies upon a common, semantically – and phonologically – sensitive process.

General Discussion

For more than two decades the question of how inflectional morphology is processed has served as a battleground for conflicting theories of language, knowledge representation, and cognitive processing. On one side of the debate have been similarity-based or single-route approaches that propose that all past tenses are formed simply through phonological and semantic analogies to existing past tenses stored in memory. On the other side of the debate are rule-based or dual-route approaches which agree that phonological analogy is important for producing irregular past tenses, but which also argue that regular inflection can *only* be explained in terms of symbolic processing.

Ramscar (in press) has shown that the one in principle objection *against* single-route accounts of

inflection – that homophone verbs based on nouns are processed on the basis of their grammatical origins, and not according to their phonological properties – is empirically unjustified: grammatical origin does not predict the past tense form of verbs, whereas phonology *and semantics* does. This paper has taken one of the strong claims *for* the dual-route theory of inflection – that the regular past tense rule is an informationally encapsulated module (see Fodor, 1983) – and subjected it to empirical scrutiny. Pinker and colleagues (e.g. Pinker, 1991, 1999, 2001; Clahsen, 1999; Kim, Pinker, Prince & Prasada, 1991) have claimed that the processing of regular inflection is driven by an innate mechanism that is unaffected by phonology, frequency or semantics. Results from the two experiments reported here fail to support this claim. Rather, they have shown conclusively that semantics does affect regular past tense comprehension, both of existing forms that may have been stored in memory, and of novel forms that needed to be interpreted on-line.

As Pinker (1999) observes, it is more than reasonable to assume that the same basic process (or processes) are responsible for both past tense production and comprehension. Ramscar (in press) showed that regular past tense production – in elicited inflection tasks – was apparently affected by semantics. The results reported here complement these findings, and extend them in that they provide an objective on-line measure of the effects of semantics on inflection (most previous studies of inflection have relied on subjective judgments and ratings to measure inflection processes, e.g. Ramscar, in press; Ullman, 1999; Prasada & Pinker, 1993; Kim et al, 1991). The results of this experiment show that – objectively – participants found regular past tense forms easier to process when the semantic contexts they were related to supported a regular form even though their subjective responses to regular forms were the same as when they were not supported by semantic context (i.e. in both cases, they considered the information carried by the regular forms to be true).

The pattern of results reported here is easily compatible with a model of inflection that assumes that past tense forms are computed (in both comprehension and production) by a process of comparison to previously stored forms, taking into consideration factors such as phonological and semantic similarity and frequency.

That these results are not compatible with the idea that regular inflection is processed independently from the contents of memory, and that it is entirely unaffected by factors such as phonological and semantic similarity and frequency (see Pinker, 1991, 1999, 2001) does not, of course, mean that the dual-route model is necessarily wrong (these results no more disprove the idea that *some* regular inflection is carried out in this context-independent manner than does the existence of still more white swans disprove the idea of orange swans). However, insofar as Ramscar (in press) has shown that one of the key reasons for positing a context-independent regular past tense rule (to deal with denominal verbs, which were supposedly regularized irrespective of their phonological and semantic properties) is unjustified, and insofar as the experiments reported here suggest that semantic and phonological comparisons in associative memory (a component that even the dual-route model accepts is necessary to model inflection) affect even the comprehension of novel inflected forms, it does seem worth considering what role it is that a context-independent rule is supposed play in a scientific account of inflection. There is an increasing body of evidence suggesting that a context-independent rule does not add anything substantive to our understanding of inflection (see e.g. Hahn & Nakisa, 2000; Ramscar, in press), and further, it appears that *any* inflection can be processed in associative memory (see Ramscar, in press and the experiments reported here) a component that even dual-route models accept is necessary to modeling inflectional morphology (see Pinker, 1991, 1999, 2001).

This evidence (and on a more mundane level, Occam's razor) militates against the inclusion of an explicit, context-independent rule in any psychological theory of inflection. At present, it appears that a similarity-based, single-route account of inflection – in which forms are processed by matching and analogous generalization according to factors such as phonological ands semantic similarity and frequency – provides a more economical explanation of, and a better fit to, the available data. To return to Rumelhart and McClelland's (1986) claim, it appears that children (and adults) may well *not* need to learn the rules of the English past tense in any explicit sense. As far as the English past tense system goes, it appears that the parser does *not* make "basically the same distinctions as the grammar" (Clahsen, 1999, p. 995). While the "grammar" of English may distinguish between irregular and regular past tense forms, these results suggest that the corresponding psychological processes that govern parsing do not make these explicit distinctions at all.

Acknowledgements

I am grateful to Lera Boroditsky, Will O'Connor, Malte Huebner and Daniel Yarlett for comments on, and discussions of, this work.

References

Anderson, S. R. (1992). A-Morphous Morphology, Cambridge: Cambridge University Press.

Berent, I., Pinker, S. and Shimron, J (1999) Default nominal inflection in Hebrew: evidence for mental variables. Cognition, 72, 1-44.

Bybee, J. L. (1988). Morphology as lexical organization. In M Hammond & M Noonan (eds.) Theoretical morphology, 119-141. San Diego: Academic Press.

Bybee, J. L. (1995) Regular morphology and the lexicon, Language and Cognitive Processes 10, 425-455.

Bybee, J. L. and D. I. Slobin. 1982. Rules and Schemas in the development and use of the English past tense. Language 58:265-289.

Bybee, J. L. and Moder, C. L. (1983). Morphological classes as natural categories. Language, 59, 251-270.

Clahsen, H. (1999). Lexical entries and rules of language: a multi-disciplinary study of German inflection. Behavioral and Brain Sciences 22 (6): 991-1013.

Fodor, J. A. (1983). The Modularity of Mind. Cambridge, MA: MIT Press.

Hahn, U & Nakisa, R. C. (2000). German inflection: Single-route or dual-route? Cognitive Psychology. 41(4), 313-360.

Harm, M.W. & Seidenberg, M.S. (1999) Phonology, reading acquisition, and dyslexia: Insights from connectionist models. Psychological Review, 106(3), 491-528.

Joanisse, M.F., Seidenberg, M.S. (1999). Impairments in verb morphology following brain injury: a connectionist model. Proceedings of the National Academy of Sciences, USA, 96(13), 7592-7597.

Kim, J. J., Pinker, S., Prince, A. and Prasada, S. (1991) Why no mere mortal has ever flown out to center field, Cognitive Science, 15, 173-218

MacWhinney, B., and Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. Cognition, 40, 121-157.

Pinker, S. (1991) Rules of language. Science, 253, 530-535

Pinker, S. (1999) Words and Rules, New York: Basic Books.

Pinker, S. (2001) Four decades of rules and associations, or whatever happened to the past tense debate? In E. Dupoux (Ed.), Language, the brain, and cognitive development: Papers in honor of Jacques Mehler. Cambridge, MA: MIT Press.

Pinker, S. and Prince, A., (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition, 28, 73--193

Plaut, D. C. and Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. Psychological Review, 107, 786-823.

Plunkett, K., and Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. Cognition, 48, 21-69.

Prasada, S. and S. Pinker (1993) Generalization of regular and irregular morphological patterns, Language and Cognitive Processes, 8, 1-56

Ramscar, M. J. A (in press). The role of meaning in inflection: Why the past tense doesn't require a rule. Cognitive Psychology.

Rumelhart D. E. and McClelland J. L (1986) On learning past tenses of English verbs. In Rumelhart D.E. and McClelland J.L (eds) Parallel Distributed Processing: Vol 2: Psychological and Biological Models. Cambridge, MA: MIT press

Ullman, M. T. (1999) Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. Language and Cognitive Processes, 14, 47-67

Inferring Unobserved Category Features With Causal Knowledge

Bob Rehder (**bob.rehder@nyu.edu**)
Department of Psychology, New York University, 6 Washington Place
New York, NY 10003 USA

Russell C. Burnett (**r-burnett@northwestern.edu**)
Department of Psychology, Northwestern University, 2029 Sheridan Road
Evanston, IL 60208 USA

Abstract

One central function of categories is to allow people to infer the presence of features that cannot be directly observed. Although the effect of observing past category members on such inferences has been considered, the effect of theoretical or causal knowledge about the category has not. We compared the effects of causal laws on feature prediction with the effects of the inter-feature correlations that are produced by those laws, and with the effect of exemplar typicality or similarity. Feature predictions were strongly influenced by causal knowledge. However, they were also influenced by similarity, in violation of normative behavior as defined by a Bayesian network view of causal reasoning. Finally, feature predictions were not influenced by the presence of correlations among features in observed category members, indicating that causal relations versus correlations lead to different inferences regarding the presence of unobserved features.

When an object has been classified as an instance of a concept, knowledge associated with that concept can be brought to bear in reasoning about the features that the object is likely to possess. But what is the nature of that knowledge, and how is it used to make inferences or predictions about unobserved features? Recent research has demonstrated that tasks such as category learning, categorization, and category-based induction are often influenced by the theoretical knowledge that one possesses. This knowledge often takes the form of causal relations between features of a category, and theories have been proposed to account for the effects of such knowledge (Rehder, 1999, 2001; Waldmann, Holyoak, & Fratianne, 1995). In this article we assess the effect of causal relations on feature inferences, and in the first of the following sections we present a formal model of causal knowledge and its predictions regarding feature inferences.

Of course, another form of knowledge that may guide feature inference is empirical information derived from the first-hand observation of category members. Prior research suggests two likely effects of such empirical knowledge on feature prediction. First, feature predictions will often be influenced by the overall similarity to the category of the exemplar with the unobserved feature. In the second section we discuss this predicted effect of similarity and show how it can run directly counter to the predictions of our formal model of causal knowledge. Second, the presence of correlations among category features may also allow one to infer the presence of a feature given knowledge about the presence of one or more other features. We discuss the effects of observed inter-feature correlations in the third section, and compare them to the effects produced by direct knowledge of causal relations—relations that were responsible for generating the feature correlations in the first place.

Feature Inference via Causal Reasoning

It is clear that causal knowledge has predictive value. For example, given knowledge of the causes of fire, one can predict, with some certainty, that a flame will appear when a match is struck, oxygen is present, and so on. Likewise, given the causal relations that hold among features of an object, the presence of an unobserved feature can be inferred by reasoning about the causes of that feature and whether those causes are present in the object at hand.

In this article we provide direct evidence of causal reasoning in feature inference, and we test a well specified theory about how this sort of reasoning might be done. This theory involves Bayesian networks—graphs in which variables are represented as nodes, and causal relations between the variables as directed links between the nodes. Figure 1 shows a simple Bayesian network in which three effect variables are dependent on a single cause variable.

Rules by which inferences can be drawn from Bayesian networks have been well developed in artificial intelligence. One important rule is the *causal Markov condition*, which states that a variable X is independent of all variables that are not themselves descendents of X given knowledge about the state of X's (immediate) parents (Pearl, 1988). In Figure 1, for example, the state of F_2 is independent of F_3 and F_4 given knowledge about F_1.

It has been proposed that Bayesian networks are good psychological models of causal knowledge—and, in particular, of the causal knowledge associated with object concepts

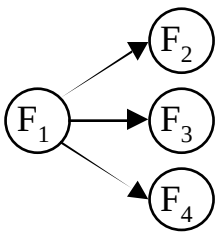


Figure 1. A common-cause causal schema.

(Rehder, 1999, 2001; Waldmann et al., 1995). On this *causal-model theory* of concepts, the model shown in Figure 1 can be used to represent a concept with four features, where feature F_1 causes features F_2, F_3, and F_4. Representing causal knowledge of category features in this way has been shown to account well for classification (Rehder, 1999, 2001), but it is an open question whether the rules of inference associated with Bayesian networks—and, in particular, the causal Markov condition—accurately describe people's inferences about unobserved features.

In this article we tackle this question by explicitly manipulating the causal knowledge about a category that a participant has and measuring the effects of that knowledge on subsequent feature predictions. Participants in a *Causal Schema* condition were told that the features of a novel category were related as in the common-cause model of Figure 1, and were then asked to make inferences about exemplars in which one of the four features was unobservable. We determined whether participants' responses were consistent with the causal Markov condition.

Feature Inference via Category Similarity

Another factor likely to influence the prediction of an unobserved feature is the similarity of the exemplar to previously observed category members (or to the category's prototype). In this case, inference is based on simple feature overlap. If an exemplar is similar to (i.e., shares many features with) many category members, and if many category members possess the unobserved feature, then the exemplar probably has the feature too. For example, if a new bird has many features typical of birds (small, sings, eats worms, etc.) it probably also flies, because it is similar to many birds and most birds fly. In contrast, a new bird with many atypical features (e.g., an ostrich) is similar to fewer birds, and so the inference to flight is less certain.

Previous research has shown that similarity plays a key role in a variety of feature inference tasks. For example, Sloman's (1993) feature-based model of the inductive projection of features across categories assumes that a feature is projected from, say, robins to falcons by computing the extent to which they have other features in common (cf. Rips, 1975). Sloman also found a phenomenon called "inclusion similarity" in which participants projected a property from an inclusive category to a subordinate (e.g., from bird to robin) more strongly when the subordinate was more typical (e.g., robin) than when it was less so (e.g., penguin). Direct evidence of the role of similarity in feature prediction (rather than projecting new features across categories) was provided by Yamauchi and Markman (2000), who taught participants artificial categories and found that exemplars that were closer to the category prototype (i.e., that possessed more features in common with training exemplars) supported stronger inferences of unobserved features.

The influence of similarity on feature inference presents a particularly stringent test of the causal Markov condition, because honoring the causal Markov condition can require one to ignore similarity. For a category with a common-cause causal schema (Figure 1), the causal Markov condition states that information about the presence or absence of F_2 and F_3 is irrelevant to inferring F_4 given knowledge of F_1. In contrast, an influence of similarity predicts that inferences to F_4 will be stronger when F_2 and F_3 are present, because the presence of F_2 and F_3 means that the exemplar is more similar to the category prototype.

In the following experiment participants in the *Control* condition were told that each feature had a 75% base rate (as were Causal Schema participants), but were not instructed on any causal relationships. Results from the Control condition will indicate an effect of similarity if feature inferences increase as a function of the exemplar's similarity to this central tendency (i.e., as a function of the number of features). Results from the Causal Schema condition will indicate whether participants are able to override this effect of similarity, as required by the causal Markov condition.

Feature Inference via Feature Correlations

The final influence on feature prediction performance we consider is the presence of within-category feature correlations. For example, many people know that birds that are small tend to sing whereas large birds do not, and on the basis of this correlation might infer the presence of a small bird upon hearing song, or the absence of singing from a large bird—and do so despite having no knowledge of the causal mechanisms that link size and singing.

Prior research confirms the intuition that the observation of within-category feature correlations can influence feature predictions, at least when participants observe category exemplars during standard classification-with-feedback training. Some studies (Thomas, 1998; Yamauchi & Markman 1998) have attributed this result to participants' similarity-matching to the training exemplars with a multiplicative similarity rule that preserves sensitivity to feature correlations. Others (Anderson & Fincham, 1996) attribute it to participants' inducing a direct representation of those inter-feature correlations (also see Wattenmaker, 1993).

A final goal of the current article was to compare the effect of causal laws on feature inference with the effect of observing the inter-feature correlations produced by those laws. In the following experiment, participants in the *Exemplars* condition were told that each feature manifested a 75% base rate, as were participants in the Causal Schema and Control conditions. But then, rather than being instructed on causal relationships, they instead observed a sample of exemplars that manifested the inter-feature correlational structure that is implied by a common-cause causal schema (i.e., exemplars with strong correlations between feature F_1 and features F_2, F_3, and F_4). Because it reflects causal laws, feature prediction performance based on this correlational structure should ideally be qualitatively similar to performance based on knowledge of the laws alone. In particular, inferences regarding the presence of an unobserved effect feature should be stronger when F_1 is present as compared to one of the other features.

Method

Materials

Six novel categories were used: two biological kinds (Kehoe Ants, Lake Victoria Shrimp), two nonliving natural kinds

(Myastars, Meteoric Sodium Carbonate), and two artifacts (Romanian Rogos [cars], Neptune Personal Computers). Each category had four binary features. For example, for the Lake Victoria Shrimp category the four binary features were "a high quantity of ACh neurotransmitter," "long-lasting flight response," "accelerated sleep cycle," and "high body weight." Each feature was described as occurring in 75% of category members. Participants in the Causal Schema condition were also taught about three causal relationships between F_1 and F_2, F_3, and F_4. Each description of a causal relationship specified the cause feature, the effect feature, and a brief description of causal mechanism linking them. For example, the $F_1 \rightarrow F_2$ causal relationship for Lake Victoria Shrimp was "A high quantity of ACh neurotransmitter causes a long-lasting flight response. The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter."

Participants

Fifty-four undergraduates or other members of the Northwestern University community received course credit or pay for participating in this experiment.

Design

Participants were randomly assigned in equal numbers to one of the six categories, and to either the Causal Schema, the Exemplars, or the Control condition.

Procedure

All phases of the experiment were conducted by computer. Participants first studied several screens of information about the assigned category at their own pace. All participants read a cover story and a description of the features and their 75% base rates. Participants in the Causal Schema condition also received a description of three causal relationships, and a diagram depicting those relationships similar to Figure 1. When ready, all participants took a multiple-choice test of this knowledge. Participants could request help, which led the computer to re-present the information about the category. Participants were required to retake the test until they made 0 errors and 0 requests for help.

Participants in the Exemplars condition then observed 48 examples of the category. Although the studies reviewed above found feature prediction performance to be sensitive to feature correlations when training exemplars were observed in a classification-with-feedback task, Wattenmaker (1991) found that participants were more sensitive to feature correlations on a transfer categorization test when they were asked simply to "look over, examine, and learn about" exemplars. Thus, category exemplars were presented sequentially at a pace determined by the participants. They observed 26, 3, 3, 3, 1, 1, 1, 2, 2, 2, and 4 instances of exemplars 1111, 1110, 1101, 1011, 0110, 0101, 0011, 0100, 0010, 0001 and 0000, respectively, where "1" denotes the presence of a feature, "0" represents its absence, and features are given in dimension order (F_1, F_2, F_3, F_4). These exemplars manifest the 75% feature base rates that participants were instructed on, and also the correlational structure that is implied by a common-cause causal schema. Specifically, the strength of the correlations between F_1 and F_2, F_3, and F_4 was $r = .62$, and the correlations among F_2, F_3, and F_4 conditional on F_1 were approximately 0. The features of each exemplar were listed in order (1–4) on the computer screen. For example, participants assigned to the Lake Victoria Shrimp category were presented with three category members that possessed "high amounts of the ACh neurotransmitter," "a normal flight response," "accelerated sleep cycle," and "high body weight." The order of the 48 exemplars was randomized for each participant.

Participants in all conditions then performed two tasks (counterbalanced for order): a feature prediction task and a categorization task. During the feature prediction task, participants were presented with 32 exemplars, each with an unobserved value on one of the four dimensions, and were asked to rate the likelihood that the feature was present on a 100-point scale. For each unobserved dimension the other three dimensions took on the eight possible combinations of values, yielding a total of 32 feature prediction problems. The features of each exemplar were listed in dimension order (1–4), with the unknown dimension designated with "???." For example, participants assigned to the Lake Victoria Shrimp category were presented with the feature list "normal amounts of the ACh neurotransmitter," "a fast flight response," "???," and "high body weight" and asked to rate on a 100-point scale whether this exemplar had an "accelerated sleep cycle." The order of the 32 feature prediction problems was randomized for each participant.

During the categorization task, participants rated the category membership of exemplars on a 100-point scale. There were 32 exemplars, consisting of all possible 16 examples that could be formed from four binary features, each presented twice. The order of the 32 test exemplars was randomized for each participant.

Results

Feature Prediction Results

Because results for those feature prediction problems in which the unobserved feature dimension was the first dimension were not directly relevant to the theoretical issues raised in this article, we report results only for those problems in which the unobserved feature was on the second, third, or fourth dimension. Figure 2 presents feature prediction ratings as a function of the total number of features in the exemplar, whether the common-cause feature (F_1) is present or absent in that exemplar, and experimental condition (Causal Schema, Control, or Exemplar). Note that the number of features in F_1-Present problems ranges from 1 to 3 whereas the number in F_1-Absent problems ranges from 0 to 2 because of the presence of F_1 itself in the F_1-Present problems.

In the Causal Schema condition (Figure 2a) feature prediction ratings were strongly influenced by the presence or absence of the common cause F_1 as compared to one of the other features. For example, problems with one feature received a much higher rating when that feature was F_1 (e.g., the feature prediction problem 1x00) than when it was one of the other features (e.g., 0x10), 70.6 versus 24.5. Similarly, problems with two features received a higher rating when one of those features was F_1 (e.g., 1x10) than when

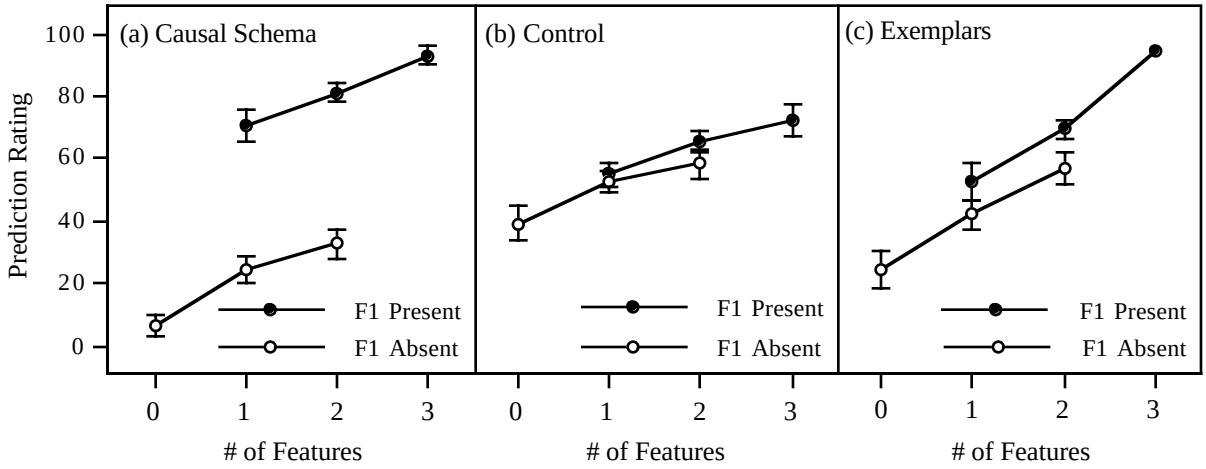


Figure 2. Feature prediction results in the (a) Causal Schema, (b) Control, and (c) Exemplars conditions

neither was F_1 (e.g., 0x11), 80.9 versus 32.8. That is, Causal Schema participants were much more likely to reason from the presence (or absence) of the common-cause feature F_1 to infer the presence (or absence) of an effect than when they were reasoning from one of the effect features.

According to the causal Markov condition, inferring the presence of an effect in a common-cause schema should not only depend on the common cause feature F_1, it should also *not* depend on any of the effect features. In fact, Figure 2a indicates that feature prediction ratings increased as the number of effect features increased. In the F_1-absent condition, feature prediction ratings were 6.7, 24.5, and 32.9 for exemplars possessing 0, 1, and 2 effect features, respectively. This occurred despite the fact that, according to the causal Markov condition, the absence of common cause F_1 makes the presence or absence of other effects irrelevant for predicting an effect feature. Likewise, in the F_1-present condition, ratings were 70.5, 80.9, and 92.8 for exemplars possessing 1, 2, and 3 features, respectively. That is, although Causal Schema participants' feature prediction ratings were strongly influenced by the causal knowledge that was provided, they also exhibited a substantial similarity effect in which more features led to stronger inferences, in violation of the causal Markov condition.

In comparison with the Causal Schema condition, results from the Control condition (Figure 2b) indicate that the effect of the presence or absence of F_1 on feature prediction ratings was not greater than the effects of the other features. This result was expected, because in the Control condition there was nothing about F_1 to make it especially predictive of an unobserved feature. However, as in the Causal Schema condition, feature prediction ratings exhibited an effect of similarity; ratings increased as a function of the number of features present in the exemplar. Ratings were 37.4, 53.6, 63.1, and 73.7 for exemplars that possessed 0, 1, 2, or 3 features, respectively. Note that this effect of similarity obtained despite the fact that the Control participants observed no members of the category, but rather just read a verbal statement of the 75% feature base rates.

These conclusions were supported by statistical analysis. Each participant's ratings were predicted from a regression equation in which the two predictors were the number of features present and a contrast code representing the presence or absence of F_1. As expected, in the Causal Schema condition the regression weight associated with the presence or absence of F_1 was both significantly greater than zero, $t(35) = 6.95$, $p < .0001$, and significantly different than the corresponding weight in the Control condition, $t(34) = 5.65$, $p < .0001$. Moreover, in both the Causal Schema condition and the Control condition the regression weight associated with the number of features was significantly different from zero, $t(35) = 4.89$, $p < .0001$, and $t(35) = 2.79$, $p < .01$, respectively. This sensitivity to number of features did not differ between the Causal Schema and Control conditions, $t < 1$.

Finally, Figure 2c presents the results from the Exemplars condition. The figure indicates that, in contrast to the Causal Schema condition, the presence of F_1 resulted in only a small increase in feature prediction ratings as compared to one of the other features. For example, whereas in the Causal Schema condition problems received a feature prediction rating that was about 50 points higher when F_1 was present in the exemplar (Figure 2a), in the Exemplars condition that difference was only 10.9 points (52.9 vs. 42.0) for one-feature exemplars and 12.4 points (69.3 vs. 56.9) for two-feature exemplars. This result obtained despite the presence of strong correlations between F_1 and the other features in the training exemplars that might have been expected to lead subjects to treat F_1 as especially predictive. In fact, the regression weight associated with the presence or absence of F_1 in the Exemplars condition was not significantly different from the Control condition, $t(34) = 1.17$, $p > .20$.

As in the Causal Schema and Control conditions, feature prediction ratings in the Exemplars condition were sensitive to the total number of features possessed by the exemplar. Ratings were 24.2, 44.2, 65.1, and 94.0 for exemplars that possessed 0, 1, 2, or 3 features, respectively. The regression weight associated with the number of features was greater than the corresponding regression weight in the Control condition, $t(34) = 2.12$, $p < .05$. In other words, the observation of exemplars that manifested the 75% feature base rates led participants to be more sensitive to similarity as compared to the Control condition, in which participants

were simply told about the 75% feature base rates.

Individual differences. On a finer level of analysis, there is some clustering in the data. Informally, the response patterns given by participants in the Causal Schema condition fell into a few different classes, and the most frequent was in fact the one that respects the causal Markov condition (uniformly low ratings when F_1 is absent, uniformly high ratings when F_1 is present). A look at the Causal Schema subjects' regression weights (see Figure 3) revealed a group of 8 "causal Markov" subjects who weighted the presence or absence of F_1 heavily and the number of features lightly, 3 "similarity" subjects who weighted the number of features heavily and F_1 lightly, and 7 "compromisers" who assigned moderate weights to both predictors. In contrast, examination of the Exemplars condition revealed 2 causal Markov subjects, 11 similarity subjects, and 2 compromisers (3 subjects weighted neither factor). That is, whereas the modal response in the Exemplars condition was similarity based, the modal response in the Causal Schema condition was consistent with the causal Markov condition.

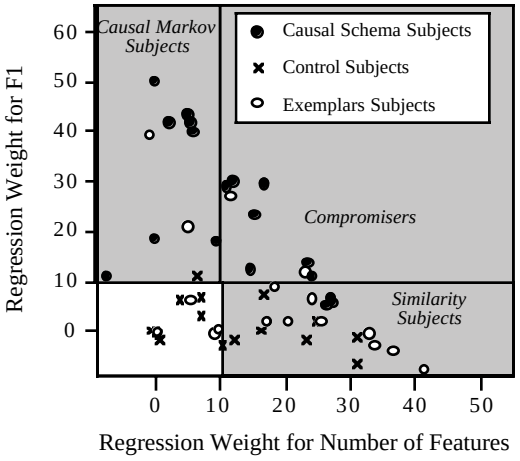


Figure 3. Regressions weights for individual subjects.

Categorization Results

One purpose of the categorization task was to ask to what extent feature inference was mediated by the goodness of category membership of the exemplar with the unobserved feature. We summarize the main findings. First, in the Causal Schema condition categorization ratings were influenced by the two factors that influenced feature predictions: They increased as the number of features in the exemplar increased (i.e., as the exemplar's similarity increased), and they increased even more when the causally central F_1 was present (i.e., the pattern shown for feature predictions in Figure 2a). However, unlike the prediction ratings, the categorization ratings were also sensitive to whether the causal relation between F_1 and each of the other observed features was confirmed or violated (i.e., whether F_1 and each of the effect features were jointly present/absent or not), a finding replicating past results (Rehder & Hastie, 2001). In other words, whereas categorization ratings showed a sensitivity to all three of the relations that constitute a common-cause schema, for purposes of predicting unobserved features participants apparently attended to only the relation in which the unobserved feature dimension was involved. The result was a dissociation between feature prediction and categorization ratings in the Causal Schema condition.

Second, in the Exemplars condition category membership ratings, like the feature prediction ratings, were sensitive to similarity but insensitive to the presence or absence of F_1 as compared to other features (cf. Figure 2c). However, unlike the feature prediction ratings the categorization ratings were also sensitive to whether the correlations between F_1 and each of the other observed features were broken or preserved.

Importantly, this latter result speaks to the possibility that Exemplars participants' insensitivity to the presence or absence of F_1 during the feature prediction task arose merely as consequence of their failing to learn and encode the correlations involving F_1. In fact, results from the categorization task indicate that participants encoded these correlations but did not make use of them in feature inference. This represents a dissociation between feature inference and categorization, just as in the Causal Schema condition.

Finally, in the Control condition both category membership and feature prediction ratings were monotonic functions of the number of features present, that is, the featural similarity of the exemplar to the category prototype.

General Discussion

The first question asked in the current article was whether causal knowledge about a category influences predictions regarding the presence or absence of unobserved features. In fact, we found that causal knowledge had a strong effect on those inferences. Reasoners were much more likely to predict the presence of an unobserved feature when its cause was present than when that cause was absent. In this respect, their reasoning was similar to the normative method of inference defined by Bayesian networks.

These results contribute to a collection of findings demonstrating the importance of theoretical or explanatory knowledge in variety of feature inference tasks. For example, Lassaline (1996) found that the projection of a new property from one category to another was stronger when causal knowledge supportive of that property was provided (also see Sloman, 1994). Rehder & Ross (2001) found that the learning of a category via a feature prediction task proceeded more rapidly when features were related on the basis of prior knowledge. However, so far as we know, the current study is the first to address the specific role of causal knowledge in inferring the presence of unobserved features in a category with known causal structure.

Although causal knowledge had a profound effect on feature predictions, we also found that normative Bayesian reasoning is not the whole story. Even when reasoners had causal knowledge, their feature inferences showed a persistent effect of overall similarity to the category, such that an exemplar with a greater number of category-associated features was deemed more likely to have an unobserved feature. This was true even though the features that contributed to similarity were conditionally independent of the unobserved feature in question. In this respect, the effect of causal knowledge on feature predictions violated the causal Markov condition associated with Bayesian networks.

The influence of similarity on feature predictions also ob-

tained in the Control condition. This effect held even though participants did not observe category members (in contrast to previous studies demonstrating similarity effects, e.g., Yamauchi & Markman, 2000), but rather were provided only with a verbal statement of the 75% feature base rates. Under these conditions one might have expected that Control participants would be especially likely to assume independence among features (i.e., base each prediction only on the base rate of the feature in question and not on the presence/absence of other features). The current results indicate people's tendency to reason on the basis of central tendency or prototype information holds even when that information is provided in summary form (and without any mention of correlations between features) rather than experientially.

On the one hand, these findings are reminiscent of other studies that have attempted—mostly in vain—to induce participants to ignore the effects of similarity (e.g., Allen & Brooks, 1991). However, our analysis of individual differences revealed considerable variation among participants in the relative importance of causal knowledge and similarity. In fact, 8 of 18 participants in the Causal Schema condition ignored similarity (that is, they honored the causal Markov condition) when predicting unobserved features, indicating that similarity-based responding is not obligatory. The question of under what conditions feature inferences are dominated by theoretical and causal knowledge versus featural similarity is one that merits further investigation.

Another question we asked was whether causal knowledge has a different effect than inter-feature correlational knowledge. Rather than being given an explicit causal model, participants in the Exemplars condition observed exemplars that manifested the correlations implied by that model. In fact, though, the vast majority (14 of 18) of Exemplars participants failed to use those correlations in making feature inferences, and their inferences were qualitatively like those of Control subjects, who had neither causal nor empirical knowledge. Instead, the effect of the empirical observations was merely to make feature predictions even more sensitive to the degree to which an exemplar was similar to the category's prototype.

This finding contrasts with previous studies in which feature prediction was found to be sensitive to inter-feature correlations, at least when those correlations were observed during a classification-with-feedback task (Anderson & Fincham, 1996; Thomas, 1998; Yamauchi & Markman, 1998). We used a different learning task, in which participants were asked merely to observe category members, but we know that the inter-feature correlations were learned and encoded during this task because they were reflected in participants' categorization ratings. That the same participants failed to use these correlations in feature inference represents a dissociation between categorization and feature inference.

More generally, in both the Causal Schema and the Exemplars conditions, we found that categorization ratings and feature inferences were sensitive to different kinds of information: Categorization but not feature prediction was sensitive to the overall causal or correlational structure instantiated in an exemplar. This implies that feature inference is not merely mediated by goodness of category membership. Instead, participants in both of these conditions used the category knowledge they possessed in different ways depending on the task at hand. That is, whereas Yamauchi and Markman (1998) have suggested that category representations will differ depending on whether they are acquired via categorization or feature prediction, the current results suggest that categorization and feature prediction tasks can also draw on different aspects of a single representation.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of the explicit rule. *Journal of Experimental Psychology: General, 120*, 3–19.

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 259–277.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 754–770.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufman.

Rehder, B. (1999). A causal model theory of categorization. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 595–600). Vancouver, British Columbia.

Rehder, B. (2001). *A causal-model theory of conceptual representation and categorization*. Submitted for publication.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323–360.

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1261–1275.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665–681.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 231–280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition, 52*, 1–21.

Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 119–143.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181–206.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 908–923.

Wattenmaker, W. D. (1993). Incidental concept learning, feature frequency, and correlated properties. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 203–222.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 776–795.

Routine Problem Solving in Groups

Torsten Reimer (t.reimer@unibas.ch)
Klaus Opwis (klaus.opwis@unibas.ch)
Anne-Louise Bornstein (anne-louise.bornstein@unibas.ch)

Department of Psychology, University of Basel,
Bernoullistr. 16, 4056 Basel, Switzerland

Abstract

Routines may help groups to effectively reduce coordination requirements when solving interdependent tasks. However, routine problem solving always involves the risk of a negative transfer, which appears if a routine is applied to novel problems even though it is inappropriate. In this experiment, negative transfer was produced by first teaching individuals a procedure for solving the Tower of Hanoi problem. Next, participants were asked to solve several transfer tasks either individually or in pairs. However, the routine could not be applied directly to the transfer tasks but led to a long detour. As expected, the individuals surpassed the dyads, who insisted more strongly on their routine. This result fits with studies that corroborate the claim that groups are prone to a "principle of inertia" when solving problems or making decisions.

Introduction

A *routine* may be defined as a well practiced problem solving procedure, which has been applied repeatedly and, therefore, does not need much planning but may be executed rather automatically (for an overview on various definitions, see Betsch, Haberstroh, & Hoehle, in press). Typically, in the problem-solving domain, routines consist of several single action steps that have to be executed in a particular order. Even though the single steps may require some planning, the sequence of the steps itself is usually highly internalized. From this automation follows that procedures or schemata that have become a routine may easily be transferred to novel tasks. However, routines are not only transferred to structurally equivalent tasks (positive transfer), but are sometimes also applied to tasks that share only surface features with the learning task. Accordingly, there is ample evidence showing that the successful use of a scheme enhances the likelihood of a *negative transfer effect*, i.e., worse performance compared to a condition in which the scheme has not been repeatedly applied before (cf. VanLehn, 1996).

In the present study, we sought to extend the literature on learning transfer to group problem solving. In particular, the study aimed at testing whether negative transfer effects are more pronounced in dyads than in individuals. On the one hand, it may be more likely that a dyad recognizes a change in task demands. On the other hand, adapting a routine in a group usually requires coordination processes, which may cause process losses (cf. Gersick & Hackman, 1990). Moreover, there are several studies showing that groups often tend to accentuate preferences or decisions that are held by a majority (cf. Hinsz, Tindale, & Vollrath, 1997). In general, if individuals are predisposed to process information in a biased way, then, groups usually tend to enhance this bias. However, if groups use strategies more reliably and consistently than individuals, then, transfer effects should also be enhanced in groups, irrespective of whether this transfer is positive or negative.

In this study, the hypothesis that negative transfer is more pronounced in groups than in individuals was tested by asking participants to solve Tower of Hanoi problems either individually or in pairs. There are several procedures that guarantee an optimal solution of the Tower of Hanoi problem (cf. Simon, 1975). In order to produce transfer effects, participants were first taught either the goal-recursion (R) or the move-pattern procedure (M). These two procedures differ in two aspects, which make them suitable for studying transfer effects within the Tower of Hanoi problem:

(1) Only the R-procedure but not the M-procedure may be directly applied to a transfer task in which the start peg is the middle peg. (2) The two procedures lead to different patterns of move latencies. Hence, differences in move latencies may be used as an indicator of the respective procedure applied and to what extent a problem solver insisted on his/her routine.

Whereas the first study tested for process losses in pairs, the second study compared homogeneous with heterogeneous pairs and additionally considered individual learning.

The Tower of Hanoi Problem

The Tower of Hanoi problem consists of three pegs and a fixed number of disks of different sizes (Simon, 1975). The original task is to move all the disks from the left to the right peg under the following constraints

(cf. Figure 1): (a) only one disk may be moved at a time, (b) only the disk that is on the top of the pyramid may be moved, and (c) a larger disk may never be placed on top of a smaller disk. A problem with *n* disks requires a minimum of 2^n–1 moves.

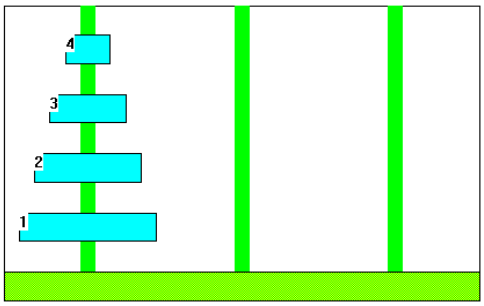


Figure 1: The Tower of Hanoi problem.

The goal-recursion procedure The *goal-recursion procedure* (R) consists in forming sub-goals. Each problem with *n >1* disks can be decomposed into three sub-problems: Into (1) a problem consisting of *n–1* disks, (2) the move of the largest disk from the start peg onto the goal peg, and (3) into a second problem consisting of *n–1* disks.

For example, the four-disk problem in Figure 1 can be decomposed into two three-disk problems and the move of a single disk: (1) First of all, the three-disk pyramid (consisting of the disks 2 to 4) has to be moved from the start peg onto the middle peg; (2) in the next step, the largest disk (number 1) can be moved onto the goal peg; and (3) finally, the three-disk pyramid has to be moved again, this time from the middle peg onto the goal peg. The three-disk problem is itself a Tower of Hanoi problem with one disk less than the original problem. Hence, the three-disk problem can be decomposed into two two-disk problems and the move of a single disk. Thus, the recursion procedure is based on a chunking strategy by dividing a problem into sub-problems until only one disk remains and there is no longer any problem.

The move-pattern procedure The *move-pattern procedure* (M), on the other hand, is based on a stimulus-driven instead of a goal-driven heuristic, without formulating any sub-goals. According to this procedure, one has to learn a particular pattern of moves, whereby attention has to be paid to the *position* of a disk and to its *parity*: *Odd*-numbered disks should always be moved from the left to the right, from the middle to the left, and from the right to the middle. *Even*-numbered disks are moved the other way round, from the left to the middle, from the middle to the right, and from the right to the left. Additionally, the same disk should never be moved twice in one go. The latter constraint guarantees that it is always clear which disk is to be moved next. For example, according to these rules, disk 4 in Figure 1 should be moved onto the middle peg because it is an even-numbered disk. Next, disk number 3 should be moved onto the right peg, because the same disk should never be moved twice in one go and *3* is an odd number.

Original vs. transfer tasks It is worth noticing that both procedures lead to the same pattern of moves and to an optimal solution, provided all rules are strictly adhered to. This functional equivalence holds for any number of disks. However, if the start peg is the *middle* peg instead of the left peg, only the recursion-procedure is optimal. The move-pattern procedure may be applied to such a *transfer task* as well, but it requires twice as many moves as the optimal goal-recursion procedure: If a transfer task is solved by applying the move-pattern procedure, the entire tower moves first onto the left peg and then onto the right peg.[1]

Differences in move-latencies Even though both procedures lead to the same pattern of moves with original tasks, they differ in the amount of planning and, therefore, result in different patterns of *move latencies* (cf. Reimer, 2001a). If the M-procedure is applied, the cognitive effort is almost the same for all moves (despite the fact that the decision as to which disk should be moved next may vary among different game situations). Thus, a player who applies the M-procedure is expected to move disks relatively regularly. According to the R-procedure though, at the very beginning as well as in situations, in which a new sub-tower has to be solved, extensive planning is required. Whereas these "first moves" should last long, subsequent moves (i.e., all other moves) should be carried out fast in order to execute the planned recursion smoothly without extensive interruptions. Hence, ideally, an application of the R-procedure results in high latencies in the first moves and short latencies when subsequent moves are performed.

In general, the extent to which a player spends more time on first moves than on subsequent moves may be quantified by the following strategy index (S):

$$S = FM / SM,$$

with FM = mean time for first moves and SM = mean time for subsequent moves. Because the M- and R-procedure differ in the extent of chunking and planning, participants who were taught the R-procedure were expected to score higher on the strategy index than participants who had been taught the M-procedure (S_R

[1] In order to meet this criterion, the original move-pattern procedure, which was described by Simon (1975), was slightly changed by linking the move patterns to the left, middle, and right peg instead of the start and the goal peg.

> S_M). Additionally, the strategy index may also serve as a measure of the extent to which participants in the M-condition change their strategy towards a chunking-strategy when solving transfer tasks.

Study 1

The first study aimed at (1) testing for differences in performance between pairs and individuals, and, in particular, to test whether the M-pairs suffer from any process losses when solving transfer tasks (*process-loss hypothesis*); (2) testing the claim that the M-pairs insist more strongly on their procedure than the M-individuals, which should result in a lower strategy index for the M-pairs than the M-individuals (*persistency hypothesis*); and (3) testing to what extent the potential process losses are mediated by the strategy index (*mediation hypothesis*).

Method

Sample and design The design consisted of three factors: Firstly, participants were individually taught either the R- or the M-procedure (factor *procedure*). Secondly, participants solved problems either individually (I) or in pairs (P) (factor *group*). Finally, type of task was varied as a within-subjects factor. Each individual and each pair had to solve two original and two transfer tasks, one four- and one five-disk problem each.[2] The 90 students who participated in the study were randomly distributed among experimental conditions (15 pairs in the conditions R-P and M-P and 15 individuals in the conditions R-I and M-I).

Procedure Each participant was first individually explained the R- or the M-procedure. Additionally, each person completed a computerized training run with 30 tasks that always required only a single move. In the *R-condition*, their task was to solve sub-problems. For this purpose, one or more disks were already marked on the computer screen. The respondent's task consisted in moving the respective sub-problem in the correct direction. In the *M-condition*, respondents were also confronted with different game situations. Here, the task consisted in executing the next move according to the M-procedure. In both conditions, immediate feedback was provided by the computer on whether the single move was correct or wrong.

In the testing phase, participants were asked to solve two original and two transfer tasks in as few moves as possible. In the pair condition, they moved in turns without communicating with each other. The opportunity to correct or to undo moves by moving a disk twice in one go was explicitly mentioned in the instructions. If a person tried to place a larger disk on top of a smaller one, an error message appeared on the screen.

Results

Number of moves Table 1 shows the mean number of moves that were required by the individuals and pairs. An ANOVA with the factors, *procedure* (R vs. M), *group* (I vs. P), and *task* (original vs. transfer tasks), revealed an interaction of *task x procedure*, which was due to negative transfer in the M-condition ($F(1,56)_{\text{task x procedure}}$= 115.51; p<.01; $F(1,56)_{\text{procedure}}$= 94.32; p<.01; $F(1,56)_{\text{group}}$= 0.13; ns; $F(1,56)_{\text{task}}$= 121.43; p<.01). In the R-condition, the original as well as the transfer tasks were solved almost perfectly ($M_{\text{original task}}$= 24.15; $M_{\text{transfer task}}$ = 24.38; $t(29)$= –.50; ns).

Table 1: Mean Number of Moves.

	R		M	
	I	P	I	P
Original tasks	24.83	23.47	25.87	23.50
Transfer tasks	25.20	23.57	39.97	46.77

However, participants who had been taught the move-pattern procedure required many more moves to solve the transfer tasks than the original tasks ($M_{\text{original task}}$= 24.68; $M_{\text{transfer task}}$= 43.37; $t(29)$= –10.2; p<.01). Moreover, as can be seen in Table 1, this negative transfer was enhanced in the group of the M-pairs. Accordingly, the two-way interaction was further qualified by a significant three-way interaction of *task* x *procedure* x *pair*, F (1,56)= 7.55; p<.05. Obviously, the M-pairs suffered much more from their routine than the M-individuals when solving transfer tasks ($t(28)$= 2.17; p<.05).

Move latencies Are these process losses caused by a higher persistency of the M-pairs? First, as expected, participants in the R-condition had much higher strategy indices than participants in the M-condition throughout (cf. Table 2).[3] This holds true for original tasks ($t(58)$= 8.3; p<.01) as well as for transfer tasks ($t(58)$= 4.28; p<.01).

Additional ANOVAs, which were conducted separately for the M- and the R-condition, confirmed the persistency hypothesis:

[2] For the following analyses, measures were aggregated across the four- and five-disk problems throughout. Problems with four disks are solvable in 15 moves and problems with five disks require 31 moves. Thus, the minimum number of moves is 23, irrespective of the task conditions, i.e., regardless of whether original or transfer tasks are solved.

[3] The distribution of latencies was positively skewed. For this reason, each latency per move was transformed first by taking the logarithm. All reported analyses are based on these transformed latencies.

Table 2: Mean Strategy Index.

	R		M	
	I	P	I	P
Original tasks	1.16	1.08	1.04	1.02
Transfer tasks	1.18	1.09	1.10	1.02

In both analyses, the main effect of *group* ($F_R(1,28)$ = 55.61; $F_M(1,28)$ = 12.88; ps<.01) was significant, indicating higher strategy indices for the individuals than for the pairs. These main effects may be explained by the time that is required by the pairs when taking turns. Secondly, there was also a significant main effect of *type of task* in both conditions: Overall, participants in the R- ($F_R(1,28)$ = 6.27; p<.05) as well as in the M-condition ($F_M(1,28)$ = 6.94; p<.05) showed higher strategy indices when solving transfer tasks compared with original tasks. However, only in the M-condition the two main effects were qualified by a significant interaction ($F_M(1,28)$ = 5.16; p<.05; $F_R(1,28)$ = 0.70; ns). As can be seen in Table 2, the M-individuals had a much higher strategy index in the transfer tasks than in the original tasks ($t(14)$= 2.70; p<.05), whereas the M-pairs did not change the way in which they structured the problem solving process ($t(14)$= 0.44; ns).

Predicting performance by the strategy index The observed persistency may also serve as an explanation for the differences in performance. In order to show that the observed process losses are due to the extent to which the M-individuals and M-pairs changed towards a chunking strategy, an ANCOVA on the number of moves was run using the strategy index as a covariate. As can be seen in Figure 2, which refers exclusively to the M-condition and to the transfer tasks, the observed process losses disappear if differences in the strategy index are controlled.

Discussion

Participants who had been taught the goal-recursion procedure did not have any serious problems solving the transfer tasks (positive transfer). Within the move-pattern condition, however, a negative transfer effect appeared. Further, the M-pairs performed worse on the transfer tasks than the M-individuals, which confirms the process-loss hypothesis. However, these process losses do not seem to be due to mere mutual distraction in the pairs, that is, it is unlikely that participants distracted each other in general when joining a dyad. If this were the case similar process losses should have been observed in the other pair conditions, too. Rather, the results confirm the persistency hypothesis: Obviously, the M-pairs did not only need many more moves to solve the transfer tasks but also insisted more strongly on using their strategy than the M-individuals, who reacted much more flexibly and tried to adopt a chunking strategy.[4]

In general, participants in the R-condition spent more time on the first than on subsequent moves, whereas participants in the move-pattern condition made their moves much more regularly. Thus, in the original tasks, the participants in the R- and M-condition did not differ in their performance, but could easily be identified on the basis of their strategy index. Moreover, differences in performance disappeared when differences in the strategy index were controlled (mediation hypothesis).

Thus, the first study confirms the assumption that negative transfer effects will be enhanced by dyads who were taught the same inappropriate routine and, therefore, share a common knowledge. However, it is reasonable to assume that it is this unanimity in particular that puts the pairs at a disadvantage. According to this interpretation, persistency was fostered by the fact that both members had learned the same inappropriate procedure. However, if this is true, then, *heterogeneous pairs* should perform much better, in particular if one member has access to an appropriate procedure. On the other hand, if participants persist in their procedure irrespective of what the other person does, such a mixed pair should not perform better than a uniform M-pair.

Study 2

This issue was addressed in the second study, in which each person belonged to a pair condition. In order to test for the heterogeneity hypothesis, a mixed pair-condition was introduced, which consisted of one M- and one R-participant (condition MR). Additionally, immediately after the learning phase and at the very end of the experiment, participants were also asked to solve several tasks individually in order to test for differences in individual learning.

Method

Sample, design, and procedure The sample consisted of 112 senior high school students who were randomly assigned to one of the three pair-conditions, MM, MR, or, RR, under the restriction of approximately equal numbers within the mixed (26) and uniform pairs (15 pairs each). First, as in experiment 1, each participant was individually taught either the M- or the R-procedure. During the testing phase, each pair had again to solve four tasks, two original and two transfer tasks, of which one problem consisted of four and one of five disks.

[4] As further evidence for the persistency hypothesis, a classification of single moves revealed that the M-pairs carried out relatively more moves that are in accordance with the move-pattern procedure than the M-individuals when solving transfer tasks (cf. Reimer, 2001a).

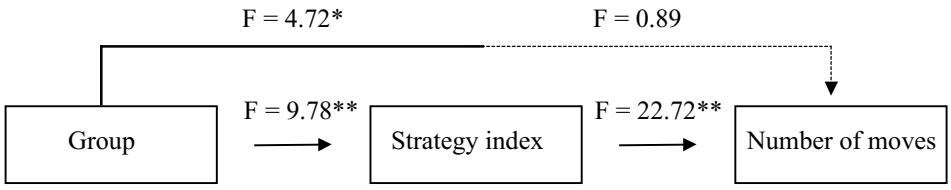


Figure 2: Path diagram: Study 1.

Additionally, participants had to individually solve two original problems with four and five disks prior to the group condition and two respective transfer problems immediately after the group condition.

Results

Single person condition I Table 3 shows the mean number of moves that were required by the individuals solving original problems immediately after the learning phase. This table is not based on the dyads but rather on the single persons as unit of analysis. For example, the condition *R / Uniform* refers to those participants who had been taught the goal-recursion procedure and who joined a person in the group condition who had access to the same procedure.

Table 3: Mean Number of Moves in the Individuals.

	R		M	
	Uniform	Mixed	Uniform	Mixed
Original tasks	29.33	26.68	28.08	26.72
Transfer tasks	28.35	28.22	52.32	49.54

An ANOVA with the factors *procedure* (R vs. M) and *group* (uniform vs. mixed) confirmed that there were no significant differences between the three experimental pair conditions on the individual level prior to the group condition (all *F*s < 2; ns). Even though it is again impossible to identify the procedures on the basis of the number of moves, the two conditions M and R may be identified on the basis of their strategy index.

As is shown in Table 4, participants who were taught the R-procedure had higher scores on the strategy index than participants in the M-condition. Accordingly, a 2x2 ANOVA revealed a strong main effect of *procedure* (F (1,108) = 121.9; p<.01; *F*s for the main effect of *group* and the interaction were less than 1).

Group condition A 2x3 ANOVA on the number of moves showed two main effects (cf. Table 5):

(1) Overall, more moves were made to solve the transfer tasks than to solve the original tasks (main effect of *type of task*: F (1,53) = 100.26; p<.01).

Table 4: Mean Strategy Index in the Individuals.

	R		M	
	Uniform	Mixed	Uniform	Mixed
Original tasks	1.17	1.17	1.05	1.04
Transfer tasks	1.16	1.17	1.08	1.05

(2) The highest number of moves was required in the uniform move-pattern condition (MM). The mixed pairs (MR) took the middle position and the RR-pairs performed best (main effect of *group*: F (2,53) = 38.08; p <.01).

(3) However, as expected, there was again a significant interaction of *type of task x group*, F (2,53) = 37.08; p <.01. As can be seen in Table 5, the differences between the pairs were almost exclusively due to the transfer tasks.

Table 5: Mean Number of Moves in the Pairs.

	RR	MR	MM
Original tasks	23.47	23.77	23.50
Transfer tasks	23.57	31.71	46.77

Overall, the observed pattern in performance may be described as follows: Whereas the original problems were solved almost optimally in each condition, there were huge differences in performance between the pairs on the transfer tasks. These difficulties were most pronounced in the MM-condition and to a much lesser extent in the MR-condition. In the RR-condition, participants had no problems with the transfer tasks at all. Interestingly, the mixed pairs, who differed significantly from the MM- as well as from the RR-pairs, performed better than the pooled uniform pairs (M_{MR} = 31.71; $M_{MM\,/\,RR}$ = 35.17; p<.05).

As expected, the pairs also differed consistently in the extent to which they applied a recursion strategy (cf. Table 6). A 2x3 ANOVA on the strategy index revealed a main effect of *group*, F (2,53) = 29.35; p <.01 (the *F*s for the main effect of *type of task* and the interaction were <1.3; ns). Within the original as well as the transfer tasks, the three pair conditions may be rank ordered on the basis of their strategy index ($S_{RR} > S_{MR} > S_{MM}$).

Table 6: Mean Strategy Index in the Pairs.

	RR	MR	MM
Original tasks	1.09	1.05	1.02
Transfer tasks	1.10	1.05	1.03

Moreover, in analogy to experiment 1, the differences in performance were, at least partially, mediated by the strategy index. If the strategy index was included as a covariate, differences in performance were reduced (F (2,53) = 38.08 vs. F (2,52) = 20.0; effect of the strategy index as a covariate: F (1,52) = 57.5; all ps<.01).

Single person condition II Table 3 and 4 (see above) also show the mean number of moves and the strategy indices in the final test of the individuals. In this test (1) the R-individuals achieved better results than the M-individuals (main effect of *procedure*: F (1,108) = 54.03; p<.01). (2) There were no significant differences in performance between the M-individuals who had joined an M- or an R-partner in the group condition (the Fs for the effects of *group* and of *procedure x group* were < 1; ns). (3) Analogous results were also observed for the strategy index (main effect of *procedure*: F (1,108) = 92.33; p<.01; neither the main effect of *group* nor the interaction were significant).

Discussion

As in the first study, the original tasks were solved almost perfectly by all pairs, even by the mixed MR-pairs. Here, the two distinct procedures *converged* to a common problem solution (cf. Reimer, 2001b). The result, i.e., that these pairs performed much better on the transfer tasks than the uniform M-pairs, suggests that group heterogeneity improved performance. Further analyses of the contributions of the M- and R-*individuals* within the mixed groups revealed that the M-participants performed significantly better when solving problems in mixed pairs than in uniform pairs. The M-participants who belonged to a mixed pair made a much higher relative number of correct moves than the M-participants who belonged to a uniform pair. This pattern was reversed for the R-participants who performed worse in the mixed pairs than in the uniform pairs. Even though this seems to support the heterogeneity hypothesis, an interesting question for further research would be whether the advantage of the M-participants who joined a mixed group had also appeared if both players had been taught *distinct inappropriate* procedures. Astonishingly, the advantage of the M-participants who belonged to the MR-pairs disappeared in the final test. However, when interpreting the results on individual learning, one should keep in mind that the pairs were not allowed to talk to each other and, therefore, had no opportunity to explain and even exchange their ideas.

Whether communication enhances or reduces the observed process losses may be another interesting issue for future research likewise the question whether the findings can be generalized to groups that consist of more than two members. In general, we can expect that the larger the groups the stronger the transfer effects supposing a group consists of homogeneous group members that share the same routine. As far as the communication issue is concerned, there is at least some evidence that communication enhances performance in pairs who have not been taught a routine but who have to develop a common strategy (cf. Reimer, 2001a).

Overall, these studies confirm the claim that groups who have routinized a problem solving procedure persist more strongly in their routine than individuals. Groups tend to behave like "ocean steamers": They often need much time and effort to work out an efficient problem solving procedure. However, once having reached a solution they are likely to persist in their routine and stick to their course irrespective of changes in the environment. The most obvious *advantage* for a group in following "the principle of inertia" consists in saving time and energy, because routines need not be actively managed and, subsequently, reduce coordination requirements (cf. Gersick & Hackman, 1990). Moreover, in situations in which the "ocean steamer" is on the right course and a routine is appropriate, groups may be expected to *surpass* individuals (cf. Hinsz et al., 1997) by better compensating for individual errors and by fostering positive transfer.

References

Betsch, T., Haberstroh, S., & Hoehle, C. (in press). Explaining routinized decision making: A review of theories and models. *Theory and Psychology*.

Gersick, C. J. G., & Hackman, J. R. (1990). Habitual routines in task-performing groups. *Organizational Behavior and Human Decision Processes*, *47*, 65-97.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*, 43-64.

Reimer, T. (2001a). Kognitive Ansätze zur Vorhersage der Gruppenleistung: Distraktion, Kompensation und Akzentuierung. *Zeitschrift für Sozialpsychologie*, 32 (2), 107-128.

Reimer, T. (2001b). Attributions for poor group performance as a predictor of perspective-taking and subsequent group achievement: A process model. *Group Processes and Intergroup Relations*, 4, 31-47.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, 7, 268-288.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, 47, 513-539.

Search, Structure or Statistics?
A Comparative Study of Memoryless Heuristics for Syntax Acquisition

William Gregory Sakas (sakas@hunter.cuny.edu)
Department of Computer Science, Hunter College, CUNY
Ph.D. Programs in Linguistics and Computer Science, The Graduate Center, CUNY
695 Park Avenue, New York, NY 10021 USA

Eiji Nishimoto (enishimoto@gc.cuny.edu)
Ph.D. Program in Linguistics, The Graduate Center, CUNY
365 Fifth Avenue, New York, NY 10016 USA

Abstract

Although several studies propose computational models of the process by which children acquire the syntax of their native language, most focus on a single algorithm applied in a single domain. Typically, the focus is *learnability* – under what conditions an algorithm can or cannot acquire the grammar of the target (native) language. Here, we present a comparative study of 12 algorithmic heuristics that are run in a domain that consists of 16 abstract languages each generated by a different grammar specified in Chomsky's principles and parameters framework. The heuristics consist of both those used in previously established models and new variations that we introduce. In contrast to a learnability study, our focus is *feasibility* – how much time and/or effort is required to achieve the target grammar. We find that the best heuristics make use of structural information obtained by parsing input sentences during the course of learning, that the performance of statistically-based heuristics are next in line, and finally, that heuristics that make use of hill-climbing search and a simple *can-parse/can't-parse* outcome from the parsing mechanism perform least well.

Background

Principles and Parameters

Chomsky (1981 and elsewhere) has proposed that all natural languages share the same innate universal principles (Universal Grammar - UG) and differ only with respect to the settings of a finite number of (binary) parameters. For example, all languages have a subject of some sort, but whether a language's grammar dictates that the subject must be overt is determined by the setting of the *null subject* parameter. The null subject parameter is set *off* in English and *on* in Spanish. The syntactic component of a grammar in the *principles and parameters* (*P&P*) framework is simply a collection of parameter values – one value per parameter. The set of human grammars is the set of all possible combinations of parameter values.

Language Acquisition The P&P framework was motivated to a large degree by psycholinguistic data demonstrating the extreme efficiency of human language acquisition. Children acquire the grammar of their native language at an early age – generally accepted to be in the neighborhood of five years old. In the P&P framework, if the linguistic theory delineates over a billion possible grammars by positing 30 parameters (2^{30}=1,073,741,824), a learner need only determine the correct 30 values that comprise the target grammar (henceforth, G_{targ}).

Given the generally accepted presupposition that a compelling theory of human language should show grammars to be easily acquirable, and since, at least on face value, parameters seem transparently learnable, it is not surprising that parameters have been incorporated into many current generative syntactic theories. However, the exact process of parameter setting has been studied only recently (cf. Bertolo, 2001 and references therein; Briscoe, 2000; Clark, 1992; Fodor, 1998a; Gibson & Wexler, 1994; Yang, 2000; among others), and although it has proved linguistically fruitful to construct parametric analyses, it turns out to be surprisingly difficult to construct a workable model of parametric syntax acquisition.

Parametric Ambiguity and the Need for Heuristics A sentence is parametrically ambiguous if it is licensed by two or more distinct combinations of parameter values. Parametric ambiguity is rampant in natural language. For example, a sequence of the form Subject-Verb-Object (SVO) is parametrically ambiguous between underlying SVO order as in English, and *verb second* (V2) order as in German.[1] Although SVO sentences can be parsed by either grammar, the derivations will be different due to the different parameter settings. By contrast, a VOS sentence is not parametrically ambiguous with respect to the V2 parameter. It can be licensed only by the -V2 value (since the second token is not a verb or auxiliary).

Ambiguity is a natural enemy of efficient language acquisition. The key problem is that, due to ambiguity, there does not exist a one-to-one correspondence between the linear left-to-right word order of an input sequence and the correct parameter values for the target grammar (as described above for an SVO sentence with respect to + or -V2). So, even if the learner hypothesizes parameter values which license the **single**, current input sentence, those values may ultimately be incorrect for G_{targ}. In the face of parametric ambiguity, efficient search heuristics must be employed to guide the learner towards the target grammar as sentences are progressively consumed by the learner. The remainder of the paper presents a comparative study of 12 search heuristics that are incorporated into current parameter-setting models of language acquisition.

Overview

A Measure of Feasibility

As a simple example of a learning heuristic and of our simulation approach, consider a domain of 4 parameters and a memoryless

[1] See Appendix for the linguistic details of how we implement the V2 parameter.

learner[2] which blindly guesses how all 4 parameters should be set upon encountering an input sentence. Since there are 4 parameters, there are 16 possible combinations of parameter settings (2^4=16), i.e., 16 different grammars. Assuming that each of the 16 grammars is equally likely to be guessed, the learner will consume, on average, 16 sentences before achieving G_{targ}. This is one measure of a model's efficiency or *feasibility*.

However, when modeling natural language acquisition, since practically all human learners attain the target grammar, the average number of expected inputs is a less informative statistic than the expected number of inputs required for, say, 99% of all simulation trials to succeed. For our blind-guess learner, this number is 72.[3] We will use this 99 percentile feasibility measure (*99% score*) for most discussion that follows, but also include the average number of inputs for completeness.

Error-Driven Learning

Although an outside oracle could ascertain when the blind-guess learner has acquired the target grammar, the learner itself has no "built-in" mechanism for identifying that it has achieved the target. Even if the correct grammar is hypothesized, the learner will most likely abandon it on the next sentence (with a probability of 15/16) and hypothesize a different (incorrect) grammar.

A standard way to build target identification into an algorithm is to dictate that the learner be *error-driven*.[4] Assuming that all inputs are grammatically correct instances of sentences that make up the target language, one could provide the learner with the ability to produce a *can-parse/can't-parse* outcome[5] given the current input sentence, the current hypothesized grammar (G_{curr}), and the rule: *Don't change G_{curr} unless there is parse failure*. With this error-driven constraint, there is no need for an outside oracle to stop the learner from relinquishing the target grammar once it is attained. Since all sentences are generated (by definition) by G_{targ}, parse success is guaranteed once $G_{targ} = G_{curr}$, and thus the learner will not be motivated to shift from its current hypothesis.

An *Error-Driven Blind-Guess* (*EDBG*) learner is our first heuristic of interest. It is easy to show that the average and 99% scores increase exponentially in the number of parameters. Clearly, human learners do not employ any strategy that performs as poorly as this.

Table 1: EDBG, # of sentences consumed

	99%	Average
EDBG	86	15.09

[2] By "memoryless" we mean that the learner processes inputs one at a time without keeping a history of encountered inputs or past learning events.

[3] The average and 99 percentile figures (16 and 72) in this section are easily derived from the fact that input consumption follows a hypergeometric distribution. See Chung (1979) for an overview.

[4] Error-driven grammar learning was first introduced by Gold (1967) and has become a standard in learnability research.

[5] We intend for a "*can-parse/can't-parse* outcome" to be equivalent to the result from a language membership test. If the current input sentence is one of the set of sentences generated by G_{curr}, *can-parse* is engendered; if not, *can't-parse*.

The Simulations

In all experiments:

- The learners are memoryless.
- The language *input sample* presented to the learner consists of only grammatical sentences generated by G_{targ}.
- The heuristics are tested in a 4-parameter, 16-language domain (see Appendix for details).
- For each heuristic, 1,000 trials were run for each start/target grammar pair.[6]
- At any point during the acquisition process, each sentence of G_{targ} is equally likely to be presented to the learner.[7]

Subset Avoidance and Other Local Maxima Depending on the algorithm and the learning domain, it may be the case that a learner will never be motivated to change G_{curr}, and hence be unable to ultimately achieve the target. This is often referred to as a *local maximum*. For example, the EDBG learner will be trapped if G_{curr} generates a language that is a superset of G_{targ}. There is a wealth of remarkable learnability literature that addresses local maxima and their ramifications.[8] However, since our study's focus is on feasibility (rather than on whether a domain is learnable) given a particular algorithm, we posit a built-in avoidance mechanism, such as the *subset principle* and/or *default values* that preclude local maxima; hence, we set aside trials where a local maximum ensues.

The Heuristics

The heuristics we present can be separated into two families based on the way they process input sentences: those that guide the learner towards the target by use of a can-parse/can't-parse outcome, and those that take advantage of information gleaned from the parse trees that are constructed by the parser. Gibson and Wexler's (1994) *Triggering Learning Algorithm* (*TLA*) and Yang's (2000) *Variational Model* make use of can-parse/can't-parse outcomes only. Fodor's (1998a) *Structural Triggers Learner* (*STL*) takes advantage of more extensive structural information obtained from sentence parsing.

TLA

The TLA incorporates two search heuristics: the *Single Value Constraint* (*SVC*) and *Greediness*. In the event that G_{curr} cannot parse the current input sentence *s*, the TLA attempts a second

[6] For the STL models presented later in the paper, the start grammar is unspecified.

[7] Not reported here are results from simulations run on several different distributions of input sentences, in particular, those where the shorter (presumably simpler) sentences occur more frequently. The relative performance of the heuristics is substantially the same; however, in all cases acquisition requires more inputs. These distributions and their relationship to the rate of ambiguity in the domain are currently being analyzed. Also, see Niyogi & Berwick (1996) for mathematical treatment of how the distribution of inputs affects TLA performance.

[8] Discussion of the problem of subset relationships among languages starts with Gold's (1967) seminal paper and is discussed in Berwick (1985) and Wexler & Manzini (1987). Detailed accounts of the types of local maxima that the learner might encounter in a domain substantially similar to the one we employ are given in Frank & Kapur (1996), Gibson & Wexler (1994), and Niyogi & Berwick (1996).

parse with a randomly chosen new grammar, $G_{new,}$ that differs from G_{curr} by exactly one parameter value (SVC). If G_{new} can parse *s*, G_{new} becomes the new G_{curr}, otherwise G_{new} is rejected as a hypothesis (Greediness). Following Berwick and Niyogi (1996), we also ran simulations on two variants of the TLA – one with the Greediness heuristic but without the SVC (TLA minus SVC, *TLA–SVC*) and one with the SVC but without Greediness (TLA minus Greediness, *TLA–Greed*). The TLA has become a seminal model and has been extensively studied (cf. Bertolo, 2001 and references therein; Berwick & Niyogi, 1996; Frank & Kapur, 1996; Sakas, 2000; among others). We will not rehash these earlier discussions here. We include the TLA in our study to present comparisons against other models.[9]

Table 2: TLA variants, # of sentences consumed

	99%	Average
TLA-SVC	117	18.12
TLA-Greed	102	19.44
TLA	227	22.56

The Variational Learner

Like other models, Yang's *Variational Learner* (*VL*)[10] incorporates the notion of a current grammar hypothesis (G_{curr}) which is applied to the current input sentence. However, unlike the other learners, the VL maintains, for each parameter, a *weight* varying from 0 to 1. Roughly, the weight of a parameter can be construed as a measure of the past successes (or failures) of either a 0 or 1 value for that parameter during prior parses of input sentences. If the weight is closer to 0, then the 0 value has been more successful; if the weight is closer to 1, the 1 value has been more successful.[11] The VL uses the weights to guide the selection of the next hypothesis.

Since Yang's emphasis was on learnability in the limit and not on feasibility, in order to compare performance against other learners, we make a minor adaptation by adding a stopping criterion (see Step 4 below). The algorithm proceeds as follows:

1. Initialize weights for all parameters to 0.5.
2. Choose a new G_{curr} randomly, but biasing the choice towards parameter values favored by the current weights.
3. If G_{curr} succeeds in parsing the current input, nudge the weights in the direction of the values that make up G_{curr} (*reward* G_{curr}). This has the effect of making those parameter values that make up G_{curr} more likely to be chosen in the future. Otherwise nudge the weights away from G_{curr}'s parameter values (*punish* G_{curr}).
4. If all the weights are within a target threshold of either 0 or 1 then learning ends. Else go to Step 2.[12]

We chose 0.01 as the threshold value for stopping the learning process. Given weights w_1, w_2, ..., w_n where n represents the number of parameters, learning ends if all weights are in the range: $0.0 \leq w_i < 0.01$ or $0.99 < w_i \leq 1.0$ (i.e., every weight is very close to either a 0 value or a 1 value). Note that this criterion could be UG-endowed; that is, an oracle is not required to stop the algorithm.

In both Yang's original VL and the version presented above, the amount that the weights are "nudged" during learning is controlled by the *learning rate* (γ).[13] We ran simulations with γ=.1, γ=.5, and γ=.75.

Yang's model is of particular interest because it is an explicit implementation of the idea to keep statistics on the effectiveness of parameter settings based on the success or failure of past learning events. The reward/punish scheme of the VL is arguably an extension of error-driven learning in that incorrect grammars are punished, but it is significantly different from the standard constraint in which G_{curr} cannot change after a successful parse. The scheme works because the VL is statistical in nature; as parameter values vie for domination, the ones most successful in the past (and hence rewarded) are most likely to be chosen in the future and will eventually prevail.[14]

However, the statistical nature of the VL, together with the stopping criterion, may lead the learner to a local maximum. A parasitic input sample might lead to an abundance of evidence (i.e., successful parses) for incorrect parameter values early in the course of learning. If the learner encounters enough misleading evidence, it would cross the stopping threshold prematurely (on parameter values that are different from those that make up G_{targ}). This effect can be prevented by keeping the learning rate low, which gives the weights elbow room to fluctuate more gently until the evidence eventually supports the correct target. Unfortunately, the cost of keeping the learning rate low can be quite severe. In our 16-language domain, the VL requires well over 100,000 input sentences[15] to achieve a 99% score with a learning rate of 0.1. When the learning rate was increased to 0.5, the expected number of inputs dropped to 242, and when the rate was increased to 0.75, the number dropped even further to 84, albeit with many more local maxima (learning failures).

To combat the extraordinarily high number of input sentences needed, we incorporated into Yang's VL a variation of the

[9] In particular, the data in Table 2 reinforce Berwick & Niyogi's (1996) conclusion that in addition to creating local maxima, SVC and Greediness reduce learning speed.

[10] We use the proper name and acronym for readability; they are not used by Yang.

[11] Of course, if the values from one or more parameters are strongly tied to values of another parameter(s), the weight does not represent a simple ratio or percentage of success. Still, Yang's intention is that the weights are an informative measure of past performance.

[12] This stopping criterion could lead the VL to converge on the wrong grammar. We consider these cases as local maxima, and as such, trials in which local maxima ensue are discarded. Also implemented was a different stopping criterion which requires that all the weights fall within a threshold of the values that make up G_{targ}. Clearly, this strategy could not be UG-endowed because it requires that the learner have advance knowledge of these values. It is also less efficient than the criterion above, though it does enforce convergence on the correct grammar.

[13] The exact amount follows the Linear reward-penalty (L_{R-P}) scheme (Bush & Mosteller, 1958). See Yang (2000) for details.

[14] See Yang (2000) for a proof of convergence as the number of inputs approaches infinity.

[15] We found that many trials required over 100,000 input sentences which was an arbitrary stopping point built into our simulation.

standard error-driven constraint.[16] The *Error-Driven Variational Learner* (*EDVL*) proceeds as follows:

1. Initialize weights for all parameters to 0.5.
2. Randomly choose a new G_{curr} to start with.
3. If G_{curr} succeeds in parsing the current input, do nothing (no reward; no punish). Otherwise, randomly choose a new grammar (G_{new}) biased by the weights, re-parse with G_{new}, and apply Yang's original reward/punish scheme to adjust the weights.
4. Set G_{curr} to values **directly indicated by the weights** – that is, if $w_i > 0.5$, then the value of parameter *i* becomes 1, if $w_i < 0.5$, then the value of parameter *i* becomes 0, if $w_i = 0.5$, then either 0 or 1 is chosen at random.
5. Go to Step 3.

Note that no stopping criterion is needed, and that there are no local maxima because a move to another grammar is always possible before G_{targ} is attained. As is shown in the table below, the addition of the error-driven constraint greatly improves performance. Surprisingly, contra Yang's original model, performance **deteriorates** as the learning rate increases. We speculate that this is because a high learning rate encourages excessive exploration of different grammars, thus superseding the (positive) conservative nature of the error-driven constraint.

Table 3: VL & EDVL, # of sentences consumed

		99%	Average
VL	$\gamma = .1$	Over 100,000	Over 33,000
VL	$\gamma = .5$	242	46.35
VL	$\gamma = .75$	84	17.91
EDVL	$\gamma = .1$	44	8.39
EDVL	$\gamma = .5$	55	9.23
EDVL	$\gamma = .75$	75	12.16

STL

Fodor's *Structural Triggers Learner* (*STL*) makes greater use of the parser than the models discussed so far. A key feature of the model is that parameter values are not simply 0 or 1, but rather bits of tree structure or *treelets*. Thus, a grammar in the STL sense is a collection of treelets rather than a collection of 1's and 0's. The STL is error-driven. If G_{curr} cannot license *s*, new treelets will be utilized to achieve a successful parse.[17] Treelets are applied in the same way as any "normal" grammar rule, so no unusual parsing activity is necessary. The STL hypothesizes grammars by adding parameter value treelets when they contribute to a successful parse.

The basic algorithm for all STL variants is:

1. If G_{curr} can parse the current input sentence, retain the treelets that make up G_{curr}.
2. Otherwise, parse the sentence making use of any or all parametric treelets made available by UG, and adopt those treelets that contribute to a successful parse.

The STL stands apart from other acquisition models in that it can detect when an input sentence is parametrically ambiguous. During a parse of *s*, if more than one treelet could be used by the parser (i.e., a *choice point* is encountered), then *s* is (possibly) parametrically ambiguous. The TLA and the VL do not have this capacity because they rely only on a can-parse/can't-parse outcome and do not have access to the on-line operations of the parser. Originally, the ability to detect ambiguity was employed in two variations of the STL: the *strong STL* (*SSTL*) and the *weak STL*.

The SSTL executes a full parallel parse of each input sentence and adopts only those treelets (parameter values) that are present in all the generated parse trees. Note that even if the surface word order of the input is ambiguous between several languages (i.e., the sentence belongs to more than one language), the SSTL can identify unambiguous parameter values (treelets) by looking at **all** of the tree structures that the parser constructs for the sentence. This makes the SSTL an extremely powerful model, and for this reason, it establishes an upper standard against which to compare other models. It is not, however, proposed as a psychologically realistic model. As with the Error-Driven Blind-Guess (EDBG) learner, it is clear that human learners do not exhibit an SSTL-like strategy. The consensus in sentence processing research is that adults are only capable of limited parallel parsing, if any (cf. Gibson, 1991). It does not seem plausible to suppose that children possess a more powerful mechanism than adults.

On the other hand, the weak STL executes a psychologically plausible left-to-right serial (deterministic) parse. One variant of the weak STL, the *waiting STL* (*WSTL*), deals with ambiguous inputs abiding by the heuristic: *Don't learn from sentences that contain a choice point.* These sentences are simply discarded for the purposes of learning. This is not to imply that children do not parse ambiguous sentences they hear, but only that they set no parameters if the current evidence is ambiguous.

Table 4: SSTL & WSTL, # of sentences consumed

	99%	Average
SSTL	14	3.35
WSTL	30	5.11

As with the TLA, these STL variants have been studied from a mathematical perspective (Bertolo et al., 1997; Sakas, 2000; Sakas & Fodor, 2001a). Although the simulation results indicate notably better performance than the other models examined thus far in this paper, previous mathematical analyses lend doubts to the ultimate success of the WSTL model. The WSTL requires some **fully** unambiguous sentences for any learning to take place. It is probably the case that fully unambiguous triggers are few and far between in the domain of human languages, and negative WSTL performance is exponentially tied to the rate of ambiguity in the domain; that is, in a more realistically ambiguous domain than the one we explored so far, the WSTL may consume and discard an extremely large number of sentences before attaining G_{targ}. This result has spurred a new class of weak STL variants

[16] The resulting algorithm is similar to one originally proposed in Fodor (1998b).

[17] In addition to the treelets, UG principles are also available for parsing, as they are in the other models discussed above. See Appendix for details that apply to the domain we use here.

which we informally call the *guessing STL* family (Sakas & Fodor, 2000).

The basic idea behind the guessing STL models is that there is some information available even in sentences that are ambiguous, and a strategy could exploit that information. We incorporate four different heuristics into the original STL paradigm:

- *Strong Oracle* (*SO*) – perform a parallel parse of the current input *s* and choose a hypothesis grammar that licenses *s* and is most similar (in terms of hamming distance) to G_{curr}.
- *Random Choice* (*RC*) – parse serially; when a choice point is encountered, randomly pick a parsing alternative and adopt the treelets that are present in the final tree structure.
- *Minimal Chain* (*MC*) – parse serially; when a choice point is encountered, pick the choice that obeys the *Minimal Chain Principle* (De Vincenzi, 1991), i.e., avoid positing movement transformations if possible.
- *Local Attachment/Late Closure* (*LAC*) – parse serially; when a choice point is encountered, pick the choice that attaches the new word to the current constituent (Frazier, 1978).

Although the MC and LAC heuristics are not stochastic, we regard them as "guessing" heuristics because, unlike the WSTL, a learner cannot be certain that the parametric treelets obtained from a parse guided by MC and LAC are correct for the target. These heuristics are based on well-established parsing preferences that adults employ, so it seems likely that children apply them also (Fodor, 1998b).

The SO heuristic was originally conceived by Fodor and Teller (2000) as an extension of the SVC. Their main point was the efficiency advantage that results from using the parser to **find** a successful parse of the current sentence, so that can't-parse trials are eliminated. Given this capability, the question of how to choose among possible parses arises. The SO criterion presupposes full parallel parsing, which is unrealistic, but our approximation to it would result from letting the parser employ new treelets only where current ones do not suffice. Our data show that the conservatism of the SO heuristic pays off: it gives the strongest performance of any learner in our study, including the SSTL.

The RC, MC, and LAC heuristics show a significant improvement over the waiting strategy (WSTL). The difference between the three variants is slight.

Table 5: guessing STL family, # of sentences consumed

	99%	Average
SO	10	2.33
RC	26	3.43
MC	26	3.44
LAC	24	3.25

Conclusions

One can expect 86 sentences to be consumed by a population of (baseline) EDBG learners before 99% of the population acquires the target grammar in our small domain of 16 languages, and TLA variants require more input than that. The implemented heuristics in the two paradigms forego information, structural or statistical, in favor of a simple mechanism – the information needed for language acquisition must somehow be available from the surface word strings that make up the languages of the domain. They will be successful only if either i) there are recognizable, unambiguous signals in the surface strings that trigger correct parameter values or ii) the distribution of cross-language ambiguity[18] in the domain being studied is conducive to the heuristics being employed. There is faint evidence for both cases. For (i), in a domain without *null subject/topic*, the fact that a VOS sentence does not have a finite verb or auxiliary as the second token is indeed secure evidence for the -V2 parameter value. Although true for the -V2 value in this case, it is unclear how other plausible syntactic descriptions will offer the same advantage for the gamut of complicated linguistic phenomena (e.g., null subject/topic) with which human languages are inundated. As for (ii), previous work by the authors (Sakas, 2000) demonstrates that the TLA is a feasible learner in *strongly smooth* domains – domains in which there is a monotonic correlation between the similarity of grammars and the languages that are generated by them. Although still an open question, linguists have argued that natural languages are not strongly smooth.

The Error-Driven Variational Learner (EDVL) is a more promising model of language learning. On average, 44 sentences will allow 99% of the population to attain the target grammar. Its success can be attributed to the use of a strategy that maintains statistics of past performance without the unreasonable requirement that the learner memorize an entire input sample.

The most efficient heuristics, however, are those that make the most use of tree structure produced by the parsing mechanism: the psychologically plausible STL variants require almost half the number of inputs consumed by the EDVL.

Conjectures and Ongoing Research

The relative success of the EDVL is important for a reason not explicated earlier. Preliminary investigation points to the fact that, uniquely among the heuristics in our study, the EDVL performs more efficiently in ambiguous domains than in unambiguous ones at the outset of learning. This could turn out to be crucial as the guessing STL family might easily be foiled by larger, more syntactically complicated domains which generate sentences that contain a multitude of choice points – the result being that the parse tree computed for an input sentence, which guides parameter setting, reveals little about the target grammar. However, it has been shown that the STL performs extremely efficiently after just a few initial parameters have been set (Sakas, 2000). One can imagine a hybrid model of a guessing STL heuristic combined with a VL-like statistical heuristic where the statistical heuristic is used to bootstrap learning, and as performance deteriorates, the structural heuristic acquires more control of the acquisition process. This is being investigated in ongoing research.

The hard data that comprise our current study, however, reveal that the utilization of structural information outperforms the statistical heuristic overall.

[18] For present purposes, cross-language ambiguity is defined as some measure based on the intersection of the sets of surface forms that make up the languages in the domain. See Fodor & Sakas (2001b) for other definitions and the effects of ambiguity on learning efficiency.

Appendix: Simulation Domain

The four parameters are: *Specifier Initial/Final*, *Complement Initial/Final*, *V2 Movement*, and *Null Subject-Topic*. We largely adopt the first three parameters from Gibson and Wexler (1994). The Specifier and Complement position parameters are non-transformational. It is assumed that the subject is base generated in Spec of IP, and that the verb moves to I in the final structure (if I is not filled with an auxiliary).

Input sequences are formed with tokens representing adverb, subject, verb, auxiliary, direct object, and indirect object. Following Gibson and Wexler, we assume that the learner can directly determine the role of a noun phrase. For example, the noun phrase *the big dog* is interpreted as a subject or one of the object types based on its role when uttered.

The following constraints on the domain are in place: all sentence types are degree-0 (i.e., no subordinate clauses); Spec of CP is to the left of C'; C always precedes IP; all adverbs are sentential (i.e., base generated in Spec of CP); and there is no transformational reordering of constituents (e.g., topicalization, wh-movement, scrambling, etc.) with the exception of the V2 movement described below.

The V2 Movement [+/-V2] parameter determines whether a finite verb moves to the second position in the root clause. That is, a "verb second" language [+V2] entails that the finite verb is transformationally fronted to C (from I) and that a topical element is moved into Spec of CP (if Spec of CP is not filled with an adverb). In a [-V2] language, the verb moves only up to I, and there is no movement of a topic into Spec of CP.

Extending Gibson and Wexler's domain, we add the Null Subject-Topic [+/-Null] parameter. This parameter represents either subject drop or topic drop depending on the value of the V2 parameter. In [-V2] languages, an overt subject is either optionally [+Null] (as in Spanish and Japanese) or obligatorily [-Null] (as in English). For [+V2] languages, the parameter works similarly, but instead of a subject, an overt topic may be optionally [+Null] or obligatorily [-Null].

Acknowledgments

We would like to thank Bob Berwick, Janet Fodor, Virginia Teller, Charles Yang, and the learnability project (CoMoLA) group at CUNY for much useful discussion and suggestions. This research was supported by CUNY Collaborative Grant #92902 and PSC-CUNY Grant #63387.

References

Bertolo, S. (Ed.). (2001). *Language Acquisition and Learnability*. Cambridge, UK: Cambridge University Press.

Bertolo, S., Broihier, K., Gibson, E., & Wexler, K. (1997). Characterizing learnability conditions for cue-based learners in parametric language systems. *Proceedings of the Fifth Meeting on Mathematics of Language*.

Berwick, R. C. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Berwick, R. C., & Niyogi, P. (1996). Learning from triggers. *Linguistic Inquiry*, *27* (4), 605-622.

Briscoe, T. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, *76* (2), 245-296.

Bush, R., & Mosteller, F. (1958). *Stochastic Models for Learning*. New York: Wiley.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chung, K. L. (1979). *Elementary Probability Theory with Stochastic Processes*. New York: Springer-Verlag.

Clark, R. (1992). The selection of syntactic knowledge. *Language Acquisition*, *2* (2), 83-149.

De Vincenzi, M. (1991). *Syntactic Parsing Strategies in Italian*. Dordrecht: Kluwer.

Fodor, J. D. (1998a). Unambiguous triggers. *Linguistic Inquiry*, *29* (1), 1-36.

Fodor, J. D. (1998b). Parsing to learn. *Journal of Psycholinguistic Research*, *27* (3), 339-374.

Fodor, J. D., & Teller, V. (2000). Decoding syntactic parameters: The superparser as oracle. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Frank, R., & Kapur, S. (1996). On the use of triggers in parameter setting. *Linguistic Inquiry*, *27* (4), 623-660.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. Doctoral dissertation, University of Connecticut.

Gibson, E. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. Doctoral dissertation, Carnegie Mellon University.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25* (3), 407-454.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10* (5), 447-474.

Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces. *Cognition*, *61*, 161-193.

Sakas, W. G. (2000). *Ambiguity and the Computational Feasibility of Syntax Acquisition*. Doctoral dissertation, City University of New York.

Sakas, W. G., & Fodor, J. D. (2000). Setting syntactic parameters: A computational analysis of child-directed speech. *CUNY Collaborative Incentive Research Grant*. City University of New York.

Sakas, W. G., & Fodor, J. D. (2001a). The Structural Triggers Learner. In S. Bertolo (Ed.), *Language Acquisition and Learnability*. Cambridge, UK: Cambridge University Press.

Sakas, W. G., & Fodor, J. D. (2001b). Learning-relevant properties of natural language domains. *The Seventh Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*. Saarbrücken, Germany.

Wexler, K., & Manzini, R. (1987). Parameters and Learnability in Binding Theory. In T. Roeper & E. Williams (Eds.), *Parameter Setting*. Dordrecht: Reidel.

Yang, C. D. (2000). *Knowledge and Learning in Natural Language*. Doctoral dissertation, MIT.

Modeling Driver Distraction from Cognitive Tasks

Dario D. Salvucci (salvucci@mcs.drexel.edu)
Department of Mathematics and Computer Science, Drexel University
3141 Chestnut St., Philadelphia, PA 19104

Abstract

Driver distraction has become a critical area of study both for research in investigating human multitasking abilities and for practical purposes in developing and constraining new in-vehicle devices. This work utilizes an integrated-model approach to predict driver distraction from a primarily cognitive secondary task. It integrates existing models for a sentence-span task and driving task and investigates two methods in which the resulting model can perform multitasking. Model predictions correspond well qualitatively to two of three measures of driver performance as collected in a recent empirical study. The paper includes a discussion of the potential for building multitasking models in the framework of a production-system cognitive architecture.

Introduction

Computational cognitive modeling continues to mature rapidly as an area for both theoretical advances in understanding cognition and practical advances in developing intelligent technology. Cognitive modeling has grown from addressing only simple cognition in basic psychological tasks to capturing integrated cognitive, perceptual, and motor processes in large-scale complex, dynamic tasks (e.g., Chong, 1998; Jones et al., 1999). This paper investigates the application of cognitive models to an extremely common yet complex task: driving. Driving involves the continual multitasking of a number of subprocesses that make use of the driver's cognition, perception, and motor movements. This rich spectrum of necessary skills makes driving an ideal task in which to investigate how humans execute complex tasks and how computational models can represent and predict the multitasking behavior in these tasks.

Driver Distraction and Cognitive Modeling

One particular aspect of driver multitasking that has received enormous attention from media and researchers alike is that of "driver distraction" -- namely, the effects of multitasking while performing some secondary task. Numerous studies have now found that performing primarily perceptual-motor tasks while driving (e.g., dialing a cellular phone) can impair driver performance (e.g., Alm & Nilsson, 1995; McKnight & McKnight, 1993). These studies generally conclude, perhaps not surprisingly, that pulling a driver's visual attention from the road and/or her hand(s) off the steering wheel degrades the driver's ability to maintain a central lane position, follow a lead car with a constant headway, or react to potential hazards. Such studies have subsequently led to legislation at all government levels to restrict the use of potentially distracting secondary-task devices. While driver distraction is generally associated with effects on perceptual-motor processes, researchers have also reported that "cognitive distraction" arising from purely cognitive secondary tasks can adversely affect driver performance (e.g., Alm & Nilsson, 1995). These results are not fully conclusive and seem to depend highly on the secondary task as well as the driving situation; nevertheless, it is clear that even purely cognitive tasks can impact driver performance in certain situations.

To better understand driver behavior and the sources of driver distraction, researchers have attempted to develop integrated driver models that capture driver behavior in a computational manner (e.g., Aasman, 1995). These models provide insight into the sources of distraction by elucidating the exact processes by which a driver attends to the external environment, processes this information cognitively, and then reacts and manipulates the environment. In addition, the computational models may be used to generate predictions about the effects of distraction on driver performance; for instance, the ACT-R driver model (Salvucci, Boer, & Liu, 2001) has been integrated with various models of cell-phone dialing to predict the impact of dialing on lane-keeping performance (Salvucci, 2001; Salvucci & Macuga, 2001). However, this previous work has addressed only primarily perceptual-motor secondary tasks with little cognitive component (like cell-phone dialing); to date, no models have demonstrated the ability to represent and generate "cognitive distraction" from primarily cognitive tasks.

Modeling "Cognitive Distraction"

This paper describes the first attempt to predict cognitive distraction with a computational cognitive model. This work employs the same methodology as in previous work for perceptual-motor distraction, namely the "integrated model approach" based in a cognitive architecture (see Salvucci, 2001). Cognitive architectures are computational frameworks that incorporate built-in, well-tested parameters and constraints on human cognitive and perceptual-motor abilities. This work focuses on a particular architecture, ACT-R (Anderson & Lebiere, 1998), that represents factual knowledge as declarative chunks and procedural knowledge as condition-action "production rules". For our purposes, the ACT-R architecture has two important benefits: (1) it facilitates reuse and integration of multiple behavioral models, and (2) it provides built-in interfaces and default parameters that facilitate *a priori* predictions of real-world

metrics of human performance (e.g., reaction times, keystrokes, eye movements). The integrated model approach takes advantage of these benefits to incorporate a model of secondary-task behavior with the ACT-R driver model to predict the effects of executing the secondary task on the primary driving task.

The initial demonstrations of this approach (Salvucci, 2001; Salvucci & Macuga, 2001) examined a primarily perceptual-motor task, namely dialing a cellular phone using different modalities (e.g., manual button input versus speech input). The work showed that an integrated driving-dialing model predicted degraded steering performance for the modalities that required the driver to look at the cell phone (i.e., manual dialing), thus drawing visual attention away from the roadway. The work presented here generalizes the previous work in two important ways. First, although it utilizes the same methodology to predict driver distraction, it predicts distraction from a primarily cognitive task — namely, a sentence-span task that involves rehearsal and recall of a sequence of words. Second, unlike the previous work, it makes use of an existing model for the secondary task (with some necessary adaptation) as well as an existing model for the primary driving task, thus demonstrating the importance and benefits of model re-use.

This paper begins with a review of the driving and secondary tasks modeled here, namely those from the empirical work of Alm and Nilsson (1995) showing effects of the sentence-span task on driver car-following performance. It then provides an overview of the integrated model approach incorporating existing models of both the driving and secondary tasks, including two methods of performing explicit multitasking between the individual task models. Finally, it compares the model's predictions with Alm and Nilsson's empirical results and discusses the broader implications of the methodology to studying multitasking in the framework of a cognitive architecture.

The Sentence-Span and Driving Tasks

The task and empirical results that will be used to validate the model predictions are taken from Alm and Nilsson (1995). Their study aimed to show exactly those effects that we are attempting to model, namely effects of cognitive secondary tasks on driver performance. For the purposes of this paper, we would like to recreate this task for the integrated model as closely as possible to facilitate later comparison between model and empirical results.

Sentence-Span Task

Alm and Nilsson (1995) employed a sentence-span task that involves the processing of sentences and the recall of words in these sentences (see Daneman & Carpenter, 1980). The task comprises two stages. In the first stage, drivers listened to five sentences of the form "X does Y" — for instance, "The boy brushed his teeth" or "The train bought a newspaper." They would also report after each sentence whether the statement was generally sensible. In the second stage, drivers were asked to state the last word of each sentence in order. For instance, for the sentences "The boy brushed his teeth" and "The train bought a newspaper," the driver would report "yes" and "no" after each sentence (respectively) and would then report the memorized list "teeth,", "newspaper," etc. The sentence-span task itself involves two concurrent activities, namely judging of sentence sensibility and memorization (and rehearsal) of final words. When combined with driving, the task puts a substantial cognitive load on drivers as they attempt to integrate the tasks.

Driving Task

As a realistic scenario in which to test interaction with the sentence-span task, Alm and Nilsson (1995) used a car-following task where the lead vehicle would sometimes perform unsafe maneuvers and leave the driver in a "safety-critical" situation. During the normal stages of the task, the lead vehicle maintained a 75 m headway distance from the driver's vehicle. Occasionally, the lead car braked suddenly with a deceleration of 4 m/s^2 until its speed reached 50 km/hr (or until a maximum of 5 s of deceleration), then accelerated at 3 m/s^2 until its speed reached 90 km/hr. The original study also included non-safety-critical situations in which the lead vehicle would indicate a right turn and eventually turn off the road; their analysis does not examine these situations in detail and they are not discussed further.

The Alm and Nilsson study provided three metrics by which they measured driver performance: (1) reaction time to the braking event, measured as the time lapse between the start of the event and the driver's initial depression of the brake; (2) lateral deviation, measured as the root-mean-squared error of the driver's vehicle position to the center of the lane; and (3) headway distance, measured as the distance between the driver's vehicle and the lead vehicle. The results section will compare the model's predictions to the empirical results from human drivers for all three metrics.

Empirical Study

Alm and Nilsson's (1995) empirical study included a total of 40 participants in two experimental groups: a *task group* that occasionally performed the sentence-span task while driving, and a *control group* that did not perform the task. In both groups, each driver negotiated four safety-critical situations in which the lead vehicle would brake suddenly. The timing of the events was randomized to either near the start or near the end of the span task (in the task group) such that drivers could not predict when the events would occur.

The driving task was performed in a high-fidelity driving simulator to give participants as realistic an impression of real-world driving as possible. The simulator included a moving-base system (based on a Saab 9000 with manual transmission), wide-angle visual system, vibration generation, and temperature regulation. The driving environment comprised a simulated 80 km two-lane highway (one lane in each direction) with oncoming traffic in the opposite lane. The highway had a very low curvature so that steering down the roadway was relatively straightforward even at high speeds. The sentence-span task was performed through an Ericsson hands-free telephone mounted on the instrument panel to the right of the steering wheel. Drivers needed only press one key to activate

(answer) the phone at the start of the task, and given practice with the phone, drivers could easily activate the phone without looking. Sentences were presented by means of a tape recorder, and driver responses were recorded on a second tape recorder.

The results of the empirical study will be discussed in a later section to facilitate comparison with model predictions. It should be noted that the original study also included both younger and older drivers to demonstrate the interactions of cognitive distraction with age. This paper only addresses the data from the younger drivers (mean age 29); the existing driver model used in this paper has been validated with data from younger drivers, and thus we expect the model to better account for the younger-driver data from the original study.

The Integrated Task Models

To model and predict the interaction of the sentence-span and driving tasks, this work utilizes the "integrated model" methodology employed in previous work (see Salvucci, 2001): Given an existing model of driver behavior, we develop or acquire a model of behavior in the secondary task, integrate the two models to perform multitasking, and finally run the integrated model to generate behavioral data. One critical element of this integration is the potential for generating *a priori* predictions — that is, rather than fitting the model to data by manipulating parameters, we carry over defaults and parameter settings from existing models and immediately use them in the integrated model. In addition, we benefit from re-use of models that have been rigorously tested in other studies. These and related issues will be discussed further in later sections. This section describes the individual task models as well as the two versions of the final integrated model.

Sentence-Span Model

The model for the sentence-span task comes from Lovett, Daily, and Reder (2000), who developed an ACT-R (Anderson & Lebiere, 1998) process model as part of their investigation of individual differences in working memory. Although the original model does not literally perform the sentence-span task, it does perform a closely related task called the MODS task in which people read strings of letters while memorizing final digits for later recall. The original model provides three critical components that are re-used in the sentence-span model: (1) the positional representation used to encode memorized items, (2) production rules that perform rehearsal of memorized items, and (3) production rules that retrieve and report the items in sequence. Interested readers can refer to Lovett, Daily, and Reder (2000) for a more detailed description of these components.

Given this core model, the sentence-span model required two modifications: (1) the addition of production rules to encode a sentence and decide whether it is sensible, and (2) the incorporation of perceptual-motor productions to hear and speak words (rather than read and type characters as in the MODS task). Table 1 shows the production rules in the final sentence-span model and indicates those rules taken from the original MODS model. While there are a number of new rules in this model, it should be noted that the first six deal with particulars of the sentence-span task involving hand movement and encoding of speech, and the final two non-MODS rules simply terminate the articulation and recall goals. The process of confirming whether or not the sentence is sensible is not modeled in any detail, but rather the model simply assumes that this process happens during the listening productions and signals a confirmation by firing the Confirm-sentence rule. In addition, the model assumes that each sentence component (subject, verb, object) requires one second of speech time.

Table 1: Sentence-span model production rules, with markings for whether they are original MODS-model rules and whether they pass control to driving in the Single-Step (SS) and Group-Step (GS) models.

Production Rule	MODS	Passes Control	
		SS	GS
Move-hand-to-phone		x	x
Press-button		x	x
Move-hand-to-wheel		x	x
Hear-sentence-subject		x	x
Hear-sentence-verb		x	x
Hear-sentence-object		x	x
Confirm-sentence		x	x
Create-memory	x	x	x
Rehearse-memory	x	x	
Done-articulate		x	x
Recall-span	x	x	
No-recall	x	x	
Say-item	x	x	
Next-item	x	x	x
Done-recall		x	x

Driver Model

The model of driver behavior is an ACT-R model that integrates control, monitoring, and decision making to navigate highway environments with traffic (Salvucci, Boer, & Liu, 2001). For control, the model employs a two-level model of steering that uses a "far point" on the road to guide predictive steering and a "near point" on the road to center the vehicle. For monitoring, the model encodes its surrounding environment using simulated eyes to maintain situation awareness. For decision making, the model checks the current situation and decides when to perform maneuvers such as lane changes. Thus, the driver model incorporates both lower-level perception and action for vehicle guidance and higher-level cognition for awareness and decision making. This driver model has been shown to account for a number of aspects of human highway driving, including nearing the inner curb during curve negotiation and switching gaze to the destination lane at the start of a lane change (see Salvucci, Boer, & Liu, 2001). Also, as mentioned earlier, the driver model has been employed to predict the effects of driver distraction from cell-phone dialing in different modalities (Salvucci, 2001a, 2001b; Salvucci & Macuga, 2001).

The complexities of the driver model make it infeasible to describe here in any level of detail. However, it is worthwhile to highlight two critical aspects of the model that are essential to the endeavor of predicting driver distraction. First, because of its implementation in the ACT-R architecture, the model is constrained to a serial line of cognitive processing. Thus, the cognitive integration of perception, action, and decision making is done in a serial loop: the model encodes relevant perceptual variables, processes these variables, then issues motor commands to manipulate the environment. When secondary tasks are added into this main loop, they naturally have some impact on the frequency with which the updating processes can occur, and thus can affect driver performance. Second, the driver model interacts with a simulated driving environment and generates a full behavioral protocol, as would a human driver navigating the same environment in a driving simulator. This faithfulness to predicting real-world data facilitates rigorous and straightforward comparison between model predictions and empirical results.

Integrated Model

In general, integration of multiple models in a production system such as ACT-R is rather straightforward: we can simply add the sentence-span memory structures to the driver model. However, two challenges arise that must be dealt with. First, the integrated model must decide how to multitask between the two component models. As in previous applications of this methodology, there does not yet exist a rigorous model of multitasking that we can employ, but we can use reasonable heuristics to guide us. Multitasking in the integrated model is performed explicitly (instead of preemptively) in that each model must pass control to the other, presumably at a fairly frequent interval. Because driving is the primary task, we are most concerned about when the secondary task model (i.e., the sentence-span model) will cede control back to driving. This paper explores two approaches for attacking this problem. The conservative approach would only allow the secondary task to fire a single production, then immediately cede control back to driving. A less conservative (though still fairly conservative) approach would allow small logical groupings of production firings to occur before passing control. These approaches were used to develop two versions of the model termed the Single-Step (SS) and Group-Step (GS) models, respectively. Table 1 indicates which productions pass control for each model. While every rule is marked for the SS model, the GS model allows certain rules to continue: the Rehearse-memory rule that rehearses memorized items in rapid succession, and the threesome of rules that combine to retrieve and report a memorized item. The choice of marked rules for the GS model is admittedly somewhat arbitrary, but at least in part guided by introspection as to how humans would perform this task. Further development on rigorous models of multitasking will help to formalize these choices in future work.

The second major challenge for model integration, not to mention model development on the whole, is the setting of parameter values. ACT-R, like similar architectures, has a number of "settable" parameters; however, all parameters have default recommended values that have withstood the test of time in modeling throughout the community. Nevertheless, the original MODS model posed an interesting problem in that it activated several learning mechanisms (e.g., learning of chunk base-level activations and production strengths) that were deactivated in the driver model. Because the MODS model had undergone more rigorous parameter testing with detailed data, it was decided to incorporate its parameter values into the integrated model, thus overriding the driver model's global settings. Fortunately (and perhaps surprisingly), this decision had no apparent adverse effects on the normal operation of the driver model, which proved rather robust to the different parameter settings and activated learning mechanisms.

Model Simulations

The driver model was made to interact with a driving simulation that mimicked the critical elements of the Alm and Nilsson (1995) car-following task. A total of 15 simulations were run: 5 runs in the *No-Task* condition without a secondary task, 5 runs in the *Task-SS* condition with the Single-Step model performing the secondary task, and 5 runs in the *Task-GS* condition with the Group-Step model performing the secondary task. Each simulation generated a detailed behavioral protocol at a rate of roughly 13 Hz including all relevant control and environmental data as well as marks for the start and end of the braking events.

Model Predictions and Empirical Results

We can now compare the model predictions with Alm and Nilsson's (1995) empirical results. It should be emphasized that the present study does not involve typical parameter estimation for fitting the model to data; rather, it centers on *a priori* predictions by simply integrating the models, running simulations, and checking the results. The goal of the study is thus to predict the effects of the secondary task on driver performance primarily in qualitative terms and, secondarily, in quantitative terms as much as possible.

Brake Reaction Time

The first and more important aspect of driver performance examined is drivers' *brake reaction time*, or the time lapse between the start of the lead vehicle's braking and the initiation of braking by the driver. Figure 1(a) shows the reaction times (means and standard deviations) predicted by the model for the No-Task, Task-SS, and Task-GS conditions. While the reaction time for the no-task condition was approximately 2.5 s, the reaction times for both task conditions were significantly higher at roughly 2.9 s, and thus the model predicts a significant impact of the secondary task on drivers' braking reaction.

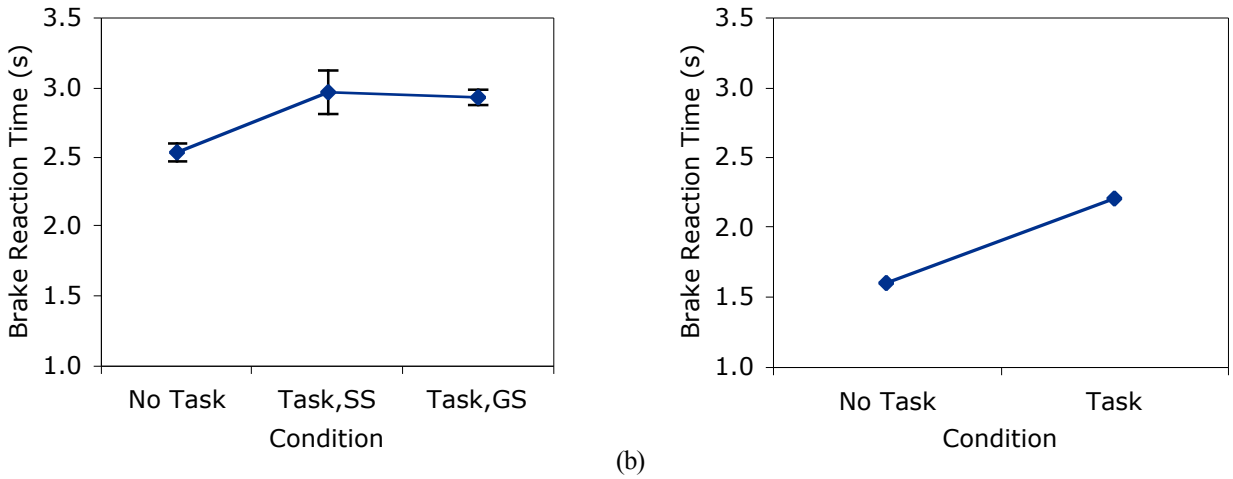


Figure 1: Brake reaction times for (a) model simulations and (b) empirical data.

Figure 1(b) shows the empirical results for brake reaction time. These results also show a clear (and significant) task effect, with an increase of reaction time from 1.6 s without the task to 2.2 s with the task. The model and empirical results therefore correspond well qualitatively. Quantitatively, the model predictions are roughly .7–.9 s too high; this discrepancy may be attributed to the fact that the model uses only distance of the lead vehicle to determines how it accelerates and decelerates, whereas the human drivers could also attend to the lead vehicle's brake lights, providing the latter with additional cues to initiate braking.

Lateral Deviation

The second aspect of performance is one of the most common measures for driver distraction studies, namely the *lateral deviation* of the driver's vehicle — defined as the root-mean-squared error of the vehicle's center with the central position in the lane. Figure 2 shows the model predictions for lateral deviation in the three conditions.

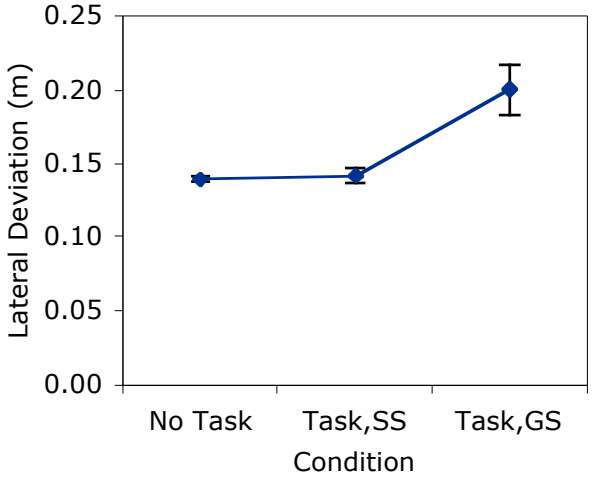


Figure 2: Lateral deviation for model simulations.

Interestingly, the SS model predicts no effect of secondary task on lateral deviation. However, the GS model predicts quite a large effect of approximately 50 cm.

Alm and Nilsson (1995) do not report specific numbers for lateral deviation; however, they do report a statistical analysis on these data that found no significant task effect on lateral deviation (against their original hypothesis). The predictions of the SS model thus support their results, demonstrating how closely interleaved multitasking can, in certain situations, have no significant effect on lateral deviation. On the other hand, the predictions of the GS model show that less conservative, "grouped" multitasking can draw cognitive attention away from the driving task enough to create a significant effect.

Headway Distance

The third aspect of performance is *headway distance*, or the distance between the driver's vehicle and the lead vehicle. While headway was maintained at 75 m during normal conditions, the lead vehicle's braking would greatly reduce this headway until the driver has a chance to react. Figure 3(a) shows the model predictions for the *minimum* headway distance, a measure of how close the driver's vehicle came to the lead vehicle. In all three conditions, the model exhibited a minimum headway of approximately 35 m.

Figure 3(b) shows the minimum distances reported in the empirical study. Here there is a clear task effect: the headway decreases significantly in the presence of the secondary task. Thus, the model predictions are not supported for the distance-headway measure. It seems that although the model clearly reacts later in the task condition, it also compensates for the late reaction by braking harder, thus eliminating any potential task effect.

Conclusions

The SS model's *a priori* predictions matched two of the three measures qualitatively, correctly predicting an effect on reaction time but no effect on lateral deviation. Given the ambiguity in the driver-distraction literature on when such

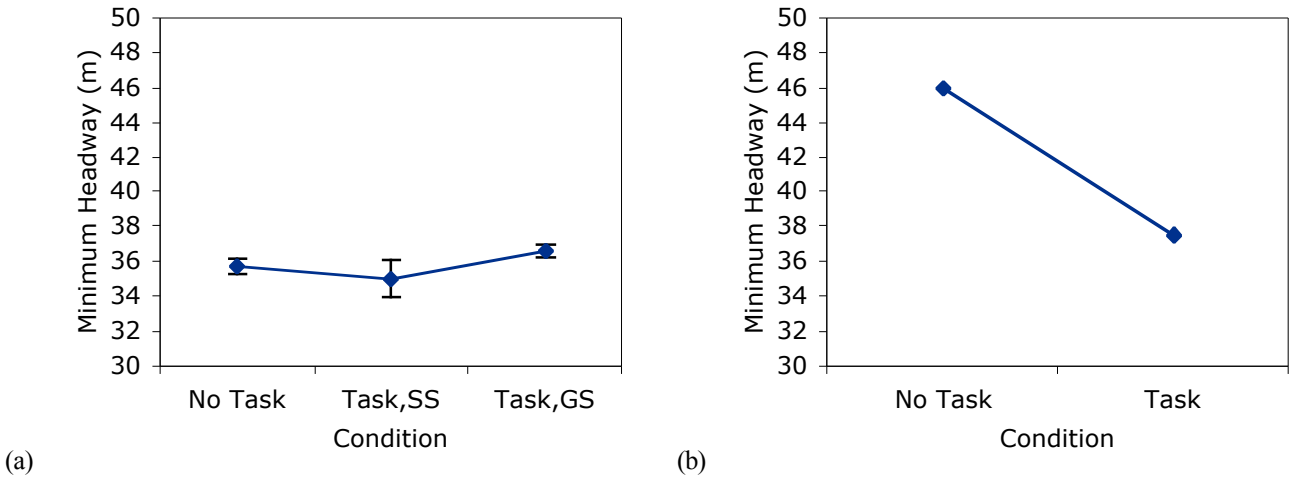


Figure 3: Minimum headway distance for (a) model simulations and (b) empirical data.

effects may occur, this is a strong result that demonstrates the model's ability to predict distraction for certain measures and not others. However, the model did not predict the effect on minimum headway, perhaps due in part to the fact that the headways were large enough that human drivers felt no dire need to closely maintain headway.

The GS model's predictions were not as good, failing to predict the absence of an effect on lateral deviation. Its failing indicates one shortcoming of this work: although the integration of models is mostly straightforward, there remain too many degrees of freedom with respect to how models can and should multitask. Combating this problem requires a more rigorous treatment of multitasking, and cognitive architectures such as ACT-R show promise in being able to account for such a process. In particular, architectures provide an opportunity to handle multitasking at the "software level" through new models implemented as production rules and/or at the "hardware level" through changes to the architecture's inner mechanisms. Recent models of complex dynamic tasks, though not yet the comprehensive models required for the long term, have already demonstrated good ability in capturing some aspects of multitasking (e.g., Chong, 1998; Jones et al., 1999).

As a related point, cognitive architectures also have the substantial benefit of facilitating re-use of models, parameters, and other components from one model to another. This study exhibits this property primarily in the re-use of two existing models for predicting distraction. However, it also opens the door to predicting numerous other aspects of behavior. For instance, Lovett et al.'s (2000) treatment of their MODS model includes an investigation of how ACT-R's *W* parameter can represent individual differences in working memory capacity. Because their work addresses mechanisms fundamental to the architecture, it can carry over directly into further investigations of the effects of capacity differences on driver distraction or even just on driving itself. This ability to share ideas and mechanisms across domains offers enormous explanatory and predictive power to architectural models in new and existing domains of study.

Acknowledgments

Many thanks to Marsha Lovett for providing the MODS model code and to Christian Lebiere, John Anderson, and Lynne Reder for helpful comments and suggestions.

References

Aasman, J. (1995). *Modelling driver behaviour in Soar*. Leidschendam, The Netherlands: KPN Research.

Alm, H., & Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, 27, 707-715.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Chong, R. S. (1998). Modeling dual-task performance improvement: Casting executive process knowledge acquisition as strategy refinement. Doctoral Dissertation, Department of Computer Science and Engineering, University of Michigan.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450-466.

Jones, R. M., et al. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine*, *20*, 27-42.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Journal of Cognitive Systems Research*, 1, 99-118.

McKnight, A. J., & McKnight, A. S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis & Prevention*, *25*, 259-265.

Salvucci, D. D. (2001). Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human-Computer Studies*, 55, 85-107.

Salvucci, D. D., Boer, E. R., & Liu, A. (2001). Toward an integrated model of driver behavior in a cognitive architecture. *Transportation Research Record*, 1779.

Salvucci, D. D., & Macuga, K. L. (2001). Predicting the effects of cell-phone dialing on driver performance. In *Proceedings of the Fourth International Conference on Cognitive Modeling*. Veenendaal, Netherlands: Universal.

The Impact of Problem Order: Sequencing Problems as a Strategy for Improving One s Performance

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Department of Applied Cognitive Psychology and Media Psychology, University of Tuebingen
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Applied Cognitive Science Department, Knowledge Media Research Center
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Abstract

Two experiments investigated the impact of problem order and problem sequencing on performance. In experiment 1 subjects were either presented with a suitable or an unsuitable presentation sequence where they were free to deviate from. Presentation sequence had an impact on performance and rearranging problems improved performance for high prior-knowledge subjects whereas low prior-knowledge subjects' performance deteriorated. Experiment 2 yielded evidence that effects of problem sequence have to be triggered by directing subjects' attention to comparing problems before working on them. Results are discussed within the framework of analogical transfer.

The Impact of Problem Order

In this paper we investigate the impact of problem order on performance when solving a sequence of mathematical problems. It has to be noted that effects of sequencing *to-be-learned materials* have been widely studied in the Sixties and the Seventies (Posner & Strike, 1976; Van Patten, Chao, & Reigeluth, 1986 for an overview), whereas effects of sequencing *to-be-solved problems* have received only little attention. Sequence effects are said to occur when performance on problem *B* varies depending on whether problem *A* had been performed before or not. This influence of solving problem *A* on performance for problem B should be specific to problem *A*, i.e., solving a problem *C* before *B* should not necessarily lead to the same performance for *B* as solving problems in the sequence *AB*. This specificity assumption distinguishes sequence effects from mere training or position effects.

Sequence effects can be analyzed as the result of two distinct cognitive processes that take place in succession, namely, learning and transfer. *Learning* refers to a change in the cognitive system of the problem solver (i.e., newly generated or modified knowledge structures) that occurs due to solving a problem *A*. *Transfer* refers to the transmission of these newly generated or modified knowledge structures to a subsequent problem *B*.

The two most prominent approaches to transfer are Singley and Anderson's rather analytical theory on transfer of cognitive skill (Singley & Anderson, 1989) and the more holistic theories of transfer by analogy (Gentner, 1983; Gick & Holyoak, 1980).

Singley and Anderson's basic assumption is that a problem is more likely to be solved the more declarative and/or procedural knowledge elements necessary to solve that problem are already known by the problem solver. Therefore, transfer among problems should increase with the number of elements being shared by the problems (Thorndike & Woodworth, 1901). Furthermore, because transfer is based on the extent of overlap between the knowledge structures necessary to accomplish two tasks a symmetrical relation between problem *A* and *B* is assumed. (Pirolli & Recker, 1994; Singley & Anderson, 1989). It is important to note, however, that sequence effects may be asymmetrically, i.e., a problem sequence *AB* might result in a different performance than a problem sequence *BA*. This asymmetry is due to the fact that the amount of what has been learned in the first place and can therefore be transferred to a succeeding problem may differ among problems. For instance, working on a difficult problem at first may result in less learning than starting with a simpler problem.

Transfer by analogy is described as the transmission of knowledge from one problem-solving situation (the source) to a target problem and consists in a number of different processes. In order to solve a target problem first a suitable source problem has to be retrieved from memory. Next, elements of the source problem have to be mapped onto the target problem. Finally, based on these mappings a solution for the target problem is generated. Research on analogy has demonstrated that structural similarity among source and target is the most important determinant of successful transfer and that this transfer is often restricted to situations where source and target are structurally equivalent. If there are structural differences between problems subjects often fail to adapt a source problem's solution to fit the requirements of the target (Reed, Dempster, & Ettinger, 1985).

With regard to sequence effects it can therefore be assumed that performance for a specific problem should improve if one solves structurally similar problems in succession. Contrarily, switching between unrelated problems might impede problem solving because this increases the probability that unsuitable preceding problems are used as sources to guide later problem solving.

Whereas there is only preliminary evidence for this assumption concerning structural similarity in a study by Novick (1988), two problem-solving studies have

investigated the effects of the second aforementioned factor that may influence the suitability of a problem sequence, namely a problem's difficulty. Reed, Ernest, and Banerji (1974) obtained no effect of problem order when studying transfer from the easier Missionary-Cannibals problem to the more difficult Jealous-Husbands problem and vice versa. Subjects who had been acquainted with the similarity relations among the problems, however, solved the problems faster in the difficult-easy sequence. Furthermore, Cook (1937) found that a difficult-easy sequence led to better performance when working on pyramid puzzles. Based on these two results it could be argued that solving difficult problems before easier ones should result in better performance than a reversed sequence. However, this may only hold for knowledge-lean problems (in the sense of VanLehn, 1989) whereas for knowledge-rich problems solving an easy problem first may support solving more difficult problems of the same problem category. This should be the case because more difficult problems often share structural elements with the easier problems. Therefore, mastering these problem components in the easier problems provides practice for solving the more difficult problems entailing these components among other new elements. This idea of mastering (subordinate) parts of a skill before proceeding to more difficult demands is in line with proposals made for the design of instructional curricula (cf. Schoenfeld, 1985; Van Patten et al., 1986).

To summarize, problem sequences that are ordered with respect to the structural similarity of the problems (similar problems being solved in succession) and with respect to the difficulty of the problems (easy-to-difficult) should result in better performance compared to either reversed or to random sequences.

Sequencing as a Metacognitive Strategy

In experimental problem-solving settings subjects are usually asked to maintain a given order when solving multiple problems whereas in more self-controlled situations they might be given the opportunity to decide on a problem sequence by themselves. In this case the question whether problem solvers strategically rearrange problems in order to improve their performance gains increasing importance.

Problem sequencing can be seen as a process that is exactly reverted to the retrieval process in analogical problem solving. In analogical problem solving a backward search is conducted to find a source problem in memory whose solution can be adapted to the to-be-solved target. Contrarily, sequencing may be described as a forward search to decide on the next to-be-solved problem (target) for which the solution of the problem being solved most recently (source) can be adapted. Conceptualizing problem sequencing in accordance with the retrieval process in analogical problem solving brings about some major advantages. In particular, findings on analogy may be used to derive hypotheses concerning problem sequencing as a metacognitive problem-solving strategy.

First, the *propensity to sequence problems* should depend on whether subjects are aware of the fact that different problem sequences may be associated with different performance outcomes and that applying knowledge used to solve one problem when approaching a next problem might foster performance. However, research in analogical problem solving has repeatedly shown that subjects often fail in using previous problem-solving experiences spontaneously when solving new problems (Reed et al., 1985) and that they need to be provided with hints in order to ensure analogical transfer (Gick & Holyoak, 1980). Additionally, the costs that result from searching for a suitable target problem that is to be solved next have to be less than the benefits that are achieved by rearranging problems deliberatively. This reasoning is in line with assumptions made by Novick (1988) or Reed et al. (1974) concerning the retrieval process in analogical problem solving. For instance, Reed et al. (1974, p. 448) postulate that "the total time to retrieve, translate, and use analogous information to find an operator should be less than the total time to find the same operator without using information from the previous problem."

Second, the *successfulness of rearranging problems* depends on whether subjects are able to identify a suitable problem sequence by themselves. In terms of analogical problem solving this relates to the question of whether subjects are able to retrieve a source problem that is structurally similar to the target. With respect to this issue research has demonstrated that subjects often face difficulties in recognizing structural problem features and that they are often misled by surface similarities of the problems (Holyoak & Koh, 1987; Ross, 1987). Novick (1988) demonstrated that the ability to retrieve a structurally similar analogue interacts with subjects' domain-specific prior knowledge with experts being more likely to find a suitable source problem than novices are. Therefore, it can likewise be assumed that the quality of problem sequencing might interact with subjects' prior knowledge in a way that only high prior-knowledge subjects benefit from self-determined sequences whereas the additional freedom of rearranging test problems might even be harmful for less advanced subjects.

In order to investigate the impact of problem order and problem sequencing on problem-solving performance two experiments were conducted. In experiment 1 subjects were provided with one of two different presentation sequences that they were free to rearrange. Contrarily, subjects in experiment 2 were confronted with predefined problem orders they could not deviate from in order to find out whether differences in problem-solving performance can still be observed when subjects are not made aware of the potential impact of problem order.

Experiment 1

Method

Participants Subjects were 76 students (49 female, 27 male) of the University of Goettingen, Germany, who participated for course credit or payment. Average age was 22.67 years.

Materials and procedure For experimentation the hypertext-based learning and problem-solving environment HYPERCOMB was used (Gerjets, Scheiter, & Tack, 2000) which contains a short introduction to the domain of combinatorics followed by a learning phase where subjects can acquire knowledge by studying worked-out examples for six problem types.

Permutation problems are about finding out the number of possibilities of bringing all elements of a set into a distinctly ordered arrangement. *Variation problems* deal with the number of possibilities for selecting a subset of elements out of a set of elements in a distinct order. *Combination problems* are about the number of possibilities for selecting a subset of elements out of a set of elements without regard to the order. All three kinds of problems can be further distinguished as being *with* or *without replacement* yielding six problem types. Replacement indicates whether the set contains undistinguishable elements or whether elements can be selected more than once, respectively. Similarity among permutations, variations, and combinations can be described with respect to the number of permutations necessary to solve a specific problem. These similarity relations among the problem types are not only expressed at this conceptual level but are also reflected at the computational level in the graded complexity of the formulas needed to solve the problems. Therefore, one can characterize transfer relations among problem types in combinatorics that are based on the overlap at the computational and conceptual level. According to this task analysis a problem sequence ranging from permutations to variations and ending with combinations should be suited best for problem solving as the problem types that are most structurally similar to each other are presented in succession.

In HYPERCOMB each problem type was illustrated by abstract information concerning its structural features and two worked-out examples. One example explained the basic application of the solution principle and the other example illustrated a more complicated situation where the solution principle in question had to be applied twice in order to solve a problem. Subjects could decide which instructional materials they wanted to study and when they wanted to quit the learning phase. In the subsequent test phase the instructional material was no longer available and subjects were asked to work on six test problems. For those test problems the solution principles which had been taught before had to be applied once for easy problems or twice for difficult problems (figure 1).

When starting the test phase subjects were informed that they would have to solve six test problems listed on a single page. They were asked to study all test problems carefully before selecting a problem they wanted to start working on. Subjects were further informed that they could solve the test problems in any order they wanted. Whenever subjects had solved a problem the initial page with all six problems was presented (including the ones already being solved) and subjects were asked to select the next problem. In order to prevent subjects from solving a problem twice solved problems could no longer be retrieved.

Easy problem: A lighthouse can flash in six different colors (red, yellow, green, blue, orange, pink) from which colors are randomly chosen to form a flare. Each flare contains two colors in succession and none of the colors can appear twice in one flare. What is the probability that the lighthouse will send a red-orange flare, i.e. it will first flash red and then flash orange?

Difficult problem: At a soccer game there are two dressing rooms for the two teams. The kickers from Oxford wear T-shirts with uneven numbers from 1 to 21 and Manchester has even numbers from 2 to 22. As the aisle from the dressing rooms is very narrow only one player at a time can enter the field. The players of the two teams leave their rooms alternately with a player from Oxford going at first. What is the probability that the first five players who enter the field have the numbers five, two, thirteen, eight, and one (i.e., the first has the number five, the second has got the two and so on)?

Figure 1: Easy and difficult test problems of problem type "variation without replacement"

Design and dependent measures As a first between-subjects variable the presentation sequence of the six test problems was manipulated. In the *suitable sequence* the problems were presented in the postulated optimal order - permutation, variation, and combination with an easy-to-difficult sequence within each problem type. In the *unsuitable sequence* variations were followed by permutations and combinations; within problem types difficult problems were presented first. As a second between-subjects variable we used subjects' *domain-specific prior knowledge* which was controlled by means of a multiple-choice questionnaire at the beginning of the experiment. A median split within the two sequence conditions was conducted to distinguish between subjects who possessed low or high prior knowledge.

As performance measures subjects error rates for easy and difficult test problems and problem-solving time were registered. For each of the six test problems subjects had to identify the correct solution principle and the values of four variables in a multiple-choice form. No calculations had to be made. A maximum of two errors was assigned for the identification of the principle and one error was assigned for each wrong answer concerning the variable values resulting in a maximum of six errors for each problem. Additionally, subjects were distinguished as to whether they rearranged problems by deviating from the given presentation sequence or not. Finally, in order to ensure that subjects were equivalent with respect to their learning behavior the example-processing time was registered and analyzed as well.

Results and Discussion

A first comparison by means of an ANOVA (presentation sequence x prior knowledge) revealed no significant differences with regard to either pretest errors ($F(1,72) = 1.29$; $MSE = 83.10$; $p > .25$) or overall example-processing time ($F < 1$) between the presentation sequences (table 1).

In order to analyze subjects' performance on the six test problems and on subjects' problem-solving time as a function of presentation sequence, prior knowledge, and sequencing behavior a third factor was entailed in the

analysis. This factor indicated whether subjects had kept the presentation order while working on the problems or whether they had deviated from it (i.e., sequencing behavior). Additionally, we used example-processing time as a covariate because this turned out to be a very important factor for predicting subjects' performance and because this measure was characterized by a high variability within each of the two presentation sequence conditions. This resulted in a three-factor ANCOVA (presentation sequence x prior knowledge x sequencing behavior) that was deployed for analyzing performance on easy and difficult problems as well as for problem-solving time. We will first report the effects for presentation sequence and prior knowledge (table 1) before having a closer look to the impact of subjects' sequencing behavior on performance (figures 2a, 2b).

Table 1: Performance (in %) and time data (in sec) as a function of presentation sequence and prior knowledge

Presentation sequence	**Suitable sequence**		**Unsuitable sequence**	
Prior knowledge	**High**	**Low**	**High**	**Low**
Pretest errors	44.3	74.1	47.9	75.2
Example-processing time	650	547	612	532
Problem-solving errors:				
- Easy problems	12.0	10.0	16.2	21.8
- Difficult problems	37.5	46.4	43.2	46.7
Problem-solving time	1075	998	1066	1052

Effects of presentation sequence and prior knowledge With regard to the number of problem-solving errors for easy test problems subjects who were presented with the suitable sequence outperformed subjects who worked in the unsuitable sequence condition as predicted ($F(1,72) = 5.02$; $MSE = 244.29$; $p < .05$) whereas there was no effect for difficult test problems ($F(1,72) = 1.50$; $MSE = 371.93$; $p > .20$). None of the effects for prior knowledge nor the interactions between presentation sequence and prior knowledge were significant (all Fs < 1). With regard to problem-solving time there were no effects for either presentation sequence or prior knowledge nor was there an interaction between the two factors (all Fs < 1). To summarize, the superiority of the suitable presentation sequence could be demonstrated for performance on easy problems independently of subjects' prior knowledge.

Sequencing behavior A question that has yet been left unanswered is whether subjects rearrange problems when being confronted with an unsuitable presentation sequence and how their sequencing behavior contributes to problem-solving performance. Analyzing the percentage of subjects who deviated from the presentation sequence an ANOVA (presentation sequence x prior knowledge) clearly revealed that subjects reacted sensitively to the quality of the presentation sequence by deviating more often from the unsuitable sequence than from the suitable one ($F(1,72) = 9.79$; $MSE = 0.23$; $p < .01$; suitable sequence/ high prior knowledge: 21% sequencers; suitable sequence/ low prior knowledge: 35% sequencers; unsuitable sequence/ high prior knowledge: 65% sequencers; unsuitable sequence/ low prior knowledge: 60% sequencers). Sequencing behavior was unaffected by subjects' prior knowledge - with the main effect and the interaction both being meaningless (both Fs < 1). Deviations from the given presentation sequence were mainly caused by subjects' preference to work on easy problems before approaching the more difficult ones - regardless of structural similarities among easy and difficult problems.

Effects of sequencing With regard to the impact of sequencing on performance for easy problems an expected pattern of results could be obtained (figure 2a). There was no main effect of sequencing behavior ($F < 1$), however, sequencing behavior interacted with subjects' prior knowledge in that high prior-knowledge subjects improved by rearranging problems whereas low prior-knowledge subjects' performance even deteriorated ($F(1,72) = 5.20$; $MSE = 244.29$; $p < .05$). Although this effect seemed to interact with presentation sequence the triple interaction was not significant ($F(1,72) = 1.31$; $MSE = 244.29$; $p > .20$), nor was there an interaction between presentation sequence and sequencing behavior ($F < 1$).

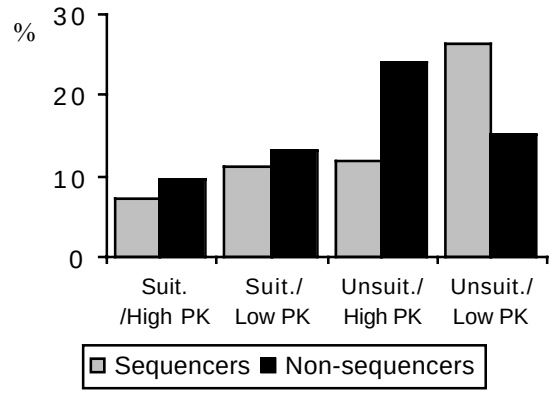


Figure 2a: Problem-solving errors for easy test problems as a function of subjects' sequencing behavior, presentation sequence, and prior knowledge

The effects for problem-solving performance on difficult test problems was different (figure 2b). Performance improved slightly by rearranging problems ($F(1,72) = 2.99$; $MSE = 371.93$; $p < .10$), whereas there were no interactions with prior knowledge or presentation sequence (all Fs < 1).

Additionally, there was no main effect for sequencing behavior on the overall time subjects needed to solve all six test problems ($F < 1$) nor were there any interactions with either presentation sequence or prior knowledge (all Fs < 1.78 and all ps > .15).

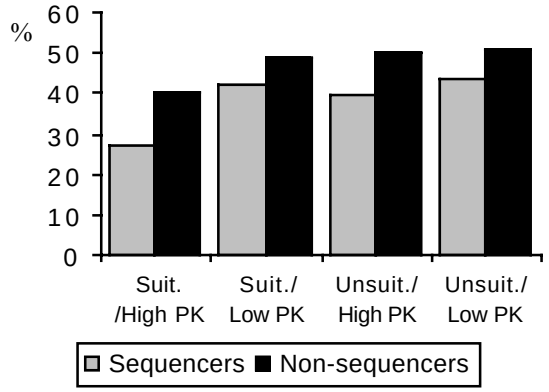


Figure 2b: Problem-solving errors for difficult test problems as a function of subjects' sequencing behavior, presentation sequence, and prior knowledge

To summarize, we found that the order in which problems are solved had an impact on problem-solving performance for *easy test problems* in that a sequence where problems were arranged according to their structural similarity and difficulty was superior to a presentation sequence not making use of this principles. Additionally, we could demonstrate that subjects tried to make use of this effect of problem sequence by rearranging problems when they were presented in an unsuitable way. However, improvements due to problem sequencing were predominant for subjects with high prior knowledge who are more likely to identify structural similarities among problems. On the contrary, low prior-knowledge subjects' performance deteriorated when they deviated from a given order of problems if these problems were unsuitably arranged.

However, the pattern of results obtained for *difficult test problems* yielded evidence for some additional speculations. In particular, sequencing improved performance on difficult test problems independently of whether subjects deviated from a suitable or an unsuitable sequence. Subjects who rearranged problems may have followed the instruction to first read all problems carefully before selecting a problem to work on. This may have focussed subjects' attention on comparing test problems and thereby displaying a deeper processing which in turn improved performance. This interpretation is related to the question on whether subjects spontaneously notice problem similarities by themselves or whether they need hints in order to make use of potential analogues relations among test problems. If the latter is true, sequence effects should only be observable when subjects are asked to process test problems thoroughly as in this experiment, but should be absent when problems are presented in predefined orders without any further instructional support. In order to address this issue a second experiment was conducted.

Experiment 2

Method

Participants Subjects were 78 students (48 female, 30 male) of the University of Goettingen who participated for course credit or payment. Average age was 24.1 years.

Materials and procedure The same learning and problem-solving material as in experiment 1 was used. However, the procedure was varied. The problems were presented in predefined sequences that subjects could not deviate from. Subjects started working on problem 1 in the sequence. After subjects had solved a problem the next problem was automatically presented. Subjects did not see any of the test problems before this automatic presentation. No return to preceding problems was possible.

Design and dependent measures As a first between-subjects variable the *presentation sequence* was varied by presenting the problems according to the same orders as in experiment 1. As a second between-subjects variable subjects' *domain-specific prior knowledge* was used. As performance measures subjects' error rates and problem-solving time were registered. Additionally, example-processing time was measured.

Results and Discussion

A first comparison by means of an ANOVA (presentation sequence x prior knowledge) revealed no significant differences with regard to prior knowledge between the two presentation sequences ($F < 1$). The effect of presentation sequence for example-processing time however almost reached statistical significance ($F(1,74) = 2.49$; $MSE = 1555590.10$; $p > .10$). Therefore, this variable was again used as a covariate in all further analyses (table 2).

Table 2: Performance (in %) and time data (in sec) as a function of presentation sequence and prior knowledge

Presentation sequence	**Suitable sequence**		**Unsuitable sequence**	
Prior knowledge	**High**	**Low**	**High**	**Low**
Pretest errors	46.8	74.7	47.2	70.8
Example-processing time	722	678	539	579
Problem-solving errors:				
- Easy problems	16.1	21.4	15.8	18.1
- Difficult problems	35.9	49.3	42.4	50.8
Problem-solving time	977	897	830	849

There was no main effect for prior knowledge on performance for easy test problems ($F(1,74) = 1.04$; $MSE = 258.55$; $p > .30$), whereas it positively influenced performance on difficult test problems ($F(1,74) = 7.20$; $MSE = 321.39$; $p < .01$). Most interestingly, there were no effects

of presentation sequence and no interactions between the two factors for any of the two performance measures (all Fs < 1). Additionally, presentation sequence had barely no impact on problem-solving time ($F(1,74) = 1.89$; $MSE = 46501.88$; $p > .10$). The main effect for prior knowledge as well as the interaction were not significant (Fs < 1). The interpretation of these results is straightforward. Simply presenting test problems in a suitable order is obviously not sufficient to improve problem-solving performance.

General Discussion

In experiment 1 a problem sequence where problems were arranged according to their structural similarity and their difficulty outperformed a problem sequence where these sequencing principles were reversed. Experiment 2 demonstrated that sequence effects only occurred when subjects were instructed to process problems carefully before working on them. This is in accordance with findings on analogy that spontaneous transfer is hard to achieve. Instead, subjects need hints that relations between problems are important in order to benefit from a suitable sequence.

Additionally, we demonstrated that subjects try to make use of this effect of problem sequence by rearranging unsuitable problem sequences. However, only subjects with high prior knowledge who are more likely to identify structural similarities seem to benefit from problem sequencing. In contrast to that, subjects with low prior knowledge do not seem to possess the skills necessary for identifying a more suitable problem sequence than the one they are initially presented with.

Several issues will be addressed in forthcoming experiments. First, the question arises whether subjects' ability to sequence problems as well as spontaneous transfer within predefined problem sequences can be fostered by deliberately directing subjects' attention to structural similarities of the problems. Second, it is of interest whether other findings of analogy-based research can likewise be transferred to problem sequencing. In particular, we want to investigate whether not only the retrieval process in analogical problem solving but also problem sequencing is vulnerable to effects of superficial similarities among problems. Third, we aim at distinguishing sequence effects that occur due to structural similarity versus sequence effects that are merely caused by the relative difficulty of problems. Additionally, a more-fine grained analysis of subjects' sequencing strategies with regard to this distinction seems promising. The results of experiment 1 provide preliminary evidence that subjects mainly sequenced problems according to their relative difficulty without paying attention to their structural interrelationships. In domains where structural similarities among problems are more evident - like algebra word problems - sequencing behavior may be quite different. Therefore, a series of experiments is currently being conducted using algebra problems.

Acknowledgements

This work was supported by a fellowship of the Deutsche Forschungsgemeinschaft awarded to the first author of the paper as a member of the Graduate College for Cognitive Science at the Saarland University, Germany.

We thank Linda Jost, Carina Kraemer, Frauke Lancker, and Julia Zimball for conducting the experiments as well as Simon Albers for programming work.

References

Cook, T. W. (1937). Amount of material and difficulty of problem solving. *Journal of Experimental Psychology, 20,* 288-296.

Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gerjets, P., Scheiter, K., & Tack, W. H. (2000). Resource-adaptive selection of strategies in learning from worked-out examples. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings from the 22nd Annual Conference from the Cognitive Science Society* (p. 166-171). Mahwah, NJ: Erlbaum.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15*, 332-340.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 510-520.

Pirolli, P., & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235-275.

Posner, G. J. & Strike, K. A. (1976). A categorization scheme for principles of sequencing content. *Review of Educational Research, 46*, 665-690.

Reed, S. K., Ernest, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology, 6,* 436-450.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 106-125.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* San Diego, CA: Academic Press.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Thorndike, E. L. & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8,* 247-261.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. Posner (Ed.), *Foundations of cognitive science*. Mahwah, NJ: Erlbaum.

Van Patten, J., Chao, C.-I., & Reigeluth, C. M. (1986). A review of strategies for sequencing and synthesizing instruction. *Review of Educational Research, 56*, 437-471.

Stochastic Independence Between Recognition and Completion of Spatial Patterns as a Function of Causal Interpretation

Wolfgang Schoppek (wolfgang.schoppek@uni-bayreuth.de)
Department of Psychology, University of Bayreuth
D-95440 Bayreuth, Germany

Abstract

A common view in the research on dynamic system control is that human subjects use exemplar knowledge of system states – at least for controlling small systems. Dissociations between different tasks or stochastic independence between recognition and control tasks, have led to the assumption that part of the exemplar knowledge is implicit. In this paper, I propose an alternative interpretation of these results by demonstrating that subjects learn more than exemplars when they are introduced to a new system. This was achieved by presenting the same material – states of a simple system – with vs. without causal interpretation. If subjects learned exemplars only, then there should be no differences between the conditions and stochastic dependence between various tasks would be expected. However, in an experiment with N=40 subjects the group with causal interpretation is significantly better at completing fragmentary system states and in judging causal relations between switches and lamps, but not in recognizing stimuli as studied. Only in the group without causal interpretation, the contingency between recognition and completion was close to the maximum memory dependence, estimated with Ostergaard's (1992) method. Thus, the results resemble those of other studies only in the condition with causal interpretation. The results are explained by the assumption that subjects under that condition learn and use a second type of knowledge, which is construed as a rudimentary form of structural knowledge. The interpretation is supported by a computational model that is able to reproduce the set of results.

Dynamic system control (DSC) is a paradigm of great interest for applied and basic research likewise. In applied contexts, researchers address questions about how human operators learn to operate new technical systems efficiently, how training should be designed, or what errors operators are likely to commit. In basic research, DSC is one of the paradigms for studying implicit learning. It has been argued that subjects control dynamic systems predominantly with exemplar knowledge about system states, part of which is considered implicit (Dienes & Fahey, 1998). This conclusion was derived from studies with systems characterized by small problem spaces, such as the "Sugar Factory" (a dynamic system with one input and one output variable, connected by a linear equation; Berry & Broadbent, 1988). However, studies with more complex systems have delivered evidence that structural knowledge (i.e. knowledge about the variables of a system and their causal relations) can be more effective for controlling these systems (Vollmeyer, Burns, & Holyoak, 1996; Funke, 1993), although it is not easy to apply and use this type of knowledge (Schoppek, 2002). But even for small systems, the question about what type of knowledge is learned in an implicit manner, is still open. Simulation studies that have proven the sufficiency of exemplar knowledge for controlling the Sugar Factory (Dienes & Fahey, 1995; Lebiere, Wallach, & Taatgen, 1998) have as yet not reproduced effects that point to implicit learning. An example of such effects is the stochastic independence between recognition of system states of the Sugar Factory as studied and performance in one-step control problems, found by Dienes & Fahey (1998). Since exemplar knowledge is typically construed as explicit rather than implicit, it cannot account for these dissociations.

This paper addresses the question if a rudimentary form of structural knowledge is acquired in addition to exemplar knowledge, albeit implicitly or explicitly. The different use of exemplar knowledge and structural knowledge in different tasks can explain dissociations between tasks. The basic strategy for separating the two

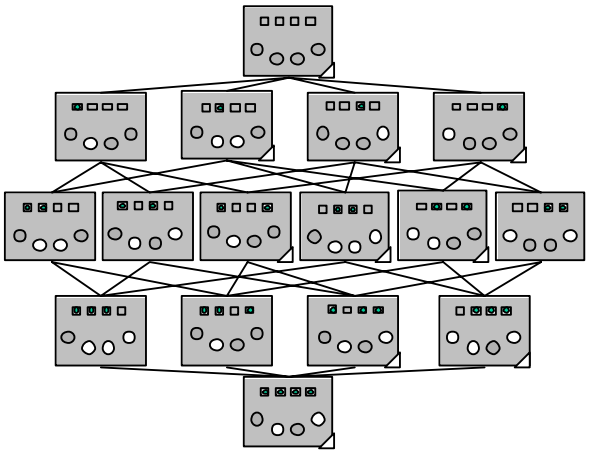


Figure 1: Problem space of the Switches & Lamps system; the states with white triangles were studied in the learning phase

knowledge types rests on using material that can be interpreted as states of a system or simply as spatial patterns. Therefore, I designed a system consisting of four lamps operated by four switches. Each switch affects one or two lamps. Two of the effects were negative, which means that the corresponding lamp is switched off when the switch is turned on. The problem space of 16 possible states is depicted in Figure 1. Subjects under both conditions (causal interpretation vs. no causal interpretation) are shown possible states and asked to memorize them.

In a previous experiment with that paradigm (Schoppek, 2001), I found positive effects of causal interpretation on recognition of patterns as studied and on causal judgment. The effects were attributed to a preliminary form of structural knowledge, namely associations between switches and lamps, acquired by the group with causal interpretation. This knowledge enables subjects to reconstruct a system state in cases where no exemplar representation of the state can be retrieved. The pattern of results was reproduced by a computational model that instantiates these assumptions. The model is written in ACT-R (Anderson & Lebiere, 1998), a cognitive architecture that distinguishes between subsymbolic and symbolic levels of processing, with associative learning residing on the subsymbolic level.

The experiment also delivered some hints that there was stochastic independence between recognition of states as studied and a completion task in the group with causal interpretation, but dependence in the other group. Again, this supports the assumption that more than one knowledge type is used in the causal condition. However, to judge the empirical contingency between tasks, it should be compared with the maximum possible memory dependence, estimated with a method proposed by Ostergaard (1992). This method requires answers to nonstudied items in the completion task, but all items that could be reasonably used in that task (i.e. all possible system states) have been studied in Schoppek (2001). Therefore in the present experiment, subjects studied only a subset of system states. This fact implies a different prediction for recognition of states as studied: The set of possible states and the set of studied states were identical in the previous experiment, whereas they are different in the present experiment. This makes the strategy of reconstructing system states susceptible to errors, because classifying any possible state as studied would result in many false alarms. Thus in the present experiment, I expected no differences in recognition performance between the two conditions.

Experiment

The experiment started with a learning phase where subjects saw 60 system states in intervals of four seconds. The sequence consisted of ten out of sixteen possible system states that were repeatedly shown in a "natural" order, i.e. only one switch changed its status from item to item. The ten states where selected such that all causal relations between switches and lamps could be concluded from them. All subjects were instructed to memorize the "states" (in the condition "causal interpretation" or ci group) or the "patterns" (in the condition "no causal interpretation" or nci group). The learning phase was followed by a speeded recognition task. 20 items, including the ten studied states, six nonstudied states, and four impossible states, had to be classified as studied or nonstudied. Next, subjects worked on the completion task, where they saw arrays of switches in certain states and were asked to complete the patterns by clicking on the correct lamps. All possible states, except the one where no switch is on, were administered. Then the subjects of the group without causal interpretation were debriefed about the meaning and the causal nature of the material. Finally, in a causal judgment task, subjects were asked to estimate the causal strength of all 16 combinations between switches and lamps on a scale ranging from -100 (strong negative relation), through 0 (no relation), to 100 (strong positive relation). N=42 students from the University of Bayreuth, participated in the experiment. One subject had to be excluded because of erroneous administration of the tasks; one other subject was excluded because he had misunderstood the instructions.

I expected medium to large effect sizes ($d \approx 0.65$) in this experiment. With the given sample size of n=20 for each group, the α-level is set to $p<0.1$ to get an acceptable power of 0.67. All significance tests were two-tailed[1].

Recognition

I expected no differences in discrimination between the two groups. This can be explained as follows. For the nci group, conditions are not much different to the previous experiment (Schoppek, 2001), except that fewer states where shown and each state was shown equally often. In the ci group, however, the fact that not all possible system states were shown in the learning phase is expected to lead to some confusion. Subjects who know about the causal structure of the material may recognize nonstudied system states as regular states and mistake them as studied. Thus, in contrast to the previous experiment, there is no advantage of knowing the causal structure. It is hard to predict if subjects use the strategy of reconstructing system states at all. An indicator for using the strategy is a longer response time.

[1] The power analysis was calculated with the G-Power program by Faul & Erdfelder (1992).

As expected, discrimination indices for recognition (calculated according to the two-high-threshold model by Snodgrass & Corvin, 1988) are almost equal in both groups (ci: d=0.46, s=0.19; nci: d=0.43, s=0.17; t=0.53). However, mean response times rt for hits are significantly longer in the ci group (ci: rt=2325 ms, s=1159 ms; nci: rt=1699 ms, s=513 ms; t=2.21, p<0.05). This result, including the difference in the standard deviations, closely replicates the findings of Schoppek (2001). It supports the assumption that at least some of the subjects in the ci group used the strategy of reconstructing system states on the basis of structural knowledge.

Completion

Since all possible system states had to be completed in this task, I expected the ci group to be better than the nci group. Subjects in the latter group have only a small chance to complete nonstudied items correctly.

Performance in the completion task is measured by summing up deviations from the correct solution over all items (variable td). For each lamp, a deviation is counted when the lamp is in the wrong state, resulting in a maximum deviation of four per item. Thus, the total deviation td ranges between 0 and 60 (4 · 15 items). The expected deviation for chance performance is 30 (0.5 · 4 · 15). As expected, there is a significant difference in total deviation between the groups: The ci group deviates less from the correct solutions than the nci group (ci: td=21.9, s=6.7; nci: td= 25.3, s=4.6; t=1.88, p<0.1). Generally, performance in the completion task was low: In terms of correct items, the ci group solved an average of 3.9 items (26%), the nci group an average of 2.9 items (19%). However, these values are close to those found by Dienes and Fahey (1998) in their one-step control problems with the Sugar Factory.

Causal Judgment

Subjects of the ci group are expected to be much better in judging causal relations between switches and lamps. At first glance, this hypothesis appears straightforward. However, if causal knowledge is learned implicitly in the form of associations between switch-events and lamp-events, it is possible that subjects of the nci group are able to judge some of the relations after they have been debriefed about the causal nature of the material.

As a measure for causal judgment, the median of the 16 absolute deviations between judgments and correct answers was calculated (variable md) for each subject. The ci group was significantly better at judging the causal relations between switches and lamps (ci: md=27.9, s=31.1; nci: md= 64.7, s=25.1; t=3.91, p<.01). This result makes it unlikely that many of the nci subjects had learned associations between switches and lamps implicitly.

Contingency analysis between recognition and completion task

If subjects used exemplar knowledge only, we expect performance in the two memory tasks to be correlated. If, however, subjects used exemplar knowledge and structural knowledge, performance in the two tasks may well be independent from each other. To judge the contingency between two memory tasks, Ostergaard (1992) has proposed a method for estimating the maximum possible memory dependence for a given data set. The method is based on the contingency tables crossing the answers in both tasks. Stochastic independence is shown when there is a significant difference between appropriate measures of the observed contingency and the contingency assuming maximum memory dependence.

The contingency analysis was applied separately for each subject, yielding distributions of observed and estimated values of the joint probability of giving a correct response to both tasks, and of the contingency measure Δp. Analyses with the data collapsed over all

Table 1: Overview over results of the experiment

	Causal interpretation (ci)	No causal interpretation (nci)	Significance
Recognition			
discrimination index	0.46	0.43	ns
response time for hits	2325 ms	1699 ms	**
Completion			
total deviation	21.9	25.3	*
Causal judgment			
median of deviation	27.9	64.7	***
Correlation			
completion – causal judgment	.62***	.21	

Significance levels: *: p<0.10 **: p<0.05 ***: p<0.01

subjects of each condition were conducted to cross-check the results. Both analyses yielded equivalent results.

In the ci group, the observed joint probability of giving a correct response to both tasks equals 0.31, a value lying right between 0.27, the joint probability of the independence model and 0.34, the joint probability of the maximum memory dependence (MMD) model. Although the absolute difference between the value for the MMD and the observed value is rather small, it is still reliable ($t(19)=2.41$, $p<0.05$). The contingency measure Δp also discriminates between the different models. The $\Delta p = 0.22$ observed in the ci group is significantly smaller than the $\Delta p = 0.37$ of the MMD model ($t(18)=2.26$, $p<0.05$).

Things are different in the nci group, where the joint probabilities of the observed data and the MMD model are 0.23 and 0.24, respectively ($t(19)=0.53$, $p=0.60$). The difference between $\Delta p=0.22$ (observed) and $\Delta p=0.23$ (MMD model) is not significant either ($t(19)=0.11$, $p=0.92$).

The result that in the ci group the observed contingency between recognizing states as studied and completing fragments of these states correctly is significantly below the maximum, indicates that different memories have been used for both tasks. In the nci group, the observed contingency between the tasks is almost at its theoretical maximum, indicating that only one type of knowledge was used for answering the items. The interpretation of these results is that both groups use exemplar knowledge in both tasks, but that subjects of the ci group also use structural knowledge, especially in the completion task. This conclusion is supported by different correlations between measures of causal judgment and completion, which are $r=0.62$ ($p<0.01$) in the ci group, and $r=0.21$ (ns) in the nci group.

Discussion

The present experiment confirmed predictions about the differential impact of causal interpretation on memory for states of a simple system. In part, these predictions were derived from a computational model that formalizes a set of assumptions about acquisition and use of two types of knowledge. Exemplar knowledge about system states is assumed to be acquired and used in all tasks, regardless of causal interpretation. With causal interpretation, subjects can additionally learn structural knowledge based on associations between switch events and lamp events (Schoppek, 2001). This knowledge can be used to reconstruct system states in cases where no relevant exemplar can be retrieved from memory. For reasons described above, this type of knowledge was expected to be useful in a causal judgment task and a fragment completion task, but not in a recognition task, resulting in stochastic independence between recognition and completion in the condition with causal interpretation.

This approach has much in common with implicit learning paradigms. Similar to those paradigms, subjects are presented with material based on a structure they do not know. In contrast to many implicit learning experiments, subjects of the nci group of the present experiment did not learn much about that structure (see the results of the causal judgment task). However, the view that structure is always learned implicitly, as soon as there is one, is not unchallenged. Wright and Whittlesea (1998) argue against the hypothesis that implicit learning is passive and independent of the intentional processes during learning. According to them, this is a misconception resulting from the fact that in most implicit learning experiments there is little or no variation in the learning phases. Wright and Whittlesea provided evidence that even small variations in the presentation of stimuli, or in the induction task can result in differences of what is learned implicitly. Causal interpretation can be viewed as one of these variations that affects processing in the learning phase.

Other examples of the effect that providing additional information about stimuli enhances memory or other kind of performance are found in classification learning (Nosofsky, Clark, & Shin, 1989) or schema acquisition (Ahn, Brewer, & Mooney, 1992). Common to all these examples is subjects' reluctance to use the additional hints. Ahn et al.'s (1992) subjects used the experimentally provided background knowledge only when they were engaged in tasks requiring the active use of that knowledge. Nosofsky et al. (1989) found that even simple rules defining a concept were only used when subjects were explicitly told to do so.

In the group with causal interpretation, the results resemble those typically found in implicit learning experiments. So does the stochastic independence in that group indicate implicit learning? It is not a new claim, but still useful to analyze the acquisition processes, the knowledge resulting from these processes, and the retrieval processes separately (Frensch, 1998), rather than calling the whole thing "implicit learning". Doing so in the present context results in a detailed web of hypotheses. According to the ACT-R model, the processes for acquiring associations between switches and lamps can be characterized as implicit, because associative learning is an autonomous process that occurs without awareness. That does not mean that it is independent from attentional processes. In fact, what associations are learned depends on the sequence in which perceptional or memory elements are focused on. In the Switches & Lamps System, the condition for acquiring useful associations is a processing sequence that focuses on the changes first (i.e. encode the switch

that has changed since the last item, then encode the lamps that have changed, then encode the rest). The assumption that such a sequence occurs more likely in the ci group, whereas in the nci group, subjects adopt other strategies such as processing the images from top left to bottom right, is plausible, although it was not tested empirically. When the critical difference between ci and nci groups lies in the processing sequence of stimuli, one can conclude the testable prediction that differences between the groups should disappear when nci subjects are instructed to focus on changes and are debriefed after the learning phase.

These deliberations are well in line with the view of Wright and Whittlesea (1998), who claim that "the only major difference between implicit and explicit learning may be that consciously knowing that a domain possesses some important structural property can cause one to learn specifically about that property, whereas the processing performed when unaware that such a property exists may focus selectively on less relevant properties" (p. 419).

As a form of subsymbolic knowledge, associations can be viewed as implicit knowledge. In ACT-R, subsymbolic knowledge exerts its influence through activation processes, but is not directly accessible by production rules. The explanatory potential of the subsymbolic level of ACT-R for implicit memory phenomena has also been demonstrated by Taatgen (1999) with a model of word recognition and completion. In his model it is the dynamics of baselevel learning rather than associative learning that accounts for dissociations.

Only at the stage of applying the knowledge a conscious strategy of utilizing the associations between switch events and lamp events is assumed, a strategy of retrieving the most active lamp event with a given switch-turned-on event as cue.

Since the system I used here was a static one, some considerations about the generalization of the results to dynamic systems are indicated. Dynamic systems are characterized by dependence on their own state, which gives them momentum. This is not the case in the Switches & Lamps System. However, similar to dynamic systems, its output variables depend on multiple input variables. The momentum is an important property that makes it hard to handle dynamic systems (Funke, 1993). This might be one of the reasons why subjects typically focus on the relations between input and output variables, often disregarding the output-output relations that establish the momentum (Schoppek, 2002). Thus, from the point of view of many subjects, the Switches & Lamp System can appear very similar to small dynamic systems like the Sugar Factory.

If one accepts "Switches & Lamps" as a model for small dynamic systems, the work presented here questions the common interpretation that controlling those systems is accomplished with exemplar knowledge only (Dienes & Fahey, 1995; Lebiere, Wallach & Taatgen, 1998). For obvious reasons, proving the sufficiency of this type of knowledge does not prove that human subjects are making do with this type, too. The findings of the group with causal interpretation parallel those of Dienes and Fahey (1998), who found stochastic independence between recognition and a completion task and arrived at similar conclusions. The present experiment extends Dienes and Fahey's approach by demonstrating that without causal interpretation the contingency between these tasks is close to its possible maximum, indicating that in that case only one type of knowledge is used. Moreover, it involves a real dissociation in the sense that the experimental manipulation affected one task (causal judgment, completion), but not another (recognition). It would be interesting to see if a variation of causal interpretation with the Sugar Factory yielded similar results.

Although many of the predictions were derived from a cognitive model, I have as yet not succeeded in reproducing the whole set of results with the model. For example, the present model overestimates discrimination between old and new states. The main reason for this is the simplified assumption that every state is encoded by three chunks in a one trial fashion: One chunk representing all switches, one representing all lamps, and one grouping the two other chunks together. This assumption has to be replaced by an appropriate theory about how humans form chunks from unfamiliar material, such as the competitive chunking theory (Servan-Schreiber & Anderson, 1990), or EPAM successors like CHREST (Gobet & Jackson, 2001). Nevertheless, even when a model does not reproduce all aspects of the data, the cognitive modeling perspective forces the analyst to explicate assumptions on all stages of processing, thus helping to draw a detailed picture of reality that goes far beyond the simple distinction between implicit and explicit learning.

Acknowledgments

I would like to thank Wayne Gray for valuable hints regarding the paradigm, and Nha-Yong Au for carrying out the experiment.

References

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 391-412.

Anderson J.R., & Lebiere C. (1998). *The atomic components of thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

Berry D.C., & Broadbent D.E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79,* 251-272.

Dienes Z., & Fahey R. (1995). Role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 848-862.

Dienes Z., & Fahey R. (1998). The role of implicit memory in controlling a dynamic system. *The Quarterly Journal of Experimental Psychology, 51A*, 593-614.

Faul, F. & Erdfelder, E. (1992). GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS [Computer program]. Bonn, FRG: Bonn University, Dep. of Psychology. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Frensch, P. A. (1998). One concept, multiple meanings: On how to define the concept of implicit learning. In M. A. Stadler & P. A. Frensch (Eds.), *Handbook of implicit learning*, pp. 47-104. London: Sage Publications.

Funke J. (1993). Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results. In G. Strube & K.F. Wender (Eds.), *The cognitive psychology of knowledge*, pp. 313-330. Amsterdam: North-Holland.

Gobet, F. & Jackson, S. (2001). In search of templates. In E.M. Altmann; A. Cleeremans; C. D. Schunn & W. D. Gray (Eds.), *Fourth international conference on cognitive modeling*. (pp. 97-102). Mahwah, NJ: Lawrence Erlbaum Associates.

Lebiere C., Wallach D., & Taatgen N. (1998). Implicit and explicit learning in ACT-R. In F.E. Ritter & R. M. Young (Eds.), *Proceedings of the Second European Conference on Cognitive Modelling (ECCM-98)*, pp. 183-189. Nottingham: Nottingham University Press.

Marescaux P.-J., Luc F., & Karnas G. (1989). Modes d´apprentissage selectif et nonselectif et conaissances acquises au controle d´un processus: Evaluation d´un modele simule. *Cahiers de Psychologie Cognitive, 9,* 239-264.

Nosofsky, R.M., Clark, S.E. & Shin, H.J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282 -304.

Ostergaard, A. L. (1992). A method for judging measures of stochastic dependence: Further comments on the current controversy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 413-420.

Schoppek, W. (2002). Examples, rules, and strategies in the control of dynamic systems. *Cognitive Science Quarterly*, *2*, 63-92.

Schoppek, W. (2001). The influence of causal interpretation on memory for system states. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 904-909, Mahwah: Erlbaum.

Servan-Schreiber, E. & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592-608.

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34-50.

Taatgen, N. (1999). *Learning without limits*. Groningen, The Netherlands: Universal Press of the Rijksuniversiteit Groningen.

Vollmeyer R., Burns B.D., & Holyoak K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20,* 75-100.

Wright R. L, & Whittlesea, B. W. A. (1998). Implicit learning of complex structures: Active adaptation and selective processing in acquisition and application. *Memory & Cognition*, *26*, 402-420.

Designing Sets of Instructional Examples to Accomplish Different Goals of Instruction

Tina Schorr (t.schorr@iwm-kmrc.de)
Virtual Ph.D. Program: Knowledge Acquisition and Knowledge Exchange with New Media
Konrad-Adenauer-Str. 40, D-72072 Tuebingen (Germany)

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Dept. of Applied Cognitive Science, Knowledge Media Research Center
Konrad-Adenauer-Str. 40, D-72072 Tuebingen (Germany)

Katharina Scheiter (k.scheiter@imw-kmrc.de)
Dept. of Applied Cognitive Psychology and Media Psychology, Institute of Psychology
Konrad-Adenauer-Str. 40, D-72072 Tuebingen (Germany)

Yiannis Laouris (laouris@cyber.cy.net)
Cyprus Neuroscience and Technology Institute
Promitheos 5, 1065 Nicosia (Cyprus)

Abstract

In this paper we discuss issues of instructional design with regard to different goals of instruction in the context of learning from examples. Two different approaches for identifying suitable instructional methods are considered: First, a cognitive task analysis is presented that examines problem-solving strategies applicable for solving mathematical word problems from a cognitive-modeling perspective. The second approach is based on a review of empirical findings on designing instructional examples. Together, these considerations lead to the selection of two instructional methods that are expected to foster learning with respect to different goals of instruction. This assumption is tested in two experimental studies presented in this paper.

Problem-Solving Strategies for Mathematical Word Problems: A Cognitive Task Analysis

When considering different options of designing instructional material it is important to take into account which instructional goal is to be accomplished. For example, in the case of learning from examples, this goal could consist in enabling a learner to solve problems that are either rather similar or dissimilar to the instructional examples. For solving such problems a learner needs to possess specific prerequisites (e.g., knowledge, time to perform different strategies). In order to enable the learner to acquire these prerequisites by studying instructional material, an instructional designer needs a precise conception of the prerequisites that are needed and therefore have to be imparted by the material. In order to acquire information on how problems in question can be solved and which prerequisites are therefore needed it is helpful for an instructional designer to perform a cognitive task analysis. This can be done by using cognitive modeling as a method of task analysis, i.e., by constructing computer models to simulate a human problem-solver's behavior based on cognitive theories (Gray & Altmann, 2001).

Cognitive models are characterized by a high precision that is achieved by the necessity of stating explicit and formalized assumptions in order to get running computer models (Zachary, Ryder, & Hicinbothom, 1998). Because of its high precision a cognitive modeling approach allows for a detailed comparison of different problem-solving strategies (Rittle-Johnson & Koedinger, 2001) as well as for the derivation of instructional methods based on the theoretical assumptions which are specified in cognitive models (Pirolli, 1999).

In the cognitive task analysis presented in this paper we examine two different problem-solving strategies applicable for solving mathematical word problems. A strategy can be characterized as a conditional sequence of subgoals and operators that is suitable to achieve a particular goal (cf. Pirolli, 1999, p. 452). This corresponds to the way strategies are usually represented in cognitive models.

Generally, problem-solving strategies can be classified as search-based, example-based, or schema-based, respectively. Search-based strategies like means-end analysis are appropriate to solve puzzle problems like the Tower of Hanoi (cf. Newell & Simon, 1972) which do not presuppose domain-specific prior knowledge (knowledge-lean problems according to VanLehn, 1989). However, more complex tasks (like solving mathematical word problems) require example-based or schema-based strategies that operate on a rather elaborated knowledge base. Example-based strategies use concrete knowledge about example problems and their solutions. Within this group of example-based strategies different strategies can be distinguished that vary in the extent they make use of example information. Compared to example-based strategies, schema-based strategies use more abstract knowledge representable as generalized, automated problem-solving schemata (e.g. Sweller, van Merrienboër, & Paas, 1998). According to VanLehn (1989, p. 545) a schema consists of „information about the class of problems the schema applies to and information about their solutions". Example-based and schema-based strategies correspond to two main

approaches in cognitive science, that is the similarity-based and the rule-based approach (Hahn & Chater, 1998).

In our task analysis we examined two strategies for solving mathematical word problems (i.e., the *keyword-strategy* and the *situation model-strategy*) by formalizing them as executable computer models within the framework of the ACT-R-architecture (Anderson & Lebiere, 1998). In the following paragraphs the two strategies will be outlined according to their subgoal structures.

- The *keyword-strategy* is an example-based strategy that uses concrete knowledge about examples when working on new problems (cf. Sowder, 1988). The strategy is characterized by *bottom-up* processing based on the mechanism of *principle-cueing* (Ross, 1987). The respective ACT-R-model starts with reading a given word problem in order to reach the top goal of solving it. While reading, a text phrase that contains certain keywords (or very similar expressions) can activate these keywords in memory. This activation process in turn may lead to the activation of those examples in memory that contain the respective keywords. These known examples will be retrieved if their activation is sufficiently high (*remindings* according to Ross, 1989) and will then be used to solve the current word problem by applying the same procedure to it that was used in order to solve the examples.
- The *situation model-strategy* is a schema-based strategy that operates on a more abstract and elaborated knowledge base and relies on *top-down* processing (cf. Reusser, 1990). The ACT-R model of this strategy again starts with reading, but at the same time it interprets a given mathematical word problem. On basis of this interpretation-process a situation model can be constructed which represents the situation described in the text in a compressed form (cf. Kintsch, 1998). This situation model is then interpreted in a mathematical fashion by matching it with domain-specific schemas representing different problems categories and their appropriate solutions. Thus, in this strategy a given word problem can be solved by applying the solution specified in a known schema that is selected and instantiated on basis of the situation model of the word problem.

A *comparative evaluation* of the two strategies by applying their cognitive models to solving word problems yielded the following results: The keyword-strategy is convenient for equivalent test problems. These are characterized by a near transfer distance because they belong to the same problem category as the instructional examples and are embedded within the same cover story (Reed, 1999). In order to solve such problems by using the keyword-strategy only a very limited amount of problem-solving time and a small knowledge base, mainly containing superficial keywords, are necessary prerequisites. For the application of the situation model-strategy on the other hand, more prerequisites are needed, in particular more time for problem-solving and a larger and more elaborated knowledge base. This base includes schemas for problem categories and knowledge on structural features of problems that determine the appropriate solution procedure. These higher demands with regard to time and knowledge, however, are accompanied by a good problem-solving performance in isomorphic test problems. These are characterized by an intermediate transfer distance, because they belong to the same problem category as the instructional examples, but are embedded within different cover stories (Reed, 1999).

To conclude, the cognitive task analysis demonstrates computationally that the goal of solving a mathematical word problem can be reached by applying different problem-solving strategies. Thereby, these strategies differ in their processing steps as well as in their prerequisites. In the cognitive task analysis, the latter were specified in terms of problem-solving time and knowledge. However, there are also differences between the strategies with regard to their appropriateness in the context of a specific instructional goal — like solving equivalentversus isomorphic word problems. For equivalent problems, the keyword-strategy is convenient whereas the situation model-strategy is more appropriate in the case of isomorphic problems. Therefore, if the goal of instruction consists in enabling learners to solve equivalent problems, the prerequisites for the keyword-strategy should be imparted by the instructional material. On the other hand, if learners are supposed to learn how to solve isomorphic problems, acquiring prerequisites for the situation model-strategy should be the focus for designing instructional materials. Hence, for an instructional designer it is important to consider the instructional goal in order to ensure that the respective prerequisites are imparted in the materials.

When possessing a precise concept of the prerequisites needed in the context of a specific instructional goal the next question that is to be answered by an instructional designer is how these prerequisites can be imparted by the instructional material. Instructional methods that are suitable in the context of the prerequisites of the two specified strategies may be identified by reviewing empirical findings on designing instructional examples.

Instructional Design: Learning from Examples

The advantages of using examples as learning material have been pointed out in many studies (for an overview cf. Atkinson, Derry, Renkl, & Wortham, 2000). It could be demonstrated that studying examples is of great help for knowledge acquisition and that especially multiple examples support schema induction (e.g., Quilici & Mayer, 1996). However, learners may also experience difficulties when learning from examples. In particular, Ross (1987) found evidence that learners face problems in discriminating between *structural features* of an example (which determine its solution procedure) and *superficial features* describing the example s cover story (which are irrelevant with regard to its solution). Some attempts have been made to counteract such difficulties by improving the design of instructional examples. In order to foster the acquisition of structural features by means of instructional design Atkinson et al. (2000) distinguish between modifying intra-example features, i.e., features concerning the format of a single example, and varying inter-example features, i.e., features related to combinations of multiple examples. An instructional method that bears on inter-examples features and that is supposed to direct learners attention to structural features is the utilization of so-called *structure-emphasizing example combinations* which are contrasted with *surface-emphasizing example combina-*

tions that guide learners attention towards superficial features (Quilici & Mayer, 1996). Both types of example combinations are based on imparting knowledge on multiple problem categories which are each illustrated by multiple examples. In the case of structure-emphasizing example combinations each example of a particular problem category is embedded within a different cover story whereas the same set of cover stories is used across problem categories (Table 1, left). Surface-emphasizing example combinations on the other hand, are characterized by the fact that all examples of a particular problem category are embedded within the same cover story which varies across different problem categories so that problem categories and cover stories are confounded (Table 1, right).

Table 1: Structure-emphasizing (A) and surface-emphasizing (B) example combinations

(A)

Cover story (CS)	*Problem category (PC)* PC1	PC2	PC3	PC4
	CS1	CS1	CS1	CS1
	CS2	CS2	CS2	CS2
	CS3	CS3	CS3	CS3
	CS4	CS4	CS4	CS4

(B)

Cover story (CS)	*Problem category (PC)* PC1	PC2	PC3	PC4
	CS1	CS2	CS3	CS4
	CS1	CS2	CS3	CS4
	CS1	CS2	CS3	CS4
	CS1	CS2	CS3	CS4

Quilici and Mayer (1996) asked their subjects to categorize new problems after they had studied examples that were either presented as structure-emphasizing or as surface-emphasizing example combinations. A clear superiority of structure-emphasizing example combinations as learning material could be demonstrated for this categorization task.

Structure-emphasizing and surface-emphasizing example combinations seem to be appropriate instructional methods for imparting the prerequisites necessary to apply the two problem-solving strategies discussed. Structure-emphasizing example combinations should help to acquire structural problem features that are required to apply the situation model-strategy which is appropriate for isomorphic problems. On the other hand, surface-emphasizing example combinations should foster the acquisition of (surface-based) keywords that can be used to apply the keyword-strategy that is suitable for equivalent problems. To test this idea we conducted a series of experiments that differ from the studies of Quilici and Mayer (1996) in several respects.

Using an example-based hypertext environment Quilici and Mayer (1996) conducted paper-pencil experiments without measuring the time that was needed to learn with the different example combinations. On basis of our cognitive task analysis we assume that the two sets of example combinations differ with regard to their time demands. The cognitive task analysis supposes an elaborated knowledge base for the situation model-strategy that rests upon structural features and schemata for problem categories. It is expected that the acquisition of such a knowledge base from structure-emphasizing example combinations demands complex cognitive processes. These processes may require more time investment than processes applied to surface-emphasizing example combinations that result in the acquisition of surface-based keywords. To test this hypothesis we implemented our experiments as computer-based experiments using a hypertext system for learning and problem-solving that allows for the concurrent automatic registration of time spent on each page visited by means of logfiles. As a result, average learning times for the two instructional methods can be determined and compared to each other.

Extending the application area to children The results of Quilici and Mayer (1996) are based on adult subjects. However, the examination of the proposed instructional methods is particularly interesting with children as subjects. The reason is that surface-emphasizing example combinations are a common learning material in mathematical school books regardless of the instructional goals. Therefore, instructional implications will immediately arise if structure-emphasizing example combinations proof superior in the context of a particular learning goal.

Using problem-solving tasks as test problems Whereas Quilici and Mayer (1996) tested the performance of their subjects mainly by administering categorizing tasks, we use problem-solving tasks that vary with regard to their transfer distance. The reason for this modification lies in the assumption that the superiority of an instructional method depends on the goal of instruction that is related to a particular transfer distance. Surface-emphasizing example combinations may enable the keyword-strategy that is appropriate for solving equivalent test problems. Structure-emphasizing example combinations may allow for the situation model-strategy that is superior in solving isomorphic problems. Because it is assumed that the prerequisites for the keyword-strategy are imparted by surface-emphasizing example combinations and the prerequisites for the situation model-strategy are facilitated by structure-emphasizing example combinations, the following results are expected: Studying surface-emphasizing example combinations has a positive impact on solving equivalent problems whereas learning with structure-emphasizing example combinations fosters the ability to solve isomorphic problems. But both instructional methods may be not very helpful if the instructional goal is to solve problems with a far transfer distance, i.e., novel problems that are characterized by a different cover story and by a different problem category compared to the instructional examples (Reed, 1999). This should be the case because both instructional methods do not impart flexible knowledge that is needed to solve novel problems.

Hypotheses

Based on the results of the cognitive task analysis as well as on the review of empirical findings on designing instructional examples we derived three experimental hypotheses about the use of structure-emphasizing and surface-emphasizing example combinations as instructional material. (1) *Time demands:* Learners using structure-emphasizing example combinations for learning need more time to study compared to learners using surface-emphasizing example combinations. (2) *Differential effectiveness:* Learners using structure-emphasizing example combinations perform better

on isomorphic problems as learners using surface-emphasizing example combinations, whereas the latter show better problem-solving results when working on equivalent problems compared to learners using structure-emphasizing example combinations. (3) *Far transfer distance:* There is no performance difference between learners using structure-emphasizing and learners using surface-emphasizing example combinations when working on novel problems.

To investigate these hypotheses we conducted two experiments using the experimental environment BASICOPERATIONS that is described in the following section.

Experimental Environment

The hypertext environment BASICOPERATIONS used for experimentation is based on the hypertext-system HYPERCOMB developed by Gerjets, Scheiter, and Tack (2000). BASICOPERATIONS deals with the domain of basic arithmetic operations and is divided into a learning and a test phase.

In the learning phase, a learner is presented with 16 worked-out examples one after another in a fixed order. Four different problem categories are illustrated by four worked-out examples each. A problem category is formed by the conjunction of two different basic operations; illustrated are the problem categories (PC1) addition/multiplication, (PC2) subtraction/multiplication, (PC3) addition/division and (PC4) subtraction/division. However, another classification of the worked-out examples can be made with regard to their cover stories. Each of the 16 examples is embedded within one of four different cover stories whereby each cover story is used in four examples. The cover stories deal with (CS1) a family on a hiking trip, (CS2) a girl getting money, (CS3) a school arranging a sports meeting and (CS4) a boy buying food. The presentation of the instructional examples is blocked according to the problem categories in a predefined sequence, i.e., all four examples of one problem category are presented subsequently before the next problem category is illustrated.

Two different versions of BASICOPERATIONS (german version on the web: *karibik.cops.uni-saarland.de/knac/ex2/zypernA_deutsch* and *../zypernB_deutsch*) were used that can be classified as providing structure-emphasizing or surface-emphasizing example combinations according to the manipulation of Quilici and Mayer (1996). In the version with structure-emphasizing example combinations all four examples illustrating one particular problem category differ in their cover stories (cf. Table 1, left); in the version with surface-emphasizing example combinations all examples illustrating one particular problem category are embedded within the same cover story (cf. Table 1, right).

In the test phase, 18 word problems had to be solved one after another in a fixed order. The instructional example combinations were no longer available during testing. According to their transfer distance with regard to the instructional examples presented in the learning phase the test problems comprised equivalent, isomorphic, and novel problems. In order to calibrate the difficulty of the test problems we conducted a baseline study where subjects had to solve the test problems without any instructional information.

Baseline Study

Method

Participants Subjects were 49 third and forth grade pupils of an elementary school in Nikosia, Cyprus.

Materials and procedure Subjects received a booklet and were instructed to solve the 18 aforementioned word problems one after another as well as 20 simple arithmetic calculations that were used to measure basic arithmetic skills. Subjects received no guidance or instructional support.

Dependent measures One error was assigned for each wrong answer in the simple arithmetic calculations as well as in the word problems.

Results and Discussion

The simple arithmetic calculations yielded an average error rate of 30.7%. With regard to the word problems, the average error rate was 48.6%. This baseline performance indicated that the word problems were sufficiently difficult to solve for the children so that an instructional support might influence the results. Therefore, this set of word problems was used in the following experiments without any modifications.

Experiment 1

Method

Participants Subjects were 44 (mainly) third and forth grade pupils of different elementary schools in Nikosia, Cyprus, who participated in the study without payment. Average age was 8.3 years. These subjects had not participated in the baseline study.

Materials and procedure At the beginning of the experiment a pretest was administered to measure basic arithmetic skills. The pretest consisted of 10 simple arithmetic calculations and of 5 simple word problems. After that, a subject started working with BASICOPERATIONS on his or her own using a PC. The experiment ended after the subject had solved the final word problem in the test phase.

Design and dependent measures As a first between-subjects variable the example combinations presented as learning material were manipulated (structure-emphasizing vs. surface-emphasizing). As a second within-subjects variable the transfer distance of the test problems was manipulated by assigning equivalent, isomorphic, and novel problems to subjects. As a first dependent variable we measured the error rates in the pretest. For every wrong answer in the simple arithmetic calculations one error was assigned. With regard to the simple word problems a maximum of two errors for each problem was assigned: one error for applying the wrong basic operation and another error for wrong calculations. Pretest errors were analyzed in order to ensure that the two experimental conditions were comparable with regard to their prior arithmetic skills. Furthermore, time spent on processing example combinations in the learning phase

was registered. In the test phase, error rates in the word problems were obtained by assigning one error for each wrong answer (yielding a maximum of 18 errors).

Results and Discussion

The average time spent with working in BASICOPERATIONS over the experimental groups was about 79 minutes, ranging from 44 to 112 minutes. The children had only little experience with solving word problems as reflected by the average error rates in the pretest for both experimental groups with 31.7% for learners in the structure-emphasizing and 38.5% for learners in the surface-emphasizing condition ($t(39) = -1.15$; $p > .20$; two-tailed).

Unexpectedly, there was no significant difference between learners of the two experimental conditions regarding the time spent on the example combinations with an average of 20.2 minutes for learners with structure-emphasizing and 22.7 minutes for learners with surface-emphasizing example combinations ($t(42) = -.85$; $p > .20$; one-tailed). Similarly, with regard to error rates for the word problems, a MANOVA (*example combinations* x *transfer distance*) revealed no main effects nor an interaction (all $Fs < 1$). A comparison of the average error rates for the word problems between subjects from experiment 1 (50.1%) and subjects from the baseline study (48.6%) showed that the instructional materials in experiment 1 had no effects on performance ($t(91) = .26$; $p > .70$; two-tailed). As an explanation for this unexpected finding one might assume that the children were too young to deal with the instructional materials and that those were not appropriate for learners with very low prior knowledge, i.e., for children just starting to learn how to apply basic operations and how to solve mathematical word problems. To examine this assumption we replicated the experiment with older children possessing a higher level of prior knowledge. For practical reasons, this experiment was conducted in Germany.

Experiment 2

Method

Participants Subjects were 51 third and (mainly) forth grade pupils of different elementary schools in the Saarland, Germany, who participated in the study without payment. Average age was 9.1 years.

Materials and procedure Materials and procedure were the same as in experiment 1. However, the materials were translated from Greek into German.

Design and dependent measures The design was exactly the same as in experiment 1 with example combinations (structure-emphasizing vs. surface-emphasizing) as a between-subjects variable and transfer distance (equivalent vs. isomorphic vs. novel word problems) as a within-subjects variable. As dependent variables pretest errors, learning time, and error rates for the test problems were measured.

Results and Discussion

The subjects in experiment 2 spent an average of about 52 minutes on working in BASICOPERATIONS, ranging from 21 to 81 minutes. With regard to the average error rates in the pretest, subjects in experiment 2 showed a significantly higher prior knowledge (10.1% errors) than those in experiment 1 with 35.0% errors ($t(90) = 8.02$; $p < .001$; two-tailed). For subjects in experiment 2, prior knowledge did not differ between the two groups with 9.6% for learners in the structure-emphasizing and 10.6% for learners in the surface-emphasizing condition ($t(49) = -.35$; $p > .70$; two-tailed).

First, we investigated whether the two experimental conditions of experiment 2 differed with regard to the learning time invested in studying the example combinations. As expected in the *time demands hypothesis*, learners in the structure-emphasizing condition spent significantly more time with the provided example combinations (17.5 minutes) than learners in the surface-emphasizing condition with 12.2 minutes ($t(49) = 1.81$; $p < .05$; one-tailed). This increased time demand can be seen as a result of the more complex processing necessary to learn with structure-emphasizing example combinations compared to less complex processes that are elicited by surface-emphasizing example combinations.

Second, we examined whether there was a differential effectiveness of the two sets of example combinations with regard to equivalent and isomorphic test problems (Figure 1).

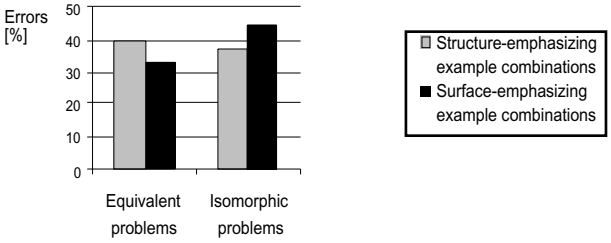


Figure 1: Mean error rates [%] as a function of example combinations and transfer distance

A MANOVA (*example combinations* x *transfer distance)* showed no main effect (all $Fs < 1$), but a significant interaction as expected ($F(1, 49) = 4.11$; $MS_e = 1144.91$; $p < .05$). Subjects who studied structure-emphasizing example combinations performed better on isomorphic test problems than subjects who studied surface-emphasizing example combinations whereas this pattern was reversed for the performance in equivalent problems. In this case the surface-emphasizing condition outperformed the structure-emphasizing condition. This finding provides evidence for the *differential effectiveness hypothesis* which assumes a difference of the two instructional methods with regard to their effectiveness depending on the instructional goal, i.e., transfer distance. Thus, to consider instructional goals may be essential when designing instructional materials.

Third, we tested the assumption that none of the two sets of example combinations supports solving novel test problems. As expected, the error rates for novel test problems did not differ between the two experimental groups with an aver-

age error rate of 36.5% for learners in the structure-emphasizing and 34.0% for learners in the surface-emphasizing condition ($t(49) = .30$; $p > .70$; two-tailed). This result provides evidence for the *far transfer distance hypothesis* which assumes that none of the instructional methods is superior in supporting to solve novel problems.

General Discussion

In this paper issues of instructional design were discussed. It could be shown that different set of instructional examples promote the acquisition of different problem-solving strategies and the accomplishment of different instructional goals.

It was argued that instructional goals must be considered when designing instructional materials. In order to specify these goals cognitive modeling as a method of cognitive task analysis was proposed. In the task analysis that was presented in this paper two specific strategies were examined and evaluated. As a result it was argued that problem-solving strategies differ in their prerequisites as well as in their appropriateness in the context of a particular instructional goal. Accordingly, instructional methods that impart the required prerequisites to apply specific problem-solving strategies also differ in their impact on problem-solving performance with regard to different instructional goals.

Based on these considerations two instructional methods, i.e., providing structure-emphasizing or surface-emphasizing example combinations, were examined in two experimental studies. It could be shown that both instructional methods were not appropriate for young learners with very low levels of domain-specific prior knowledge. But for learners possessing higher levels of prior knowledge, both, structure-emphasizing and surface-emphasizing example combinations proved to be appropriate methods to foster learning and problem-solving depending on the current instructional goal.

Acknowledgements

We thank Simon Albers, the Ministry for Educational, Cultural and Scientific Affairs of the Saarland, the Friedrich-Ebert-Stiftung as well as the Deutsche Forschungsgemeinschaft for their support and three unknown reviewers for their helpful comments.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*, 181-214.

Gerjets, P., Scheiter, K., & Tack, W. H. (2000). Resource-adaptive selection of strategies in learning from worked-out examples. In L. R. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 166-171). Mahwah, NJ: Erlbaum.

Gray, W. D., & Altmann, E. M. (2001). Cognitive modeling and human-computer interaction. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (Vol. 1). New York: Taylor & Francis.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? exhaustive? empirically distinguishable? *Cognition, 65,* 197-230.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Pirolli, P. (1999). Cognitive engineering models and cognitive architectures in human-computer interaction. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of Applied Cognition*. Chicester: Wiley.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88*, 144-161.

Reed, S. K. (1999). *Word problems*. Mahwah, NJ: Erlbaum.

Reusser, K. (1990). From test to situation to equation: Cognitive simulation of understanding and solving mathematical word problems. In H. Mandl, E. De Corte, N. Bennett, & H. F. Friedrich (Eds.), *Learning and instruction in an international context* (Vol. II). New York: Pergamon Press.

Rittle-Johnson, B., & Koedinger, K. R. (2001). Using cognitive models to guide instructional design: The case of fraction division. In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 633-638). Mahwah, NJ: Erlbaum.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Ross, B. H. (1989). Remindings in learning and instruction. In S. Vosnaiadou, & A. Rotony (Eds.), *Similarity and analogical reasoning*. Cambridge, MA: Cambridge University Press.

Sowder, L. (1988). Children's solutions of story problems. *Journal of Mathematical Behavior, 7*, 227-238.

Sweller, J., van Merrienboër, J. J., & Paas, F. W. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press.

Zachary, W. W., Ryder, J. M., & Hicinbothom, J. H. (1998). Cognitive task analysis and modeling of decision making in complex environments. In J. A. Cannon-Bowers, & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training*. Washington, DC: American Psychological Association.

Learning by Solved Example Problems: Instructional Explanations Reduce Self-Explanation Activity

Silke Schworm (schworm@psychologie.uni-freiburg.de)
Department of Psychology, Educational Psychology, Engelbergerstr.41
79085 Freiburg, Germany

Alexander Renkl (renkl@psychologie.uni-freiburg.de)
Department of Psychology, Educational Psychology, Engelbergerstr.41
79085 Freiburg, Germany

Abstract

Learning from worked-out examples is of major importance for initial skill acquisition in well-structured domains. In addition, research has provided knowledge in regards to structuring worked-out examples and how to effectively combine self-explanation activity and instructional explanations. The goal of the present project was to develop a computer-based learning environment in which teachers can learn how to use worked-out examples. Examples of favorably and unfavorably designed worked-out examples were the primary source of information for the teachers. The examples (of worked-out examples) were not in themselves worked-out examples if one views them from a design perspective as the (design) solution steps were not given. We have labeled this type of examples "solved example problems." We investigated to what extent learning from such solved example problems could be fostered by self-explanation prompts and by providing instructional explanations. The results of our 2x2 design (80 student teachers) showed that prompting self-explanations in particular had favorable effects. Hence, self-explanations fostered learning not only from worked-out examples but also from solved example problems. Supplementary instructional explanations only partially enhanced learning and at times they were even detrimental.

Introduction

This study applies the results of cognitive science research (i.e., worked-out example and self-explanation research) to the design of a computer-based learning environment. An empirical study about this learning environment, in turn, contributes to the research on example-based learning and self-explanations.

Learning from worked-out examples is of major importance for the acquisition of cognitive skills in well-structured domains such as mathematics or physics (for an overview see Atkinson, Derry, Renkl, & Wortham, 2000). However, worked-out examples do not guarantee effective learning. One moderating factor is the learner's self-explanation activity. Only when a learner actively self-explains the rationale of the worked-out solutions to her/himself will s/he gain an understanding of the solution procedures. Another factor is the provision of instructional explanations. In the study presented below, teacher students learned in an example-based computer learning environment how to effectively structure and combine worked-out examples. It was intended to foster their learning by the employment of self-explanation prompts and by supplementary instructional explanations.

Learning by Worked-Out Examples

Worked-out examples consist of a problem, solution steps and the complete solution to the problem. Usually they can be found in mathematics and physics schoolbooks. In most cases, a principle or law is introduced in the beginning followed by a worked-out example. The worked-out example shows how the principle can be applied to problem solving. Then, problems to be solved by the students are given.

Learning by worked-out examples is not meant to refer to the short learning phase between the introduction of a principle and problem-solving. It means, instead, that the example phase is prolonged. Several studies have shown that such example-based learning is more effective for skill acquisition than the standard procedure of studying just one example and then solving problems (for an overview see Sweller, van Merriënboer, & Paas, 1998).

Of course, the use of worked-out examples do not guarantee effective learning. Learning outcomes are influenced mainly by (1) the learner's self-explanation activity and the provided instructional explanations and (2) how the learning materials (examples and problems) are structured (cf. Atkinson et al., 2000). These two aspects are discussed in the following sections.

Self-Explanations and Instructional Explanations

The extent to which learners benefit from the study of worked-out examples depends on how well they explain the rationale of the presented solutions to themselves ("self-explanation effect", Chi et al., 1989; Renkl, 1997; Renkl, Stark, Gruber, & Mandl, 1998). It is especially useful to make the "meaning" of specific operations explicit by reifying the relationship between (sub-)

goals and operators or with the principles underlying a specific operation.

Whereas self-explanations are of major importance, research has shown that the effects of instructional explanations are often disappointing (e.g., Brown & Kane, 1988; Chi, 1996). It seems to be more effective to prompt self-explanations than to offer instructional explanations. On the other hand, it has to be taken into account that relying solely on self-explanations also has several disadvantages. For example, at times the learner is not able to self-explain a specific solution step or the given self-explanations are incorrect.

Renkl (in press) developed a set of instructional principles to support the spontaneous self-explanation activity by providing instructional explanations. Two central principles are (1) the priority of self-explanations (instructional explanations should just be used as type of support) and (2) the furnishing of instructional explanations on learner demand. The study of Renkl (in press) showed that such instructional explanations heightened the average learning outcomes.

However, instructional explanations may not only have positive effects. Due to their feedback functions, a specific problem can occur (e.g., Kulhavy, 1977). If feedback containing the correct answer (here: the explanation) is easily available, learners typically reduce their effort in attempting to find the answer on their own. They tend to look up the right answer instead of coming up with the answer themselves – which reduces learning outcomes. Thus, the provision of instructional explanations may reduce self-explanation activities. Aleven and Koedinger (2000) found, in the context of computer-based learning, that in more than 80% of the cases their learners did not use available help which additionally required self-explanation activity. The learners asked directly for the help which contained the final solution.

In summary, self-explanations are of major importance when learning from worked-out examples. Instructional explanations can foster learning but often they do not. What is left open is the question as to what extent the findings on (self-) explanation can be generalized to non-mathematized content areas.

Design of Worked-out Examples

Researchers (e.g., Catrambone, 1996; Mwangi & Sweller, 1998) have suggested that the design of worked-out examples plays a critical role in their effectiveness (intra-example features). For example, featuring sub-goals prominently fosters learning outcomes (Catrambone, 1996). In the present study, we focused on the so-called *integrated format*: Examples are constructed which integrate all sources of information into one source (e.g., diagrams and text). Splitting learners' attention across multiple, non-integrated informational sources causes irrelevant cognitive load and impairs learning (Mwangi & Sweller, 1998; Ward & Sweller 1990).

Beyond the structure of single worked-out examples the combination of multiple examples is of significance for learning outcomes (inter-example features). In general, multiple examples that contain different surface features (e.g. figures, objects) should be used. This aids the learner's ability to recognize the common underlying structure when asked to compare examples.

Often there are two or more different structures to be learned. It is important to emphasize the structural aspects by using very similar surface features for different problem types. Learners frequently do not recognize the difference of the underlying structure because they concentrate on the similarity of the surface features. Therefore, they often choose the wrong solution. Thus, when dealing with different but interrelated problem types, multiple examples should be combined in such a way that the relevant structural features are apparent to the learner (*structure-emphasized example set*). This can be achieved, as stated above, by using different surface features within one problem type and similar surface features between the different problem types (Quilici & Mayer, 1996).

Of course, there are many other example features which influence the effectiveness of learning from worked-out examples, but in the first module of our computer-based learning environment, we implemented only two features: the integrated format and the structure-emphasized example set.

Research questions

Taking into account the current research on learning from worked-out examples, two important questions can be formulated:

(1) How can we teach teachers the knowledge about the effective use of worked-out examples in their classroom? To reach this goal means to bridge research findings into practice and to improve the quality of instruction.

(2) Do self-explanations and additional instructional explanations foster learning when learning from solved example problems, that is, examples that do not provide solution steps but only the problem and the final solution? Such a solved example problem would be, for example, a well-written essay (literature) or the picture of an intriguing mask (arts). So far it remains an open question as to what extent the results of the worked-out example research, which were mainly obtained within the context of learning mathematics and physics, can be transferred to learning by solved example problems in other topic areas.

For addressing the first question we have developed an initial module for a computer-based learning environment. It is the first part of a future web-based learning environment in which teachers can learn how to use worked-out examples in their classrooms. Due to its intended net-based use, not only "objective" learning outcomes were of interest but also the program's acceptance and the perceived learning results. Those

aspects are of major importance when offering a facultative learning opportunity to practitioners. Acceptance is the basis of the usage of the program itself whereas the perceived results are a predictor for its implementation in classrooms.

In this module, future teachers learn about the design of worked-out examples by studying cases of well and poorly designed worked-out examples. It is important to note that the design of worked-out examples is not an algorithmic process with specific solution steps. Therefore, the examples of the program are, as viewed from the teacher's perspective, not worked-out examples (they do not contain any solution steps) but instead solved example problems (there is simply the problem and its solution) (see Fig. 1).

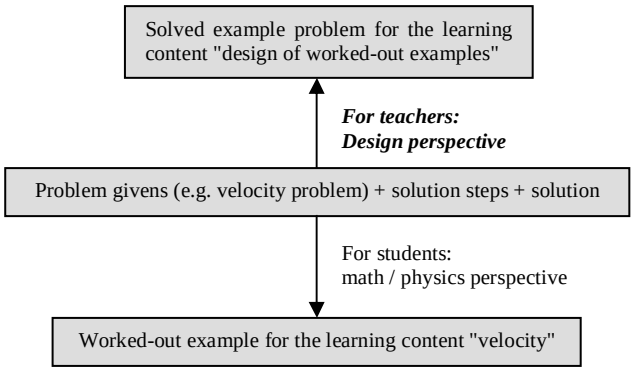


Figure 1: Contents of the learning program from different perspectives

The second research question of this study addresses the effectiveness of self-explanations and instructional explanations when learning by solved example problems: Do self-explanation prompts and additional instructional explanations upon learner demand foster learning?

The following specific research questions were addressed:

1. Is there – as expected – a positive effect of prompting self-explanations on learning outcomes?
2. Is there a positive effect of providing instructional explanations?
3. Do the two instructional means combine additively or non-additively?
4. Do the different instructional treatments influence acceptance and perceived leaning outcomes?

Methods

Sample and Design

80 student teachers from two different colleges volunteered to take part in this study (mean age: 22.3 years; 52 female and 28 male participants). One part of the participants were future teachers in German low-track and medium-track schools (n = 47), the other part were future teachers in high-track schools (n = 33). The participants were randomly assigned to the four experimental conditions of a 2x2-factorial design (n = 20 in each group). In the experiment, the participants learned in a computer-based learning environment how to effectively design worked-out examples by studying solved example problems (factor 1: prompting self-explanations [with and without], factor 2: instructional explanations [with and without]). In the analysis of the written self-explanations (reactions to the prompts), 6 participants are missing due to technical problems (3 per cell). Technical problems also led to 4 missing values with respect to time-on-task which was recorded as a control variable.

Learning Environment

The program contained a short introduction about learning with worked-out examples. Afterwards examples of worked-out examples or sets of examples were displayed. They were taken from the domains of geometry and physics.

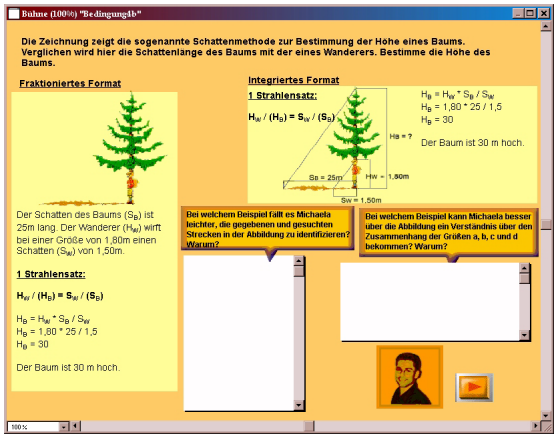


Figure 2: Screenshot of the learning environment

Figure 2 shows a screenshot of the learning environment with a solved example problem with an integrated format. Two worked-out examples were presented to the participants—one required the mapping of different sources of information (graphic, calculation, and text – left side) whereas the other was designed in an integrated format (right side). The self-explanation prompts asked the learners to comment on why one of these two worked-out examples was more favorable. Self-explanations had to be provided to all 13 prompts but their extensiveness (number of elaborations) was self-regulated by the learner. The provided instructional explanations were—in a sense—answers to the self-explanation prompts. They were presented verbally when the learners clicked on a button. The button contained a picture of an expert teacher introduced earlier in the program (see Fig. 2). A "text"-button, which appeared after the acoustic presentation, enabled the learner to view a written explanation. The instructional explanations were

demanded 4.13 times on average (*SD* = 3.31). The frequency of demands did not correlate with learning outcomes. Therefore, this variable was not further considered.

Procedure

The participants worked in individual sessions of approximately 3 hours. In order to provide basic knowledge to allow the participants to be able to understand the solved example problems, an instructional text containing the basic principles of the worked-out example design was given to the participants. Afterwards they studied several solved example problems dealing with an integrated format and structure-emphasized example set in the domain of geometry and physics. The different domains were chosen to foster the transfer of the acquired knowledge. All participants were instructed to think aloud during this period in order to assess (oral) self-explanations (these data have not yet been analyzed). The group with prompted self-explanations had to write down their explanations in note-boxes. During the study of examples, the time spent on different pages of the learning program was registered. After studying the solved example problems, the participants worked on the post-test (learning outcomes). Lastly, the participants filled out a questionnaire regarding the perceived usefulness of the program.

Instruments

Post-test: Assessment of the Learning Outcomes. The first part of our post-test consisted of selection tasks. The participants had to choose one of several given worked-out examples (integrated format) or they had to combine four examples to a structure-emphasized example set. The domains used were geometry, physics (near transfer), and arithmetic (far transfer). For near and far transfer there were three tasks with increasing complexity. Depending on the complexity of the task 1, 4 or 6 points could be achieved (selection part: maximum of 22 points). The second part was a generation task: The participant had to create a structure-emphasized example set in an integrated format. The quality of this task solution was rated by three raters according to specified criteria (e.g., using integrated format in all examples). An entirely correct solution was awarded with 12 points.

Questionnaire. Included in the questionnaire were demographic questions as well as questions concerning the acceptance of the learning environment and the perceived learning results. The items were to be answered on a rating scale from 1 to 6. For the acceptance scale (19 items; e.g., "The content of the program was easy to understand."), we obtained a Cronbach's α of .86. There were four items that assessed the perceived learning results (e.g., "How would you judge your current knowledge about worked-out examples?"; Cronbach's α: .72).

Written Self-Explanations. In the learning program, the learners in the self-explanation groups were prompted 13 times. The written self-explanations were analyzed using a specifically developed coding system. The main categories were as follows:

(1) *Connection between the design principles of worked-out examples and the solved example problem presented in the program* (e.g., "The variables are written next to the lines, therefore there will be less mapping problems.").

(2) *Linkage to the learning-goals* (integrated format or structure-emphasized example set) (e.g., "A different surface does not automatically require a different solution method.").

(3) *Mathematical content of the solved example problems* (e.g., "In both examples I have to determine the speed.").

(4) *"Side-aspects"* These remarks were correct but did not refer to the learning goals of the program (knowledge about integrated format or structure-emphasized example set) (e.g., "… provided that the second [example in integrated format; comment by the authors] could be drawn clearer.").

(5) *Metacognition* (e.g., "I would have solved the problem in the same way.").

The written reactions to the self-explanation prompts were segmented with the coding categories in mind. Often more than just one elaboration (category) was coded in a reaction to a prompt. The coding categories were distinct and there were no inclusions of segments. A few utterances did not fall into any of the categories (e.g., statements about specifics of the learning program), so they were not taken into account.

We aggregated the codings in two respects. First, all categories were summed up together to an overall score of elaboration activity. The single categories occurred relatively infrequently so that corresponding scores would have not been reliable. Second, the elaborations in reaction to the 13 prompts were aggregated.

Results

Pre-Analyses

In the post-test, a maximum of 34 points could be achieved (M = 23.20; SD = 5.67). The two subtests (selection and generation) were summarized, due to their similar result patterns and their positive correlation (r = .31; p < .05).

The post-test correlated significantly with the perceived learning results (r = .41; p < .05). It did not significantly co-vary with the acceptance scale (r = .06; p > .10). There was a significant correlation between the acceptance and perceived learning results (r = .51; p < .05).

The amount of elaborations in the written self-explanations was found to be an important predictor of the learning outcomes (r = .55; p < .05). There was no significant influence of time-on-task on learning outcomes (r = .19; p > .10); even in the single subgroups there were no significant relationships between time-on-task and learning outcomes. This pattern in the results indicates that it was not primarily the quantitative aspect of learning (learning time), but the qualitative aspect (elaborations) which influenced learning outcomes. Table 1 summarizes the descriptive results of time-on-task, written self-explanations, acceptance, perceived usefulness, and learning outcomes (posttest).

Effects of Self-Explanation Prompts and Instructional Explanations

In order to determine the effects of our experimental variations, we performed an ANCOVA controlling for the type of student teachers. This variable significantly influenced learning outcomes (low- and medium-track teachers: M =22.05, SD = 5.76; high-track teacher M = 24.83, SD = 5.18; $t(78)$ = 2.21; p < .05). As the regression slopes of the experimental groups did not differ significantly, the type of student teachers could be used as a covariate.

The analysis of the posttest showed a significant main effect for the prompting self-explanations, that was of medium to high practical significance ($F(1,75)$ = 8.68; p < .05; η^2 =.11). There was no significant main effect for the instructional explanations ($F(1,75)$ = 0.37; p > 0.1). The interaction effect reached the level of significance and was of medium practical significance ($F(1,75)$ = 4.91; p < .05; η^2 = .06; see Fig. 3).

Table 1: Means and standard deviations of the experimental groups.

.	no self-expl. & no instructional expl.	self-expl. & no instructional expl	no self-expl. & instructional expl.	self-expl. & instructional expl
Time-on-task	23.75 (8.82)	46.91 (10.09)	29.64 (9.14)	50.31 (16.76)
Elaborations	--	19.82 (4.51)	--	16.88 (4.51)
Acceptance	4.51 (0.39)	4.39 (0.58)	4.67 (0.65)	4.44 (0.61)
Perceived learning results	3.58 (0.64)	3.98 (0.89)	4.16 (0.77)	3.76 (1.17)
learning outcomes	20.36 (5.46)	25.80 (4.44)	22.75 (5.30)	23.86 (6.29)

As expected, the group without self-explanation prompts and without any instructional explanations performed the worst. Offering instructional explanations fostered learning when self-explanations were not prompted. However, when self-explanations were prompted, supplementary instructional explanations impaired learning. Hence, the most successful group received prompting of self-explanations, but no instructional explanations.

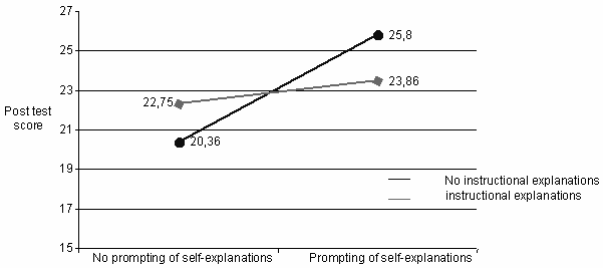


Figure 3: Interaction between prompting self-explanations and instructional explanations with respect to learning outcomes.

As stated in the introduction, the option to request instructional explanations could reduce self-explanation activities. Indeed, the group with self-explanation prompts and additional instructional explanation did elaborate less than the self-explanations only group (about 17 versus 20 elaborations). The difference was significant ($t(32)$ = 1.72; p < .05; one-sided).

Did the groups differ in their time-on-task? Table 1 displays remarkable differences between groups (24 minutes versus 50 minutes). There was a main effect of prompting self-explanations ($F(1,72)$ = 64.91; p < .05), with prompting increasing time-on-task. The main effect of instructional explanations reached only a significance level of 10% ($F(1,72)$ = 2.93; p < .10). There was no interaction effect (F< 1). Even though there were differences in time-on-task between the groups, the experimental effects could not be interpreted as mere time-on-task effects. Firstly, typing the self-explanations per se requires time; secondly, as mentioned above, our analyses showed that the quality and not the quantity of learning activities determined the learning outcomes. However, the results indicated that fostering learning by prompting self-explanations requires additional learning time.

An analysis of the treatment effects on perceived learning results (using the type of student teachers as a covariate; there were no significant differences in the regression slopes of the experimental groups) yielded no main effects in the prompting of self-explanations (F < 1) and instructional explanations ($F(1,75)$ = 1.54; p > .10), but a significant interaction ($F(1,75)$ = 6.36; p < .05). The two groups with self-explanation prompts showed similar perceived learning results of medium size. The group without such prompts and without instructional explanations evaluated their learning results as low. When only instructional explanations were provided, the participants perceived the highest learning results.

The various conditions did not differ in their acceptance by the learners. There is neither a main effect in the prompting of self-explanations (F (1,76) =

1.91; $p > .10$) nor a main effect for instructional explanations ($F < 1$) nor an interaction ($F < 1$).

Discussion

An informal look at the posttest results shows that all student teachers learned substantially about the design of worked-out examples. Hence, we made a significant step towards answering the question as to how to teach teachers knowledge about example-based learning. The extent of learning, however, varied significantly with the experimental conditions. The most effective method was to prompt self-explanations. Instructional explanations were detrimental, at least if they were combined with prompting self-explanations, because they reduced self-explanation activity and thereby the learning outcomes.

Nevertheless, it can be stated that instructional explanations without self-explanation prompts leads to better learning outcomes than leaving the students completely to their own devices. This result is consistent with the findings of Renkl (in press) on the effectiveness of instructional explanations. In Renkl's experiment, no self-explanation prompts were employed.

In contrast to the objective learning outcomes, the highest perceived learning outcomes were found in the instructional explanations-only group. There is an obvious contrast between the real learning outcome and the perceived one. While fostering the learners' own activities objectively leads to the best results, the learners seem to prefer having the explanations presented to them. Apparently, learners do not value instructional measures which require their own activity.

An important consequence of the results presented here is the evident relevance of self-explanation activity for learning outcomes, not only when learning with worked-out examples but also when learning with solved example problems. For this reason the various results concerning the self-explanation-effect are probably transferable to content areas where no worked-out examples can be sensibly constructed. At the same time instructional explanations seem to be less important than self-explanations – equivalent to the results of worked-out example research. However, further research has to be performed using other types of solved example problems.

Looking at the learning processes it is important to note that the mere amount of elaborations substantially predicts learning outcomes. In the near future, the thinking aloud protocols will be analyzed which will give us further insight into the underlying learning processes in the different experimental groups.

Acknowledgments

This study was funded by the German Research Foundation (DFG; RE 1040/5-1).

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 292-303). Berlin: Springer.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181-214.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from examples. *Cognitive Psychology, 20,* 493-523.

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1020-1031.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* S33-S49.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182.

Kulhavy, R.W. (1977). Feedback in written instruction. *Review of Educational Research, 47,* 211-232.

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction, 16,* 173-199.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144-161.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21,* 1-29.

Renkl, A. (in press). Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction.*

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90-108.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251-296.

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80,* 424-436.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7,* 1-39.

The Psychological Implausibility of Naturalized Content

Sam Scott (sscott@ccs.carleton.ca)
Department of Cognitive Science
Carleton University, Ottawa, ON K1S 5B6

Abstract

Conceptual Atomism (CA) is the view that most concepts are represented psychologically as atoms, with no internal structure and CA Atomism on its own is a psychological/semantic theory, but from its inception, it has been mixed up with the separate, meta-semantic project of naturalizing content. I will show that this combined project is forced to end in the self-defeating position of positing non-atomic structures for a large number of concepts. I suggest that a better way out would be to separate the two projects, and allow each to develop on its own.

Introduction

For the last two decades, a number of psychologically minded philosophers have been pursuing a project aimed at naturalizing mental content (Dretske, 1981; 1986; Fodor, 1987; 1990; Millikan, 1984; 1989; 1990). This is a meta-semantic project that seeks an explanation of how meaningful states can arise from non-meaningful ordinary matter. The leading players in this project are also proponents of Conceptual Atomism, the view that concepts are atoms, with no internal structure or necessary relations to other concepts. Conceptual Atomism is a psychological/semantic project, for which the project of naturalizing mental content is supposed to provide a meta-semantics. This combined project – call it Naturalized Conceptual Atomism (NCA) – is still very much a going concern today (Fodor, 1998; Laurence and Margolis, 1999; Margolis 1998; Millikan, 1998; Usher 2001).

The meta-semantic project has a big problem with what I will call 'unacquainted content' (defined below). Proposed solutions to this problem either do not work, or lead to a psychological/semantic position that proponents of NCA have explicitly rejected in the past – namely, that a large number of lexical primitives correspond to complex (non-atomic) concepts. I will look at the three main attempts to naturalize mental content and show how they all either fail or lead to a non-atomic structure for large numbers of concepts. The remedy for this situation, as I see it, is to separate the meta-semantic project from the psychological/semantic project, and let each develop, for the time being, independently of the other.

A Few Definitions

Concept

Following the standard psychological usage, I am using the term 'concept' to mean a sub-propositional mental representation. (This is in contrast to the standard philosophical usage in which a concept is more like an abstract object.) For the present purposes, I will stick to examples of concepts (mental representations) that are about objects or natural kinds.

Unacquainted Content

Unacquainted content is the Achilles heel of NCA. It is the kind of content that a concept has if its bearer has had no direct experience with the represented object or kind. For example, anyone who has experience with dogs (i.e. almost everyone reading this) will have a normal DOG concept. But most North Americans who have heard of, but never directly experienced, wombats have a WOMBAT concept with unacquainted content.[1]

The term 'unacquainted content' also covers many kinds of hypothesized, future or fictional content. For instance the concept, UNICORN, has unacquainted content because the concept bearers could not possibly have directly experienced the nonexistent objects to which it refers.

Nonexistent Object

Nonexistent objects are what concepts with unacquainted content seem to refer to. Maybe nonexistent objects are objects in possible worlds, maybe they have some kind of Meinongian nonexistent being,[2] or maybe they don't exist at all and references to them are vacuous. I don't intend to take a position on this ontological issue, because the main question of the paper is not whether there are unicorns, but whether there are UNICORNs (atomic representations for unacquainted content).[3]

The Problem of Unacquainted Content

The main proponents of NCA are Dretske, Millikan and Fodor. All three are engaged in a philosophical project that seeks (a) a naturalized account of (b) external content, and all three tend to assume that (c) concepts are atoms with no internal structure. Their three different brands of NCA differentiate around (d) the special problems posed by misrepresentation. I will briefly discuss these four points of agreement and then I will discuss the differences between

[1] A word in small caps (e.g. WOMBAT) refers to a concept, while a word in single quotes (e.g. 'wombat') refers to a lexical item.
[2] Alexius Meinong was the German philosopher and psychologist credited with proposing this solution (Meinong, 1904).
[3] My hunch is that everyday common sense is pseudo-Meinongian, and therefore my description of a unicorn as a nonexistent object will be perfectly intelligible to all but the most dogmatic readers.

the three proposals, focusing on the special problem posed by unacquainted content.[4]

(a) A Naturalized Account. To naturalize content would be to find a coherent story to tell about how the intentional nature of concepts arises from the non-intentional nature of ordinary matter. In practice this has typically meant grounding the meaning of a symbol in some kind of causal or information-bearing relationship between the symbol and the object it represents.

(b) External Content. Proponents of NCA follow Putnam (1975) in insisting that there has to be an external or broad component to representational content. Meaning is not (only) in the head.

(c) Conceptual Atomism. Dretske, Millikan, and Fodor all make the assumption that concepts and other meaningful mental states must be both syntactically and semantically atomic. A concept simply refers to an object in the world. Semantically speaking, no part of a concept's meaning derives from any relationship it may have with other concepts. Syntactically speaking, if the concept had an internal structure of some kind, it would raise the question of what the individual parts of the structure refer to, and it's doubtful whether that is even a meaningful question to ask in this context. If, for example, DOG is satisfied by all and only dogs because of a causal relationship between DOGs and dogs, then there is just no internal structure in the equation that needs to be explained.

(d) Misrepresentation. If the meaning of DOG is just dog, and if DOG gets its meaning in virtue being caused by dogs, what do we do with the fact that sometimes DOG tokens might be caused by things other than dogs? For example, a cat on a dark night might cause a DOG token. If so, this seems to imply that DOG means the same as 'dog or cat on a dark night', which is intuitively wrong. In fact, this "disjunction problem" is much bigger than that. Pictures of dogs can also cause DOG tokens. So can the word 'dog', thoughts about pets, and so on. So the meaning of DOG, on this account, would actually be an infinite disjunction including things like dogs, cats on dark nights, 'dog' tokens, PET tokens, LEASH tokens, and so on. It is in attempting to solve this problem that the three accounts proposed by Dretske, Millikan and Fodor diverge.

Dretske on Misrepresentation

Dretske was the first to formulate a version of NCA built on information theory (Dretske, 1981). According to this approach, a concept C represents some X in the world only if C carries information about X. More specifically, if X and only X causes C then C represents X. The formulation is meant to be counterfactual supporting. So if X and only X *would cause* C, then C represents X. Left like this, Dretske's theory suffers from the disjunction problem as badly as any causal theory possibly could – the condition that only X would ever cause C is far too strong to apply to real cognitive agents in noisy environments.

Dretske's proposed solution (Dretske, 1986)[5] begins by making a distinction between simple and complex organisms. Simple organisms have only one route to a representational state. As an example, he points to marine bacteria that contain magnetic sensors called magnetosomes. These sensors detect the surrounding magnetic field and allow the bacterium to align itself with magnetic north. Since in the northern hemisphere, the lines of the magnetic field are inclined downwards, the bacterium can use the signal from its magnetic sensors to swim upwards or downwards in the water. The bacteria die in the oxygen-rich water close to the surface, so bacteria living in the north are naturally selected to use their sensors to swim downwards (towards magnetic north). If they are transplanted to the southern hemisphere where the field lines incline upwards, they will kill themselves by swimming into oxygen-rich water.

Dretske thinks that simple organisms like the magnetosome bacteria cannot accidentally misrepresent, because the information contained in whatever representations they form is ambiguous. In its natural environment, the bacterium's magnetosome representations reliably causally covary with the direction of oxygen-free water. Hence it is tempting to say that when the bacterium is moved to the southern hemisphere, it begins to misrepresent that direction. But on the other hand, the magnetosome representations also reliably causally covary with the direction of magnetic north, and this does not change no matter where on earth the bacterium is moved to. So on this latter view, it is not a case of misrepresentation that causes the northern bacteria to kill themselves when moved to the south. The magnetosome mechanism still reliably indicates magnetic north, but something else is going wrong inside the organism that causes it to swim in that direction and kill itself. Dretske concludes from this that where there is only one causal route to a representation, misrepresentation cannot occur because the informational content of the representation (i.e. what the representation is *supposed* to mean) is indeterminate. Unless it is possible to unambiguously determine a representation's informational content, it is not possible to determine whether it has been tokened in error.

In more complex organisms, there can be more than one route to a representation. For instance, a human being can detect a hamburger by seeing it, smelling it, tasting it, feeling it, and so on. There are multiple sensory routes that end in the same representation, H. If, on the contrary, one could only detect a hamburger by smelling it, H would reliably causally covary with both the hamburger and the

[4] Sometimes the term 'misrepresentation' is used to include representations of nonexistent objects and states of affairs as well as representations tokened in error. But nonexistent objects are a kind of unacquainted content. So for my purposes, a misrepresentation is a representation that was supposed to correctly represent an existing object or state of affairs, but, for some reason, failed to do so.

[5] In later work, Dretske (1988) pursued a different solution that shares more in common with Millikan's approach, discussed below.

odor. So the content of H, on Dretske's story, would be indeterminate. But since there are at least four sensory routes (in a human) to H, the content can be fixed. A token of H caused by seeing a hamburger does not causally covary with the odor of the burger, so the odor can be ruled out as part of H's content. Now we can see how misrepresentation is possible. Any one of the senses can be tricked into causing a token of H when there is no hamburger present, but since the content of H is fixed by the intersection of multiple causal routes, the resulting token H can sensibly be considered to accidentally misrepresent.

Dretske and Unacquainted Content

Information-based NCA of this kind suffers from a big problem with unacquainted content. In Dretske's version, the problem is, in many cases, one of indeterminacy. Take Jay Leno, the host of the tonight show. Like most people with a LENO concept, I have watched him for hours on TV. I know both what he looks like and what he sounds like, so I have two causal routes to my LENO concept. If I ever saw Jay Leno in person, it's reasonable to suppose my LENO concept would be tokened through one or more of these causal routes. So the condition that Leno would cause LENO tokens is satisfied. But the condition that *only* Leno would cause LENO tokens is violated – recordings of Leno also cause LENO tokens. Unfortunately, the multiple causal routes story is no help here because I have only two causal routes to LENO tokens and they would both be engaged whether I saw him live or on TV. It's possible that this problem can be set aside by noting that there is a causal relationship of some sort between the real Leno and the TV Leno, but going down this road will likely produce more problems than it solves. There is a causal relationship between a certain type of bacteria and pimples, but it should not follow, at least in any Conceptual Atomist story, that any part of the content of my PIMPLE concept is a type of bacteria.

The problem gets worse when there is no possibility at all of a direct sensory causal route to a token, as is the case for nonexistent objects like the fictional detective, Sherlock Holmes, or the Second Shooter hypothesized in certain theories about the assassination of John F. Kennedy.[6] I do know a lot of facts about what these two nonexistent objects are, having heard the conspiracy theory about the Kennedy assassination and read the stories about Sherlock Holmes. But it does not follow that either of these individuals (should they turn out to exist after all) would cause appropriate tokenings in me if I ever saw them, because I have no history of a direct sensory link with them, and therefore no tokens with the appropriate informational content.

[6] I have no opinion about these theories. Let's just say for the sake of argument that there was no Second Shooter.

Millikan on Misrepresentation

Millikan approaches the problem of misrepresentation from another direction. One way of looking at misrepresentation is to say that it arises when a given representation fails to perform its proper function. For example, if DOG is tokened in response to a cat, we can intuitively say that the mechanism that outputs DOG tokens has failed to do its job properly. The DOG token is only supposed to represent dogs, but it is being tokened (in this case, accidentally) in response to a cat. So all the approaches to explaining misrepresentation within a theory of NCA have in common that they want to find some naturalistic way to describe the proper function of a given representation. Millikan meets this challenge head on by trying to find a teleological solution rooted in the theory of natural selection (Millikan 1984; 1989; 1990; also see Dretske, 1988).

Consider the human heart. Intuitively, we would like to say that its proper function is to circulate blood, but where do we get the authority to say such a thing? Millikan answers that we can say the heart has the function of circulating blood if we can show that's what hearts were naturally selected for. Applying this idea to mental representations, Millikan suggests that we can say that, for instance, DOG refers to dogs if we can show that's what DOG tokens were naturally selected for.

To determine what a representation was selected for, Millikan urges us to focus on the systems within the organism that make use of the representation (Millikan, 1989). For example, the representations produced by the navigation mechanism within a magnetosome bacterium are consumed by some other part of the organism that uses the information to pick the current swimming direction. If we assume that these various mechanisms were selected for their ability to propel the bacterium away from oxygen-rich water, then the proper function of the magnetosome representations must be to represent the direction of such water. So when we transplant the bacterium, it can truly said to be Accidentally Misrepresenting that direction. Millikan's solution has the advantage of allowing us to say what we intuitively want to say about the bacteria – that in normal conditions they represent, and in abnormal conditions they misrepresent.

A tempting way of looking at this solution is that it is the same as Dretske's information-based solution, but with the causal covariation occurring on an evolutionary time scale rather than over the lifetime of a single organism. In fact, Dretske (1981: 234) does toy with the idea of innate representational content produced in just such a way – representations that are selected for the informational content they carry. But reflection on the case of the magnetosome bacteria shows the real difference in the two theories. Recall that Dretske (1986) was forced to conclude that the content of the magnetosome mechanism's representations were indeterminate – there were just too many things the representations causally covaried with to judge which was the 'proper' informational content. Exactly the same argument would apply on an evolutionary scale.

But by focusing on the naturally selected proper function of the representations, Millikan avoids this indeterminacy.

Millikan and Unacquainted Content

As appealing as Millikan's solution to misrepresentation may seem, it has problems with unacquainted content that are at least as bad as those associated with Dretske's approach. As Dretske himself has pointed out (Dretske, 1986), the theory cannot explain representational content for anything that a species either has not encountered during its evolutionary history, or has encountered but had no need or use for. If no member of the species, or any ancestor species, ever encountered a particular type of object, then no part of the organisms that comprise that species could possibly have been selected for the purpose of representing that content. This denies representational content to almost any representation of a nonexistent object, and many representations of real things such as works of art, new pieces of technology, or anything that is recent enough to have played no role in the evolutionary history of the species. Millikan has a problem with unacquainted content on an evolutionary scale.[7]

Fodor on Misrepresentation

Arguably, the most promising version of NCA comes from Fodor (1987; 1990; 1994; 1998). For the last 15 years or so, he has been pushing a theory of Asymmetric Causal Dependence (ACD) theory to explain how an information-based semantics could deal with, among other things, misrepresentation.[8] In his essay, "A theory of content II", he combines Dretskian informational semantics (a concept C means X if it's a law that X's cause C's) with an a symmetric dependence condition (Y's that cause C's only do so because X's cause C's and not vice versa). This takes care of misrepresentations such as cats on dark nights causing DOG tokens (this state of affairs is dependent on dogs causing DOG tokens but not the other way around), and it is also extendible to explain various kinds of "robust" tokenings (non-X-caused C tokenings that are nevertheless not error cases – for instance, DOG tokens that are caused by pictures of dogs or thoughts about leashes.)[9]

[7] It could be objected that the representation of hypothetical or nonexistent things *in general* is very useful, and thus could have been selected for. But Millikan's theory is supposed to provide an explanation for the specific content of specific representations, and this is what it fails to do for unacquainted content.

[8] Lately, Fodor (1998) prefers to talk about concept acquisition as a process of "locking on" to a relevant property. The new formulation addresses some concerns about nativism and ontology, but Fodor is clear that however locking on works, the meaning of the resulting concept is still grounded in an informational relationship, and ACD remains his most mature attempt to characterize that relationship.

[9] Note that I am actually describing what Fodor (1990) called the "pure" version of ACD. He also suggests the possibility of a "mixed" version in which he adds the condition that C must have actually been caused by X at least once. This mixed version will obviously fail for unacquainted content, so I will only deal with pure ACD here. And to be fair, Fodor (1994: Appendix B) has made it pretty clear that he doesn't think much of the mixed theory anyway.

Fodor and Unacquainted Content

The problem of unacquainted content for pure ACD is immediately apparent, particularly for nonexistent objects. For example, how can non-unicorn-caused tokenings of UNICORN be asymmetrically dependent on unicorn-caused tokenings when there are no existing unicorns? Fodor thinks that this objection can be answered, by reminding us that, like Dretske, he is telling a nomic story:

> It can be true that the property of being a unicorn is nomologically linked with the property of being a cause of UNICORNs even if there aren' t any unicorns… There wouldn't be non-unicorn-caused UNICORN tokens but that unicorns would cause UNICORN tokens if there were any unicorns. (Fodor, 1990, p101, italics removed and single quotes changed to small caps for consistency).

Fodor has been attacked on the unicorn front before. For instance Baker (1991) constructed a detailed argument based on unicorns and "shunicorns" (a creature of her own design) that requires us to speculate about which of various possible worlds containing unicorns and/or shunicorns is "closer" to our own. If your mind boggles at this kind of talk, I will now offer what I hope is a slightly simpler explanation below for why unicorns are a big thorn in the side of the pure version of ACD.

In this unicorn-free world, all valid UNICORN tokenings must be robust tokenings – they are caused by things other than unicorns. The acquisition of the concept UNICORN in the absence of unicorns comes from exposure to representations (visual or verbal) of unicorns. Having learned about unicorns from books and stories, if a unicorn suddenly popped into existence in front of you, it would likely cause a UNICORN token. So we have two valid causal routes to UNICORN tokens: one from representations of unicorns, and one from possible real unicorns that you might encounter in the future (if unicorns began to exist). To apply ACD, we have to know what would happen if we broke either of these two causal links. Would breaking the causal link between future unicorns and UNICORN tokens break the link between representations of unicorns and UNICORN tokens? My intuition is that this scenario doesn't even make sense, but suppose for the sake of argument that breaking the unicorn/UNICORN link would break the representation/UNICORN link. Then UNICORN tokens are causally dependent on (future) unicorns.

But what would happen if we broke the causal link between representations of unicorns and UNICORN tokens? According to ACD, if UNICORN is to mean unicorn, then this should not affect the causal link between future unicorns and UNICORN tokens. But it obviously does. In a world without unicorns, if you don't learn about them from representations of them then you don't learn about them at all. This means that if a unicorn suddenly popped into existence in front of you, you wouldn't know what it was. Maybe it would cause tokens of HORSE, HORN or whatever,

but it wouldn't cause a UNICORN token because you wouldn't have one for it to cause. So in the best case, causal dependence runs both ways and ACD doesn't apply. In the worst case (where you don't buy the story about breaking the link between future unicorns and UNICORN tokens) ACD runs in the wrong direction and UNICORN ends up having representations of unicorns as its content. But this must be false – UNICORN has unicorns as it's content.[10] Notice that you can run exactly the same argument for any type of unacquainted content, such as my LENO concept. Tokenings of LENO in the presence of Leno are causally dependent on tokenings of LENO in response to representations of Leno.

There is a way out of this trap for an extreme radical nativist. Fodor (e.g. 1998) entertains, though he does not endorse, the possibility that we are born with a stock of atomic concepts waiting to be triggered by the right sort of content-fixing experiences. Applying this idea to unacquainted content, if we all have built-in UNICORN token types that just need to be "triggered" somehow, then maybe our first encounter with a unicorn would cause a UNICORN token after all. Of course we wouldn't have a word for this token, but that is irrelevant. So ACD would be satisfied by assuming that we are born with a lifetime supply of tokens that already have their nomic triggering conditions fixed.

But radical nativism is not a popular option in cognitive science. Though Fodor correctly points out that whether (or to what extent) nativism is true is an empirical question, it seems very unlikely to most researchers that the empirical facts will bear the theory out. Furthermore, if the project is to naturalize content, then all radical nativism does is open up new questions. We are now owed a naturalistic account of how it can be the case that an individual is born with a large stock of mental states that already have the appropriate nomic connections. Given the problems with both Dretske and Millikan's evolutionary accounts, it seems unlikely that such a story is forthcoming. Without the story, all we have reduces to the statement that UNICORN means unicorn because it has a set of properties that causes it to mean unicorn.

The Non-atomic Way Out

All three attempts to construct a theory of NCA seem to fail for unacquainted content. However there is still a way out that is consistent with a slightly weakened version of Conceptual Atomism. This solution, proposed by Fodor (1990: 124) and Dretske (1981: 222, 230) is to allow some concepts to be non-atomic, structured entities built out of atomic components.[11] So UNICORN, LENO, and so on actually unpack into phrasal entities in the language of thought, assembled out of primitive atoms. That is, they are *definitions*.[12] Fodor fails to provide any serious defense of the position, except to state that he thinks the situation in which a complex concept would be required is "*very, very rare*" (1990:124, his italics). Dretske proposes the same solution, but like Fodor, balks at defending it: "I hope [the compositional solution] is sufficiently plausible not to *need* argument" (1981:222, also his italics).

But contrary to Fodor, concepts with unacquainted content don't seem to be particularly rare at all. And contrary to Dretske, the definitional solution does need an argument, having been judged implausible, at least as a general account of conceptual structure, by a wide consensus of Cognitive Scientists.[13] Almost any standard account of the recent history of empirical research into conceptual structure begins with a recounting of the demise of so-called definitional theories (e.g. Komatsu, 1992; Laurence and Margolis, 1999; Smith and Medin, 1981). The most commonly cited reasons for abandoning of a definitional account of conceptual structure are that: a) there is a widespread consensus that most concept words of any interest are not rigorously definable (see Laurence and Margolis, 1999); b) no attempt to find psychological data that might reveal a definitional structure for simple lexical items has succeeded (e.g. Kintsch, 1974); and c) the well-established psychological phenomenon of typicality ratings, or "goodness of example" effects (e.g. Rosch, 1973) is extremely difficult to account for within a definitional theory (see Smith and Medin, 1981).

Conclusion: A Better Way Out?

Dretske, Millikan, and Fodor have no solution to the problem of unacquainted content, unless we take one of two rather unpalatable options: a) accept a radical concept nativism in which tokens like UNICORN are an innate part of our psychological make-up; or b) accept that many concepts, including UNICORN, WOMBAT, LENO, and so on must have a definitional structure. Nobody seems wants to take option (a) seriously, and it begs the question anyhow, so we're left with option (b), which not only has no empirical support, but also contradicts the whole spirit of the Conceptual Atomist enterprise. What do we do now?

Recall that there are at least two projects here: the meta-semantic project of naturalizing content, and the psychological/semantic project of Conceptual Atomism. The first project is stalled by the problem of unacquainted content, and in attempting to save itself, has wreaked havoc on the second project. My suggestion is that we do not accept this conclusion, and that we separate the projects from now on. Let those interested in the meta-semantic

[10] There is a persistent notion that UNICORN must refer to an idea or to a representation. But a unicorn is not an idea or a representation; it's an animal that looks like a horse with a horn on its head. Ideas and representations are not animals and they have neither heads nor horns. So ideas and representations are the wrong sorts of things to serve as the content for UNICORN.

[11] Fodor proposes this (somewhat apologetically) only for cases of nonexistent objects, but it is easily extendible to any unacquainted content.

[12] "… the idea that many terms express concepts that have internal structure is tantamount to the idea that many terms have definitions." (Fodor, 1981: 289)

[13] Ironically, this consensus includes Fodor himself (e.g. Fodor, 1998; Fodor, Fodor and Garrett, 1975; Fodor, Garrett and Walker, 1980).

problem try to solve it on its own terms, and leave Conceptual Atomism to develop on its own. That way Conceptual Atomism can be consistent with itself in claiming that UNICORN and WOMBAT are atomic, just like DOG and COW. This is essentially the Language of Thought hypothesis (Fodor, 1975) with a referential semantics, but without the causal-historical meta-semantics. UNICORN refers to unicorns, but how, exactly, it comes to do that is an issue to be resolved (or not) by the separate project of meta-semantics.

I suspect that there will be some skepticism as to whether Conceptual Atomism can survive without its accompanying meta-semantic theory. Therefore, I will end with two reasons why I think that it can.

1. No competing theory is tied to a similar meta-semantic project. For example, neither the prototype theory nor the theory-theory of concepts attempts to say anything about how meaning arises from non-meaningful stuff. Neither do most modern versions of the definitional theory. And, after all, why should they? At this early stage, a psychological/semantic theory should be judged on its own merits, not by standards set at some other level of analysis.

2. Conceptual Atomism is still a decent theory even without the meta-semantic project. There is no psychological evidence for definitional structure, and the evidence that drives the prototype and theory theories can be accounted for within Conceptual Atomism – the former by supposing that typicality effects arise from a separate categorization mechanism, and the latter by supposing that people do have theories that guide their behavior, but that these theories are *about* the concepts they involve, rather than being *constitutive* of them. And above all, Conceptual Atomism is arguably one of the most natural fits to the computational theories of mind that are still so popular.

Acknowledgements

Many thanks to Rob Stainton for repeated discussion of this material. Thanks also to Andy Brook, Craig Leth-Steensen, and two anonymous reviewers for very helpful comments.

References

Baker, Lynne Rudder. (1991). Has content been naturalized? In Barry Loewer and Georges Rey (Ed.) *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Dretske, Fred. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, Fred. (1986). Misrepresentation. In R. Bogden (ed.) *Belief, Form, Content and Function*. Oxford: Oxford University Press.

Dretske, Fred. (1988). *Explaining Behavior*. Cambridge, MA: MIT Press.

Fodor, Janet D., Jerry A. Fodor, and Merrill F. Garrett. 1975. The psychological unreality of semantic representations. *Linguistic Inquiry, 4*, 515-531.

Fodor, Jerry A. (1975) *The Language of Thought*. New York: Crowell.

Fodor, Jerry A. (1981). The present status of the innateness controversy. In *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Fodor, Jerry A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, Jerry A. (1990). A theory of content II. In *A Theory of Content and Other Essays*. Cambridge, MA: The MIT Press.

Fodor, Jerry A. (1994). *The Elm and the Expert*. Cambridge, MA: MIT Press.

Fodor, Jerry A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, Jerry A., Merrill F. Garrett, E. Walker and C. Parkes. 1980. Against definitions. *Cognition, 8*, 263-367

Kintsch, Walter. (1974). Lexical decomposition: Compression and memory. In *The Representation of Meaning in Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Komatsu, Lloyd K. (1992). Recent views of conceptual structure. *Psychological Bulletin*. 112(3): 500-526.

Laurence, Stephen and Eric Margolis. (1999). Concepts and cognitive science. In Eric Margolis and Stephen Laurence (Ed.) *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Margolis, Eric. (1998). How to acquire a concept. *Mind and Language, 13*, 347-369.

Meinong, Alexius. 1904. The theory of objects. In Alexius Meinong (ed.) *Untersuchungen zur Gegenstandstheorie und Psychologie*. Reprinted in Roderick M. Chisholm (ed.) 1960. *Realism and the Background of Phenomenology*. New York: The Free Press. 76-117.

Millikan, Ruth Garrett. (1984). *Language, Thought and other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, Ruth Garrett. (1989). Biosemantics. *Journal of Philosophy, 86(6)*, 281-297.

Millikan, Ruth Garrett. (1990). Compare and contrast Dretske, Fodor, and Millikan on teleosemantics. *Philosophical Topics, 18(2)*, 151-161.

Millikan, Ruth Garrett. (1998). A common structure for concepts of individuals, stuffs, and real kinds; more mamma, more milk and more mouse. *Behavioral and Brain Sciences, 9*, 55-100.

Putnam, Hilary. (1975). The meaning of meaning. In *Mind, Language, and Reality*. Cambridge, UK: Cambridge University Press.

Rosch, Eleanor H. (1973). On the internal structure of perceptual and semantic categories. In Timothy E. Moore (Ed.) *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Smith, Edward E. and Douglas L. Medin. (1981). *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Usher, Marius. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind and Language, 16(3)*, 311-334.

Counterfactual Undoing in Deterministic Causal Reasoning

Steven A. Sloman (Steven_Sloman@brown.edu)
Department of Cognitive & Linguistic Sciences, Box 1978
Brown University, Providence, RI 02912 USA

David A. Lagnado (David_Lagnado@Brown.Edu)
Department of Cognitive and Linguistic Sciences, Box 1978
Brown University, Providence, RI 02912 USA

Abstract

Pearl (2000) offers a formal framework for modeling causal and counterfactual reasoning. By virtue of the way it represents intervention on a causal system, the framework makes predictions about how people reason when asked counterfactual questions about causal relations. Four studies are reported that test the application of the framework to deterministic causal and conditional arguments. The results support the proposed representation of causal arguments, especially when the nature of the counterfactual intervention is made explicit. The results also show that conditional relations are construed in different ways.

Introduction

Many questions are decided by causal analysis. In the law, issues of negligence concern who caused an outcome and, at least under common law, the determination of guilt requires evidence of a causal chain leading to a crime. Evidence that might increase the probability of guilt (e.g., an accused's race) is impermissible if it doesn't support a causal analysis of the crime. Some legal scholars (Lipton, 1992) claim that legal analyses of causality are in no sense special, that causation in the law derives from everyday thinking about causality. Causal analysis is just as prevalent in science, engineering, politics, indeed in every domain that involves human prediction and control.

Causal analysis is often difficult because it depends not only on what happened, but also on what *might* have happened (Mackie, 1974). Thus the claim that A caused B will often imply that if A had not occurred, then B would not have occurred. Likewise, the fact that B would not have occurred if A had not often suggests that A caused B.

This explains a fundamental law of experimental science: Mere observation can only reveal a correlation, not a causal relation. That's why causal induction requires manipulation, control over an independent variable such that changes in its value will determine the value of the dependent variable whilst holding other relevant conditions constant. Everyday causal induction has these same requirements. Causal inductions in everyday contexts are aided by manipulation of potential causes, by people *intervening* on the world rather than just observing it (the conditions favoring intervention are spelled out in Pearl, 2000; Spirtes, Glymour, & Scheines, 1993).

If we already have some causal knowledge, then certain causal questions can be answered without actual intervention. Some of those questions can be answered through mental intervention, by imagining a counterfactual situation in which a variable is manipulated and determining the effects of change. People attempt this, for example, whenever they wonder "if only..." (if only I hadn't made that stupid comment... If only my data were different...).

Pearl (2000) offers a causal modeling framework that covers such counterfactual reasoning. The framework makes predictions about how people reason when asked counterfactual questions about causal relations. Pearl's analysis extends to relations of probabilistic causality but this paper is limited to studies of deterministic arguments. Before describing those studies, we briefly review the relevant aspects of Pearl's analysis.

Observation vs. Causation (Seeing vs. Doing)

Seeing

In general, observation can be represented using the tools of conventional probability. The probability of observing an event (say, that a logic gate is working properly) under some circumstance (e.g., the temperature is low) can be represented as the conditional probability that a random variable G, representing the logic gate, is at some level of operation g when temperature T is observed to take some value t:

$$\Pr\{G = g|T = t\} \text{ defined as } \frac{\Pr\{G = g \,\&\, T = t\}}{\Pr\{T = t\}}.$$

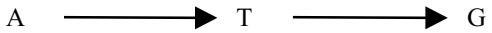
Conditional probabilities are symmetric in the sense that, if well-defined, their converses are well-defined too. In fact, given the marginal probabilities of the relevant variables, Bayes' rule tells us how to evaluate the converse:

$$\Pr\{T = t|G = g\} = \Pr\{G = g \mid T = t\} \frac{\Pr\{T = t\}}{\Pr\{G = g\}}. \qquad (1)$$

Doing

To represent action, Pearl proposes an operator *do(•)* that controls both the value of a variable that is manipulated as well as a graph that represents causal dependencies.

do(X=x) has the effect of setting the variable X to the value x and also changes the graph representing causal relations by removing any directed links from other variables to X (i.e., by cutting X off from the variables that normally cause it). For example, imagine that you believe that temperature T causally influences the operation of logic gate G, and that altitude A causally influences T. This could be represented in the following causal diagram:



Presumably, changing the operation of the logic gate would not affect temperature (i.e., there's no causal link from G to T). We can decide if this is true by acting on the logic gate to change it to some operational state g and then measure the temperature; i.e., by running an experiment in which the operation of the logic gate is manipulated. We could not in general determine a causal relation by just observing temperatures under different logic gate conditions, because observation provides merely correlational information. Measurements taken in the context of action, as opposed to observation, would reflect the probability that T=t under the condition that *do*(G=g):

$$\Pr\{T = t | do(G = g)\}$$

Obtained by, first, constructing a new causal model by removing any causal links to G:

The rationale for this is that if I have set G=g, then my intervention renders other potential causes of g irrelevant. I am overriding their effects, so I should not make any inferences about them. Now I can examine the probability distribution of T in the causal graph. But in doing so, I should not take into account the prior probability of g, because I have set its value, making its value certain by virtue of my action. Because the *do* operation renders T and G probabilistically independent, the result is that:

$$\Pr\{T = t | do(G = g)\} = \Pr\{T = t\}.$$

The *do* operator is used to represent experimental manipulations. It provides a means to talk about causal inference through action. It can also be used to represent *mental* manipulations. It provides a means to make counterfactual inferences by determining the representation of the causal relations relevant to inference if a variable had been set to some counterfactual value.

Do we "do"?

Consider the following Causal Argument (1) in which A, B, C, and D are the only relevant events:

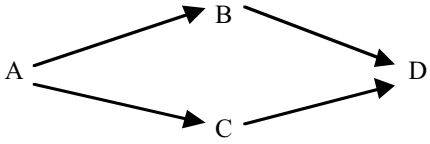
A causes B.
A causes C.
B causes D.
C causes D.
D definitely occurred.

On the basis of these facts, please answer the following 2 questions:

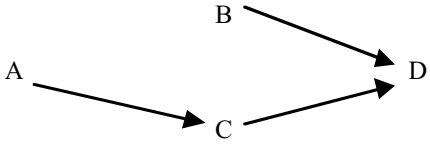
i. If B had not occurred, would D still have occurred?___ (yes or no)

ii. If B had not occurred, would A have occurred?___ (yes or no)

Pearl (2000) gives the following analysis of such a system. First, we can graph the causal relations amongst the variables as follows:



You are told that D has occurred. This implies that B or C or both occurred, which in turn implies that A must have occurred. A is the only available explanation for D. Thus, all 4 events have occurred. When asked what would have happened if B had not occurred, we should apply the *do* operator, *do*(B = did not occur) with the effect of severing the links to B from its causes:



Therefore, we should not draw any inferences about A from the absence of B. So the answer to the counterfactual question ii. above is "yes" because we already decided that A occurred, and we have no reason to change our minds. The answer to counterfactual question i. is also "yes" because A occurred and we know A causes C which is sufficient for D.

Other theories of propositional reasoning, mental models theory (Johnson-Laird & Byrne, 1991) and any theory based on logic (e.g., Rips, 1994), don't really make predictions in this context because the argument uses causal relations and therefore lies outside the propositional domain. The closest they can come is to posit that causal relations are interpreted as material conditionals (an assumption made by Goldvarg & Johnson-Laird, 2001). To see if such an interpretation of Causal Argument (1) is valid, we can consider Abstract Conditional Argument (1):

If A then B.
If A then C.

If B then D.
If C then D.
D is true.

The corresponding questions were:
i. If B were false, would D still be true? ___ (yes or no)
ii. If B were false, would A be true? ___ (yes or no)

The causal modeling framework makes no particular prediction about such an argument except to say that, because it does not necessarily concern causal relations, responses could well be different from those for the causal argument. The predictions made by a "material conditional" account will depend on assumptions about how people interpret the questions; i.e., how they modify the original set of premises. To answer question i. people may suppress the statement that D is true, whilst adding the statement that B is false. If they do, the truth of D is indeterminate, because it is not entailed by the falsity of B. Alternatively, people might not suppress D. The answer would then be "yes" because the original premises state that D is true. Such an account yields a less ambiguous answer to question ii. Once people suppose that B is false, they are licensed to infer, by modus tollens, that A is false. If these "material conditional" theories make any prediction for the causal arguments, these should correspond to their prediction for comparable conditional arguments.

Experiment 1

Method. 238 University of Texas at Austin undergraduates were given one of the two arguments shown and asked the listed questions.

Results. Responses are shown in Table 1. The predictions of the causal modeling framework were supported for the causal arguments but not for the conditional arguments. The predominance of "yes" responses in the causal condition implies that for the majority of participants the supposition that B didn't occur did not influence their beliefs about whether A or D occurred. This is consistent with the idea that these participants mentally severed (undid) the causal link between A and B and thus did not draw new conclusions about A or about the effects of A from a counterfactual assumption about B. Responses to the conditional argument were more variable: no one strategy for interpreting and reasoning with conditional statements dominated.

Table 1: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (1).

Question	Causal	Conditional
i. D holds	80%	57%
ii. A holds	79%	36%

These results were replicated with two additional arguments that used an identical causal or logical structure but added semantic content to the problems. For example, one pair of arguments concerned a robot. Here is the causal version of that problem (Robot Causal Argument 1):

> A certain robot is activated by 100 (or more) units of light energy. A 500 unit beam of light is shone through a prism which splits the beam into two parts of equal energy, Beam A and Beam B, each now travelling in a new direction. Beam A strikes a solar panel connected to the robot with some 250 units of energy, causing the robot's activation. Beam B simultaneously strikes another solar panel also connected to the robot. Beam B also contains around 250 units of light energy, enough to cause activation. Not surprisingly, the robot has been activated.

i. If Beam B had not struck the solar panel, would the robot have been activated?
ii. If Beam B had not struck the solar panel, would the original (500 unit) beam have been shone through the prism?

The same 238 undergraduates were given either the causal or conditional version of this problem. Their responses are shown in Table 2.

Table 2: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (1).

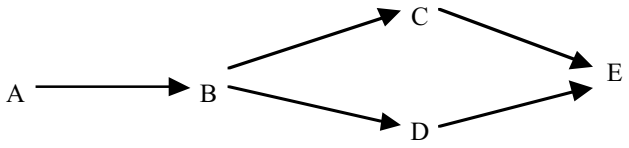
Question	Causal	Conditional
i. robot activated	80%	63%
ii. beam shone	71%	55%

The results are very close to those of the abstract problem except that a higher percentage of participants said "yes" in the conditional version of this problem, $z = 2.83$; $p < .01$. This may have occurred because a larger proportion interpreted the "if-then" connectives of the conditional version as causal relations. The clear physical causality of the robot problem lends itself to causal interpretation.

Experiment 2

One might argue that the difference between the causal and conditional arguments in the previous examples is not due to a greater tendency to counterfactually decouple variables from their causes in the causal over the conditional context, but instead to different pragmatic implicatures of the two contexts. In particular, perhaps the causal context presupposes the occurrence of A more than the conditional context presupposes the truth of A. It's more plausible that D would be true in the conditional arguments even if A were false than that D would have occurred in the causal arguments even if A had not. If so, then the greater likelihood of saying "yes" in the causal scenarios could be due to these different presuppositions rather than different likelihoods of undoing.

To control for this possibility as well as to replicate the effect, we examined causal and conditional versions of arguments with the following structure:



Participants were told not only that the final effect, E, had occurred, but also that the initial cause, A, had too. This should eliminate any difference in presupposition of the initial variable because its value is made explicit. To illustrate with one of the problems shown, here is the causal version of the abstract problem (Causal Argument 2):

A causes B.
B causes C.
B causes D.
C causes E.
D causes E.
A definitely occurred.
E definitely occurred.

i. If D did not occur, would E still have occurred?
ii. If D did not occur, would B still have occurred?

The causal modeling framework predicts that a counterfactual assumption about D should disconnect it from B in the causal context so that participants should answer "yes" to both questions. Participants should only answer "yes" in the conditional context if they interpret the problem causally. Once again the predictions of a material conditional account depend on assumptions about how the questions modify the premises. A plausible assumption is that only statements mentioned in the question are suppressed. Thus in answering question ii., belief about the truth of D and B might be suspended and not-D supposed. However, this leads to a conflict because not-D implies not-B (via modus tollens) but the premises state A and thus imply B (via modus ponens). It is thus unclear whether or not they should infer B. In any case, a material conditional account must predict no difference between the causal and conditional contexts.

Method. Twenty Brown University undergraduates received either the causal or conditional versions of the abstract and robot problems described above.

Results. The results, shown in Tables 3 and 4, are comparable to those from the earlier problems, although the proportion of "yes" responses tended to be lower in the causal condition, especially for the likelihood of the beam shining if the solar panel had not been struck (only 55% in Table 4).

Table 3: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (2).

Question	Causal	Conditional
i. E holds	70%	45%
ii. B holds	74%	50%

Table 4: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (2).

Question	Causal	Conditional
i. robot activated	90%	75%
ii. beam shone	55%	45%

A difference between causal and conditional arguments again obtained for Abstract arguments, $z = 2.20$; $p = .01$, but not for Robot ones, $z = 1.18$; n.s. The difference for Abstract arguments suggests that the earlier results cannot be attributed entirely to different pragmatic implicatures from causal and conditional contexts. The overall reduction in "yes" responses could be due to either a different participant population, some proportion of participants failing to establish an accurate causal model with these more complicated scenarios, or participants not implementing the undoing operation in the expected way (i.e., not mentally disconnecting B from D).

Failure to undo is not entirely unreasonable for these problems because D's nonoccurrence is not definitively counterfactual. The question said "If D did not occur" which does not state why D did not occur; the reason is left ambiguous. One possibility is that D did not occur because B didn't. Nothing in the problem explicitly states that the nonoccurrence of D should not be treated as diagnostic of the nonoccurrence of B.

Experiment 3

The causal modeling framework predicts that the connection between B and D should be mentally undone whenever D is explicitly prevented; when an intervention (mental or physical) outside the model determines the value of D. To simulate such a situation, we repeated Experiment 2, but made the interventional prevention of D explicit.

Method. Participants saw exactly the same sets of premises in both causal and conditional contexts, but were asked different questions, questions that made the external prevention of D explicit (Causal and Conditional Arguments 2EP). For the abstract causal context, the questions were:

i. If somebody stepped in to prevent D from occurring, would E still have occurred?
ii. If somebody stepped in to prevent D from occurring, would B still have occurred?

For the abstract conditional context, the questions were:

i. If somebody stepped in and changed the value of D to false, would E still be true?
ii. If somebody stepped in and changed the value of D to false, would B still be true?

For the robot context, the questions in the causal and conditional versions were identical (only the paragraphs describing the situation differed):

i. If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel, would the robot have been activated?
ii. If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel, would the original (500 unit) beam have been shone through the prism?

Responses were obtained from either 18 or 20 Brown undergraduates.

Results. Results are shown in Tables 5 and 6. The probability of saying "yes" was higher in the explicit prevention context than in its absence, but not significantly so, z = 1.16 and 1.39 for Abstract and Robot arguments, respectively. The two may not differ statistically because the probability of saying "yes" was already so high in the causal condition of Experiment 2. In any case, the great majority of participants acted as if explicitly preventing D caused it to have no diagnostic value for its cause (B), and that therefore other effects of the cause (E) still held. In other words, the effect of explicitly preventing D is well captured by the *do* operator.

Table 5: Percentages of participants responding "yes" to Abstract Causal and Conditional Arguments (2EP), prevention of the antecedent explicit.

Question	Causal	Conditional
i. E holds	75%	50%
ii. B holds	80%	67%

Table 6: Percentages of participants responding "yes" to Robot Causal and Conditional Arguments (2EP).

Question	Causal	Conditional
i. robot activated	75%	83%
ii. beam shone	75%	67%

An unexpected byproduct of explicit prevention was to increase the proportions of "yes" responses in even the conditional context, z = 1.80; p < .05. This probably occurred because the explicit prevention context made it more likely that the arguments would be construed causally. For example, a question beginning "If a lead barrier were placed in the path of Beam B to prevent it from striking the solar panel," may well have suggested to participants that they should construe the situation in terms of physical causation and reason about the situation using causal logic.

One implication of this observation is that the interpretation of conditionals varies with the theme of the text that the statements are embedded in. Conditionals embedded in deontic contexts are well known to be interpreted deontically (Manktelow & Over, 1990). The Abstract Conditional Arguments (1) and (2) above show that when the theme is ambiguous, the interpretation will be highly variable. Robot Conditional Argument (2EP) shows that when the theme is causal, conditionals will be interpreted causally.

Experiment 4

The final experiment attempts to replicate the observations made thus far by showing the undoing effect as well as the enhancement of the effect in an explicit prevention context. Moreover, it does so using an if-then statement in order to show that a conditional statement can be treated as causal in an appropriate context.

Method. The following scenario was described to 78 Brown undergraduates:

> All rocketships have two components, A and B. Component A causes component B to operate. In other words, if A, then B.

The scenario assumes the simplest possible causal graph:

Notice that the relation between A and B is stated using an if-then construction. Approximately half the participants, in the non-explicit prevention condition, were then asked:

i. Suppose component B were not operating, would component A still operate?
ii. Suppose component A were not operating, would component B still operate?

The remaining half, in the explicit prevention condition, were asked:

i. Suppose component B were prevented from operating, would component A still operate?
ii. Suppose component A were prevented from operating, would component B still operate?

The causal modeling framework predicts the undoing effect, that participants will say "yes" to question i., Component A will continue to operate if B isn't because A should be disconnected from B by virtue of the counterfactual supposition about B. It also predicts the proportion will be higher in the explicit than non-explicit prevention conditions because the nature of the intervention causing B to be nonoperative is less ambiguous. No other framework, logical or otherwise, makes either of these predictions. Finally, the causal modeling framework predicts that people should respond "no" to the second question regardless of condition. If A is the cause of B, then B should not operate if A does not.

Results. The results are shown in Table 7. The 68% giving an affirmative answer to the first question in the Non-explicit Prevention condition replicates the undoing effect seen in the previous studies. The even greater percentage (89%, z = 2.35; p < .01) in the Explicit condition replicates the finding that the undoing effect is greater when the reason that a variable has the specified value is made explicit. Responses to the second question were almost all negative, demonstrating that people are clearly

understanding that the relevant relation is causal. This rules out an alternative explanation for the earlier studies, that people were treating causes and effects as disconnected because they didn't interpret the relations as causal but merely as correlational.

Table 7: Percentages of participants responding "yes" to questions in the Rocketship scenario given questions with antecedents non-explicitly or explicitly prevented.

Question	Non-explicit Prevention	Explicit Prevention
i. if not B, then A?	68%	89%
ii. if not A, then B?	2.6%	5.3%

Discussion

These data show that most people obey a rational rule of counterfactual inference, the undoing principle. When reasoning about the consequences of a counterfactual supposition of an event, most people do not change their beliefs about the state of the normal causes of the event. They reason as if the mentally changed event is disconnected and therefore not diagnostic of its causes. This is a rational principle of inference because an effect is indeed not diagnostic of its causes whenever the effect is not being generated by those causes but instead by mental or physical intervention from outside the normal causal system. To illustrate, when an experimenter manipulates the brightness of a computer monitor, one should not assume that the monitor needs replacing.

The demonstrations all described a deterministic causal system. The undoing principle also applies to probabilistic causes however.

These data support the psychological reality of a central tenet of Pearl's (2000) causal modeling framework. The principle is so central because it serves to distinguish causal relations from other relations, such as mere probabilistic ones. The presence of a formal operator that enforces the undoing principle, Pearl's *do* operator, makes it possible to construct representations that afford valid causal induction and inference -- induction of causal relations that support manipulation and control and inference about the effect of such manipulation, be it from actual physical intervention or merely counterfactual thought about intervention. The *do* operation is precisely what's required to distinguish representations of probability like Bayes' nets from representations of causality.

More generally, the findings are consistent in a qualitative sense with the view of cognition assumed by Pearl (2000) following Spirtes, Glymour, and Scheines (1993). Their analysis starts with the assumption that people construe the world as a set of autonomous causal mechanisms and that thought and action follow from that construal. The problems of prediction, control, and understanding can therefore be reduced to the problems of learning and inference in a network that represents causal mechanisms veridically. Once a veridical representation of causal mechanisms has been established, learning and inference can take place by intervening on the representation rather than on the world itself. But none of this can be achieved without a suitable representation of intervention. The *do* operator is intended to allow such a representation and the studies reported herein provide some evidence that people are able to use it correctly.

Representing intervention is not always as easy as forcing a variable to some value and cutting the variable off from its causes. Indeed, most of the data reported here show some variability in people's responses. People are not generally satisfied to simply implement a *do* operation. People often want to know precisely how an intervention is taking place. A surgeon can't simply tell me that he's going to replace my hip. I want to know how, what it's going to be replaced with, etc. After all, knowing the details is the only way for me to know with any precision how to intervene on my representation, which variables to *do*, and thus what can be safely learned and inferred.

Causal reasoning is not the only mode of reasoning. But the presence of a calculus for causal inference removes any doubt that it's an important one.

Acknowledgments

This work was funded by NASA grant NCC2-1217. We thank Brad Love for his help and Daniel Mochon, Ian Lyons, and Peter Desrochers for collecting data. Josh Tenenbaum provided an important insight.

References

Lipton, P. (1992). Causation outside the law. In H. Gross & R. Harrison (Eds.), *Jurisprudence: Cambridge Essays.* Oxford: Oxford University Press.

Goldvarg, E., & Johnson-Laird, P.N. (2001). Na ve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25, 565-610.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991) *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mackie, J.L. (1974). *The cement of the universe*. Oxford: Oxford University Press.

Manktelow, K.I., & Over, D.E. (1990). Deontic thought and the Selection task. In K.J. Gilhooly, M. Keane, R.H. Logie, & G. Erdos (Eds), *Lines of Thinking, Vol. 1*, Chichester: Wiley.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge: The MIT Press.

Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

Formalizing Affordance

Mark Steedman (steedman@cogsci.ed.ac.uk)
Informatics, University of Edinburgh, 2 Buccleuch Place
Edinburgh EH8 9LW, Scotland UK

Abstract

The idea that to perceive an object is to perceive its affordances—that is, the interactions of the perceiver with the world that the object supports or affords—is attractive from the point of view of theories in cognitive science that emphasize the fundamental role of *actions* in representing an agent's knowledge about the world. However, in this general form, the notion has so far lacked a formal expression. This paper offers a representation for objects in terms of their affordances using Linear Dynamic Event Calculus, a formalism for reasoning about causal relations over events. It argues that a representation of this kind, linking objects to the events which they are characteristically involved in, underlies some universal operations of natural language syntactic and semantic composition that are postulated in Combinatory Categorial Grammar (CCG). These observations imply that the language faculty is more directly related to prelinguistic cognitive apparatus used for planning action than formal theories in either domain have previously seemed to allow.

Introduction

The notion of an *affordance* (Gibson 1966) has in its most basic sense of an invariant supporting perception been extremely helpful in directing attention to non-obvious properties of the sensory array relevant to visual and haptic perception, and motor control (Lee 1980; Turvey 1990). In its more general sense of an interaction with the world that a perceived object mediates (Gibson 1979) it has proved equally attractive to a wide range of theoretical positions that have emphasized the fundamental role of the notion of *action* in human cognition (Norman 1988, 1999). This is the sense in which a door "affords" egress and ingress, a knife affords cutting and scraping, and the like. The attraction of this notion is that it seems to offer a way in which perceptual learning can be linked to the goals and actions upon the environment of the learner, an idea that has been followed up by E. Gibson and Spelke (1983), among others. However, its influence in these domains has been limited by two difficulties.

One has been the controversial idea of "direct perception". This is the idea that the perception that a mailbox " affords letter-mailing to a letter-writing human in a community with a postal system" (Gibson 1979, p.139, citing Gibson 1950) is as directly related to properties of the sensory array as time-to-impact is to characteristics of the optic flow field for a diving gannet. It is certainly hard to believe that the perception of such affordances is "direct" in this strong sense, although recognition of mailboxes, like that of everything else, is undoubtedly mediated *in part* by such Gibsonian invariants of the optic array as relative spatial frequency spectra, and acquisition of the mailbox artefact concept unquestionably depends upon the association of such invariants with affordances in the more general sense. I shall ignore the perceptual aspect of affordances here.

A more serious obstacle to the exploitation of the idea of affordances in this general sense has stemmed from the very fact that many such affordances are actions or events. A formal theory of events in their relation to objects that is applicable to such perceptual categorization and/or conceptual representation of artefact concepts—that is, a theory of what the affordance itself actually is, and how it actually works as a basis for effective action in the world—has been lacking.

The Linear Dynamic Event Calculus

The Linear Dynamic Event Calculus (LDEC) combines the insights of the Event Calculus of Kowalski and Sergot (1986), itself a descendant of the Situation Calculus of McCarthy and Hayes (1969) and the STRIPS planner of Fikes and Nilsson (1971), with the Dynamic and Linear Logics that were developed by Harel (1984), Girard (1987) and others.

Dynamic logics are a form of modal logic in which the $\Box$ and $\Diamond$ modalities are relativized to particular events. For example, if a (possibly nondeterministic) program or command α computes a function F over the integers, then we may write the following:

(1) $n \geq 0 \Rightarrow [\alpha](y = F(n))$

(2) $n \geq 0 \Rightarrow \langle\alpha\rangle(y = F(n))$

The intended meaning of the first of these is "in any situation in which $n \geq 0$, after every execution of α that terminates, $y = F(n)$". That of the second is (dually) that "in any situation in which $n \geq 0$, there is an execution of α that terminates with $y = F(n)$".

We can think of these modalities as defining a logic whose models are Kripke diagrams in which accessibility between possible worlds is represented by events. Such events can be defined as mappings between situations or partially specified possible worlds, defined in

terms of conditions on the antecedent which must hold for them to apply (such as that $n \geq 0$ in (1)), and consequences (such as that $y = F(n)$) that hold in the consequent.

The particular dynamic logic that we are dealing with here is one that includes the following dynamic axiom, which says that the operator ; is *sequence*, an operation related to *functional composition* over events, viewed as functions from situations to situations:

(3) $[\alpha][\beta]P \Rightarrow [\alpha;\beta]P$

Using this notation, we can conveniently represent, say, a plan for *getting outside* as the composition of *pushing* a door and then *going through* it, written $push'; go\text{-}through'$.

Composition is one of the most primitive *combinators*, or operations combining functions, which Curry and Feys (1958) call **B**. It can be defined by the following equivalence with a lambda term:

(4) $\mathbf{B}\alpha\beta \equiv \lambda s.\alpha(\beta s)$

Plans like $push'; go\text{-}through'$ could be written in Curry's notation as $\mathbf{B} push' go\text{-}through'$

Situation/Event Calculi and the Frame Problem

The situation calculi are heir to a problem known in the AI literature as the Frame Problem (McCarthy and Hayes 1969). This problem arises because the way that we structure our knowledge of change in the world is in terms of event-types that can be characterized (mostly) as affecting just a few fluents among a very large collection representing the state of the world. (Fluents are facts or propositions that are subject to change). Naive event representations which map entire situations to entire other situations are therefore representationally redundant and inferentially inefficient. A good representation of affordances must get around this problem.

To avoid the frame problem in both its representational and inferential aspects, we need a new form of logical implication, distinct from the standard or intuitionistic $\Rightarrow$ we have used up till now. We will follow Bibel et al. (1989) and others in using *linear* logical implication $\multimap$ rather than intuitionistic implication $\Rightarrow$ in those rules that change the value of fluents.

For example, we can represent events involving doors in a world (simplified for purposes of exposition) in which there are two places *out* and *in* separated by a door which may be *open* or *shut*, as follows:[1]

(5) a. $shut(x) \multimap [push(y,x)]open(x)$
b. $open(x) \multimap [push(y,x)]shut(x)$

(6) a. $in(y) \multimap [go\text{-}through(y,x)]out(y)$
b. $out(y) \multimap [go\text{-}through(y,x)]in(y)$

Linear implication has the effect of building into the representation the update effects of actions—that once you apply the rule, the proposition in question is "used up", and cannot take part in any further proofs, while a new fact is added. The formulae therefore say that if something is shut and you push it, it becomes open (and vice versa), and that if you are in and you go through something then you become out (and vice versa).

To interpret linear implication as it is used here in terms of proof theory and proof search, we need to think of possible worlds as states of a single updatable STRIPS database of facts. Rules like (5) and (6) can then be interpreted as (partial) functions over the states in the model that map states to other states by removing facts and adding other facts. Linear implication and the dynamic box operator are here essentially used as a single state-changing operator: you can't have one without the other.

The effect of such systems can be exemplified as follows. If the initial situation is that you are in and the door is shut:

(7) $in(you) \wedge door(d) \wedge shut(d)$

—then the linear rules (5) mean that an attempt to prove the proposition in (8) concerning the state of the door in the situation that results from pushing the door will fail because rule (5a) has removed the fact in question from the database that results from the action $push(you,d)$:[2]

(8) $[push(you,d)]shut(d)$

On the other hand, attempts to prove the following will all succeed, since they are all facts in the database that results from the action $push(you,d)$ in the initial situation (7):

(9) a. $[push(you,d)]open(d)$
b. $[push(you,d)]door(d)$
c. $[push(you,d)]in(you)$

The advantage of interpreting linear implication in this way is that it builds the STRIPS treatment of the frame problem (Fikes and Nilsson 1971) into the proof theory, and entirely avoids the need for inferentially cumbersome reified frame axioms of the kind proposed by Kowalski (1979) and many others (see Shanahan 1997).

Using linear implication (or the equivalent rewriting logic devices or state update axioms of Thielscher (1999) and Martí-Oliet and Meseguer (1999)) for STRIPS-like rules makes such frame axioms unnecessary. Instead, they are theorems concerning the linear logic representation.

Even in this extremely simplified world, we need a little more apparatus to represent our knowledge about doors in a way which will allow us to make plans in-

[1] We follow a logic programming convention that all variables appearing in the consequent are implicitly universally quantified and all *other* variables are implicitly existentialy quantified. Since in the real world doors don't always open when you push them, box must be read as *default* necessity, meaning "usually".

[2] We follow the logic programming convention of negation by failure, according to which a proposition is treated as false if it cannot be positively proved to be true.

volving them. We also need to state preconditions on the actions of pushing and going through. Here ordinary non-linear intuitionistic implication is appropriate:[3]

(10) a. $door(x) \wedge open(x) \Rightarrow possible(go\text{-}through(y,x))$

b. $door(x) \Rightarrow possible(push(y,x))$

These rules say (oversimplifying wildly) that if a thing is a door and is open then it's possible to go through it, and that if a thing is a door then it's possible to push it.

We also need to define the transitive property of the possibility relation, as follows, using the definition (3) of event sequence composition:

(11) $possible(\alpha) \wedge [\alpha] possible(\beta) \Rightarrow possible(\alpha;\beta)$

This says that any situation in which it is possible to α, and in which actually doing α gets you to a situation where it is possible to β, is a situation in which it is possible to α *then* β.

If we regard actions as functions from situations to situations, then this rule defines *function composition* as the basic plan-building operator of the system. Composition is one of the simplest of a small collection of combinators which Curry and Feys (1958) used to define the foundations of the λ-calculus and other applicative systems in which new concepts can be defined in terms of old. Since the knowledge representation that underlies human cognition and human language could hardly be anything *other* than an applicative system of some kind, we should not be surprised to see it turn up as one of the basic operations of planning systems.

This fragment gives us a simple planner in which starting from the world (12) in which you are *in*, and the door is *shut* and stating the goal (13) meaning "find a possible series of actions that will get you *out*," can, given a suitable search control, be made to automatically deliver a constructive proof that one such plan is (14), the composition of *pushing*, and *going through*, the door:

(12) $in(you) \wedge door(d) \wedge shut(d)$

(13) $possible(\alpha) \wedge [\alpha] out(you)$

(14) $\alpha = push(you,d);go\text{-}through(you,d)$.

One way to produce this proof, which is suggested as an exercise, is via *backward-chaining* from the goal (13) on the consequents of rules (10) using the transitivity rule (11). The situation that results from executing this plan in the start situation (7) is one in which the following conjunction of facts is directly represented by the database:

(15) $out(you) \wedge door(d) \wedge open(d)$

This calculus is developed further in Steedman 1997, 2002 in application to more ambitious plans, such as the "monkey and bananas" problem, and a number of generalizations of the frame problem, using on a novel analysis of *durative* events extending over intervals of time, which are ignored here.

However, we have said nothing yet about the problem of *search* implicit in searching for and identifying such plans.

Formalizing Affordance using LDEC

Although the example is simplified for purposes of exposition (in particular, with respect to the problem of durativity), it provides the basis for a quite general calculus of events. (See Shanahan (1997), Thielscher 1999, and Steedman (1997, 2000b) for related proposals including discussions of ramification, qualification, delayed action, simultaneity, nondeterminism and other standard problems that such representations have to deal with.)

In fact the representation of actions and events in terms of an association of preconditions and consequences with the core event is a very generally applicable one. If the precondition is a conditional stimulus such as a light, and the consequence is a reward, such as food, while the action concerned is pecking or pressing a bar, then it can be considered as a representation of an *operant* in the cognitive sense of Rescorla and Wagner (1972), itself a notion closely related to that of an affordance.

It also provides the basis for a formalization of the relation between objects and their affordances, of the kind that we need in order to talk about perceptual and cognitive learning in non-linguistic animals and prelinguistic children. For example, the facts in (5) and (6) strike me as a pretty good representation of what my cat knows about the affordances of doors. Of course, the representation is perfectly neutral concerning the invariants that afford the perception of doors in the first place, their relation to bodily properties like the size of the cat's head, and aspects relevant to learning such as motor embedding of the actions of pushing and going through, and so on. It is a representation of what sort of thing it is that is perceived and learned. Nevertheless, the representation could be used to explain the transition she made in her perceptual learning from a stage where doors afforded her (6) (going through for purposes of egress and ingress) but not (5) (pushing to open and close), homing in via a set of superstitious and rapidly extinguishing spurious affordances to a correct affordance (5) supporting the motor plan (14) and its internalization as yet another affordance of doors. The representation also suggests a basis for experimentally investigating precise details of the cat's representation of the affordances of doors. (For example, do they afford her the ingress and egress of other cats?) Many of these experiments have already been done—most notably, by Köhler (1925), in his investigations of tool use and planning in Chimpanzees.

One of Köhler's most thought-provoking observations concerning such planning was the following. A chimpanzee which was perfectly capable of consistently using a tool such as a stick to reach otherwise unattainable objects—one to whom sticks afforded reaching—was unable to enact such a plan unless the stick was ac-

[3]The version of linear logic mixing linear and standard implication is is closely related to "Bunched Implication Logic" (see Pym 2001, which gives an extensive treatment of its semantics and proof theory, including a cut elimination theorem).

tually present in the problem situation. Mere availability of a stick in an adjoining room—even one which the ape had recently explored—was not enough to trigger the relevant knowledge and cause the ape to fetch the stick.

This observation suggests that for non-linguistic animals, including those closest to us in evolutionary terms, access to the affordances of objects is tied to immediate perception of the objects themselves, as Gibson believed. For an animal, this is quite a good way of running your planner. If you don't have much control over your physical environment, it is probably better to look at those plans the situation affords, rather than backward chaining to conditions that there may be no way for you to satisfy, say because of the time of year. This in turn suggests, uncontroversially, that affordances like egress are indexed in such animals by object-concepts like *door*, rather than by end-states like being *out*, and that planning proceeds *reactively* by forward chaining from what is the case, rather than backward chaining from the goal.

We can represent such indexing by first defining actions like *pushing* and *going through* as functions like the following derived from (5) and (6):

(16) a. $push(y,x) \rightsquigarrow \left\{ \begin{array}{l} shut(x) \multimap open(x) \\ open(x) \multimap shut(x) \end{array} \right\}$

b. $go\text{-}through(y,x) \rightsquigarrow \left\{ \begin{array}{l} in(y) \multimap out(y) \\ out(y) \multimap in(y) \end{array} \right\}$

(Here $\rightsquigarrow$ reads as "yields". The linear implication symbol $\multimap$ is overloaded to signify linear mapping of state to state accompanied by deletion and addition of facts. Implication is so closely related to functional mapping, and the functions in question are so closely related to the state update or rewrite axioms of the proof theory that this overloading seems unlikely to cause confusion.)

The set of such functions $Affordances(door)$ constitutes the affordances of doors:

(17) $Affordances(door) = \left\{ \begin{array}{l} push \\ go\text{-}through \end{array} \right\}$

The Gibsonian affordance-based door-schema $door'$ can then in turn be defined as a function mapping doors into (second-order) functions from their affordances like pushing and going-through to their results:

(18) $door' = \lambda x_{door}.\lambda p_{Affordances(door)}.px$

The operation of turning an object of a given type into a function over those functions that apply to objects of that type is another primitive combinator called $\mathbf{T}$ or *type raising*. As in the case of composition (4), the effect of this combinator can be defined by equivalence to the corresponding λ-term:

(19) $\mathbf{T}x \equiv \lambda p.px$

Accordingly, (18) can be rewritten:

(20) $door' = \lambda x_{door}.\mathbf{T}x$

Such a concept of doors is useful for reactive planning, and one can add more affordances to $Affordances(door)$ as one's experience increases. It seems quite likely that this is close to the way cats or at least chimpanzees conceptualize doors.

However, in human terms it is a somewhat stultifying representation, in that it restricts the concept to previously encountered events involving doors that one has somehow stumbled across. One would like to have the advantages in terms of efficiency of planning that thinking of objects in terms of their affordances allows, while also being able envisage novel uses for doors—for example, using one as a table, or as a raft—when circumstances demand it. In other words, one would like to be able to generalize (18) over a wider range of affordances, such as the affordances of natural kinds such as flat movable objects, or of other things that you can push and/or go through. However, there are reasons to think our ability to generalize very far beyond natural kinds and directly experienced affordances is quite limited. (For example, people find considerable difficulty in solving those irritating conundrums which require one to see that a pair of pliers affords the weight for a plumbline, or that the box that thumbtacks are packaged in affords a bracket that can be thumbtacked to the wall to provide a support for a candle.) It seems likely that the basis for such limited generalization is partly perceptual, and partly embedded in our modes of interaction with objects, as Gibson insisted.

Combinatory systems that include both composition and type raising are quite expressive—see Smullyan (1985, 1994) for discussion. They have the character of calculi for rebracketing and permuting terms in expressions. Such calculi are closely related to linear logic itself—see Lambek (1988) for discussion. In this connection it is interesting that the theory of Combinatory Categorial Grammar (CCG, Ades and Steedman 1982, Steedman 2000a) implies that the grammar of all languages involves both type-raising of argument categories and composition of predicates.

Combinatory Grammars

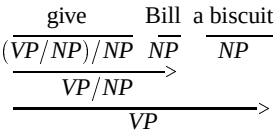
CCG, like other varieties of Categorial Grammar, is a theory in which all linguistic elements are categorized or typed as either functions or basic types, and in which syntactic derivation is achieved by syntactic rules corresponding to directionally and categorially restricted versions of a small number of combinators prominently including composition $\mathbf{B}$ and $\mathbf{T}$. Thus it is a theory that makes language look as if it has been built on a preexisting sytem for planning action in the world, and thereby seem less unique as a cognitive faculty than is usually assumed.

While readers must be directed elsewhere for a full presentation, it may suffice for present purposes to merely note that in CCG elements like verbs are associated with a syntactic "category" which identifies them as *functions*, and specifies the type and directionality of their arguments and the type of their result. For example, a ditransitive verb (DTV) is a function from (indirect object) NPs on the right into transitive verbs (TV)—that is, into functions from (direct object) NPs on the right into

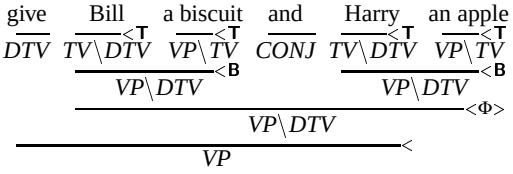
VP:[4]

(21) give := $(VP/NP)/NP$

Such a DTV is a (curried) function that can apply to its arguments to yield VP, as follows:

(22)


However, the involvement of further combinatory operations engenders a wide variety of coordination phenomena characteristic of all languages of the world, including English "argument-cluster coordination", "backward gapping" and verb-raising constructions in Germanic languages, and English gapping. The first of these is illustrated by the following analysis, from Dowty (1988):

(23)


The type-raising and composition rules, indicated by **T** and **B** repectively, guarantee that the semantics of non standard constituents like *Bill a biscuit* is such as to reduce appropriately with a ditransitive verb like *give*. It is in fact a prediction of the theory that such a construction can exist in English, and its inclusion in the grammar requires no additional mechanism whatsoever.

The earlier papers show that no *other* non-constituent coordinations of dative-accusative NP sequences are allowed in any language with the English verb categories, given the assumptions of CCG. Thus the following are ruled out in principle, rather than by stipulation:

(24) a. *Bill to Sue and introduce Harry to George
b. *Introduce to Sue Bill and to George Harry

Examples like (23) have often been described in terms of very powerful mechanisms of "deletion under identity" of missing elements like the verb *give* in the right conjunct. However, unlike CCG, such proposals fail to explain the observation that such deletions preserve word order, in the sense that in both coordinate and canonical sentences of English, *verbs are to the left of their complements*.

This observation is merely the English specific manifestation of a generalization concerning Universal grammar, due to Ross (1970), who noted that when verbs are "deleted" in this way in languages with other "basic" word orders, such as verb-final (SOV) and verb initial (VSO) languages, they always do so in a way that preserves the canonical left-to-right ordering of verb and argument, thus:[5]

(25) VSO: *SO and VSO VSO and SO
SOV: SO and SOV *SOV and SO

Logical and Neurological Relations between Language and Action

The ubiquitous appearance of composition **B** and type-raising **T** in both affordance-mediated action planning of the most elementary sort on the one hand, and universal grammar on the other, strongly suggests that the language faculty in its syntactic aspect is directly hung onto a more primitive set of prelinguistic operations including these combinators, originally developed for motor planning. This hypothesis has strong implications for the theory of evolution and the child's acquisition of language, for which there is considerable circumstantial evidence from neurological and neuroanatomical observations.

The Linear-Dynamic Event Calculus and related linear and STRIPS-like systems offer a way of representing actions in ways that are useful for planning action. This in turn offers a way of capturing affordances of objects, a notion that is relevant to doing so efficiently, and which is therefore relevant to perceptual categorization and concept learning relevant to tool-use. Two combinatory operations of composition and type-raising play a central role in this process. Those same combinators appear in syntactic guise in natural language, where they provide the basis for an explanatory account of language-specific constructions and cross-linguistic universal generalization, and where a considerable body of evidence from neuroanatomy and child development that has been adduced in support of the Motor Theory suggests that planning and language are closely related. LDEC and CCG make that relation look direct enough to explain the fact that the evolutionary advance in question appears to have been very rapid indeed.

It is interesting to speculate upon what such an evolutionary step might be based. One strong candidate is the attainment of the modal and propositional attitude concepts that are necessary to support a theory of other minds—that is, functions over propositional entities. (We have so far glossed over an important distinction between plans, which compose actions of type *state* → *state*, and grammar, which composes functions of type *proposition* → *proposition* or *property* → *property*.)

It is propositional functions that induce true recursion in both conceptual structures and grammar. There is no evidence that apes entertain such concepts. In particular, the most successful attempts to teach apes to use language, notably those involving ASL and other manipulative languages, show a lack of recursive syntax coupled

[4] We here use the "result leftmost" notation in which a rightward-combining functor over a domain β into a range α are written α/β, while the corresponding leftward-combining functor is written α\β. (α and β may themselves be function categories.) There is an alternative "result on top" notation, according to which the latter category is written β\α.

[5] Interestingly, SVO languages like English pattern with verb initial languages in this respect, rather than with verb final. This fact and certain apparent exceptions to Ross's generalization arising in languages with more than one "basic" word order are discussed in Steedman 2000a.

with an almost autistic paucity of conversational initiative. Perhaps it is *only* the lack a theory of mind and the associated propositional attitude concepts that holds apes back from developing human language on the basis of their planning abilities, a suggestion consistent with the views of Tomasello 1999.

Acknowledgments

Thanks to to Ric Alterman, Silvia Gennari, Joyce McDonough, Michael Ramscar, Matthew Stone, and Rich Thomason for comments and advice on a draft. Various stages of the research were funded in part by EPSRC grants GR/M96889 and GR/R02450 and EU (FET) grant MAGICSTER.

References

Ades, A. and Steedman, M. (1982). On the order of words. *Linguistics and Philosophy*, 4:517–558.

Bibel, W., del Cerro, L. F., Fronhfer, B., and Herzig, A. (1989). Plan generation by linear proofs: on semantics. In *German Workshop on Artificial Intelligence - GWAI'89*, volume 216 of *Informatik-Fachberichte*, Berlin. Springer Verlag.

Curry, H. B. and Feys, R. (1958). *Combinatory Logic: Vol. I*. North Holland, Amsterdam.

Dowty, D. (1988). Type-raising, functional composition, and nonconstituent coordination. In Oehrle, R. T., Bach, E., and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 153–198. Reidel, Dordrecht.

Fikes, R. and Nilsson, N. (1971). Strips: a new approach to the application of theorem proving to problem solving. *AI Journal*, 2:189–208.

Gibson, E. and Spelke, E. (1983). The development of perception. In Mussen, P., editor, *Handbook of Child Psychology, vol. 3: Cognitive Development*, pages 1–76. Wiley, New York.

Gibson, J. (1950). *The Perception of the Visual World*. Houghton-Mifflin Co., Boston, MA.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Houghton-Mifflin Co., Boston, MA.

Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Houghton-Mifflin Co., Boston, MA.

Girard, J.-Y. (1987). Linear logic. *Theoretical Computer Science*, 50:1–102.

Harel, D. (1984). Dynamic logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume II, pages 497–604. Reidel, Dordrecht.

Köhler, W. (1925). *The Mentality of Apes*. Harcourt Brace and World, New York.

Kowalski, R. (1979). *Logic for Problem Solving*. North Holland, Amsterdam.

Kowalski, R. and Sergot, M. (1986). A logic-based calculus of events. *New Generation Computing*, 4:67–95.

Lambek, J. (1988). Categorial and categorical grammars. In Oehrle, R. T., Bach, E., and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 297–317. Reidel, Dordrecht.

Lee, D. (1980). The optic flow field: The foundation of vision. *Philosophical Transactions of the Royal Society, Series B*, 290:169–179.

Martí-Oliet, N. and Meseguer, J. (1999). Action and change in rewriting logic. In Pareschi, R. and Fronhöfer, B., editors, *Dynamic Worlds*, pages 1–53. Kluwer, Dordrecht.

McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence*, volume 4, pages 473–502. Edinburgh University Press, WEdinburgh.

Norman, D. (1988). *The Psychology of Everyday Things*. Basic Books, New York.

Norman, D. (1999). Affordance, conventions, and design. *Interactions*, 6:38–43.

Pym, D. (2001). *The Semantics and Proof Theory of the Logic of Bunched Implications*. to appear.

Rescorla, R. and Wagner, A. (1972). A theory of pavlovian conditioning. In Black, A. and Prokasy, W., editors, *Classical Conditioning, II*. Appleton-Century-Crofts, New York.

Ross, J. R. (1970). Gapping and the order of constituents. In Bierwisch, M. and Heidolph, K., editors, *Progress in Linguistics*, pages 249–259. Mouton, The Hague.

Shanahan, M. (1997). *Solving the Frame Problem.* MIT Press, Cambridge.

Smullyan, R. (1985). *To Mock a Mockingbird*. Knopf, New York.

Smullyan, R. (1994). *Diagonalization and Self-Reference*. Clarendon Press, Oxford.

Steedman, M. (1997). Temporality. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 895–938. North Holland, Amsterdam.

Steedman, M. (2000a). *The Syntactic Process*. MIT Press, Cambridge, MA.

Steedman, M. (2000b). *The Productions of Time*. Ms., University of Edinburgh, `http://www.cogsci.ed.ac.uk/~steedman/`.

Steedman, M. (2002). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25:(to appear).

Thielscher, M. (1999). From situation calculus to fluent calculus: State update axioms as a solution to the inferential frame problem. *Artificial Intelligence*, 111:277–299.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.

Turvey, M. (1990). Coordination. *American Psychologist*, 45:938–953.

Providing Distinctive Cues to Augment Human Memory

Jeanine K. Stefanucci (jks8s@virginia.edu)
Department of Psychology, 102 Gilmer Hall, Box 400400
Charlottesville, VA 22904-4400

Dennis R. Proffitt (drp@virginia.edu)
Department of Psychology, 102 Gilmer Hall, Box 400400
Charlottesville, VA 22904-4400

Abstract

Previous research in our lab (Tan, Stefanucci, Proffitt & Pausch, 2001) demonstrated that a multimodal prototype computer system, the InfoCockpit, could increase users' memory of information compared to a standard desktop computer. Displaying information on multiple monitors with ambient visual and auditory dispays engages context-dependent memory and memory for location, thus facilitating recall. We replicate this finding and isolate the memory cues to find whether the combination of contextual information and spatial location is necessary to obtain this memory advantage. Our findings show that contextual information alone provides users with the best strategy for later recall.

Introduction

In the past years, computer interfaces have been designed with the goal of promoting usability. These interfaces have a consistent "look and feel" that fosters usability but does not help the user remember information learned on the system. Our research examines a newly built interface, termed the InfoCockpit, which supports and aids human memory and performance while preserving usability.

The design of the InfoCockpit is based on psychological research that has uncovered many ways of improving memory through the use of spatial and environmental memory cues. These cues are incorporated into the InfoCockpit so that users can more easily recall information that they learn on the computer. This system provides users with "locations" and "places" to hook their memories onto without compromising usability.

Creating Place

Memories are tied to the environmental context in which they take place (Smith, Glenberg, & Bjork, 1978). For example, one might try to help a friend remember a conversation by referencing the context of that conversation (e.g. "don't you remember we talked about this at the coffee shop downtown?"). Having recalled the place of the conversation, the friend can more easily remember what was said. This strategy recruits an important cue for human memory; the context or "place" is a reference to start a search for the information discussed. Being in places, or referencing them, evokes memories and increases the chances of remembering information.

Psychologists have researched the use of environmental context as a cue for memory for the past few decades (Godden & Baddeley, 1975; Smith, Glenberg & Bjork, 1978). Smith (1979) found that people associate information and the environmental context in which it is learned. Although these associations are often incidental, they can be useful retrieval cues when recalling information. Smith (1982) also had participants encode information in multiple learning environments or different "places". He showed that the amount of information recalled increases when learning takes place in different contexts. In further studies, however, Smith (1984) found that recall performance in multiple learning contexts was not significantly improved when participants returned to the place that they were in at the time of encoding. Diverse learning environments provide a memory advantage over a single learning environment but this advantage is not contingent upon reinstatement of the context at retrieval.

In addition to the number of learning environments, contexts that are distinctive can also increase memory performance. Places that draw attention are the most effective in producing a memory advantage (Smith, Vela, & Williamson, 1988). Learning information through different sensory modalities can create a distinctive context. In addition to visual cues, ambient three-dimensional sounds can serve as distinctive cues for memory. It has been shown that ambient sounds enhance memory for visual information presented in their context (Davis, Scott, Pair, Hodges, & Oliverio, J., 1999).

Providing Location

Memories are also tied to a location in space (Gordon, 1903). Whereas we use "place" to denote an ambient environmental context, "location" refers to the position of information within that "place". We cannot help but

notice, for example, the position of an object in a room. The location of that object in space is processed preattentively and remembered almost automatically (Logan, 1998). Similarly, most people comment that they can remember where on a page they read something without remembering the information they read. Several studies confirmed this anecdote by showing the reliability of spatial location as an important memory cue (Rothkopf, 1971; Zechmeister & McKillip, 1972). Given the evidence above, it follows that spatially distributed information is easier to remember than information presented in a single location (Gordon, 1903).

Combining Location and Place: The InfoCockpit

The InfoCockpit (see Figure 1) uses multiple projectors to display a panoramic image of a "place" onto large screens. It provides a context for users to reference when they are retrieving information from memory. Ambient three-dimensional surround sound is added to immerse the user in the place. For example, panoramas of a woodland scene are projected with consistent 3D sounds like leaves rustling, birds chirping, and insects. The InfoCockpit provides spatial cues by presenting information to users on multiple monitors. When learning the information, users inadvertently notice on which monitor information is presented. We hypothesized that users would be more likely to remember the information if they could recall on which monitor it was presented.

Figure 1. – The InfoCockpit uses multiple monitors to provide "location" cues and ambient visuals with three-dimensional sounds to create "place" cues.

This system stands in stark contrast to current desktop systems, which present all information to the user on a single monitor, and do not display a place cue. Desktop users do not have to orient themselves to information; windows simply bring information to them. There are no spatial cues encoded with the information and no way of easily retrieving information by remembering the context in which it was seen.

Previous research has not attempted to construct environments that present "location" and "place" cues to systematically examine whether a large effect can be obtained. In our lab, we combined these cues (location and place) to see if they produced a greater memory advantage than when presented independently (Tan, Stefanucci, Proffitt, Pausch, 2001). Tan et al. found that users of the InfoCockpit had a 56% improvement in memory performance when compared to users of a standard desktop computer.

The current paper addressed whether users of the InfoCockpit systematically relied on one cue over the other or if the combination of "location" and "place" was the best way to promote later recall. Each of the cues was examined in isolation to assess its solitary contribution to the larger effect. Based on our previous findings, we assumed that participants using the InfoCockpit would be able to remember more than participants using a standard desktop computer. And indeed, this was true. In addition, we hypothesized that "more is better"; participants in the InfoCockpit condition would remember significantly more word pairs than participants who received "location" or "place" cues in isolation. Contrary to our hypothesis, our experiment revealed that participants receiving only "place" cues performed significantly better than participants in all other conditions.

Method

Participants

Eighty University of Virginia students (40M, 40F) participated in the experiment. Participants were paid $20 for their participation. They were naïve to the purposes of the experiment and had not participated in a previous memory experiment like this one.

Apparatus

The apparatus used to display materials, the InfoCockpit, is a large multiple screen display system (see Figure 1). The displays are run from a Dell Precision Workstation with 620 Pentium III Xeon dual processors. Installed in the Dell are two Appian Jeronimo Pro 4-port graphics cards that allow the computer to drive the six display screens. Two sets of displays are arrayed three across, with NEC 18" LCD monitors directly below the projection surfaces. The LCD monitors serve as the main working area on which users interact with information. The projection displays provide a horizontal viewing angle of approximately 145 degrees and are used to immerse the user in a particular place. Three Sharp Notevision 6 projectors (2200 lumens) display the context images on the projection screens.

We created and played back audio contexts on a Macintosh G4 using a Digidesign Pro Tools Mix24

digital audio workstation. The contextual environments were comprised of 6 channels of sound. Speakers were placed surrounding the user at ear level and at 4 feet above ear level, +/- 30 degrees at ear level, and +/- 120 degrees at 4 feet above ear level. The ear level speakers were 5 feet away from the user while the speakers above were 8 feet away.

Procedure

The experimental design consisted of two phases: a training phase and a testing phase.

Training Phase Participants learned three lists of words, each list containing ten pairs of words (all common, high frequency nouns). All participants learned the lists one at a time. Each list consisted of 10 cue words and 10 target words. The 10 cue words were the same for the 3 lists, but the target words varied from list to list (i.e. 'plate-passenger', 'plate-string', and 'plate-scientist'). We named the lists Lawn, Museum, and History to help the participants parse the lists in memory. For the participants in the InfoCockpit or Context conditions, these names referred to projected places.

The training phase consisted of both a study period and a learning period. During the study period participants were presented with each pair of words once (for 5 seconds each). After study was completed, the learning task began. One of the cue words from a pair was randomly presented to the participants. Participants then typed the target word that went with the cue. Feedback was given to the participants. If they were incorrect, the correct word was presented. Another cue word would then appear and they would have to type its target. This went on iteratively until participants had recalled all of the pairs correctly in one iteration (meeting 100% criteria). This ensured that all participants' knowledge of the material was equivalent before testing. Participants had unlimited time to finish the learning portion. The procedure for learning the word lists was explained to participants before training began.

Participants were assigned to one of four conditions (Desktop, InfoCockpit, Spatial, or Context) defined by the display configuration on which they learned pairs of words. An equal number of males and females were randomly assigned to each condition. Participants in the InfoCockpit group studied and learned the word lists, each on a different monitor and associated with different contextual images and sounds. For example, participants would see images and hear sounds from a museum while they were learning the word pairs for the "Museum" list on the middle monitor.

The Desktop group performed the task on a standard desktop computer with a single monitor. They learned the same three lists (Lawn, Museum, and History) on one screen, with no projected context images or three-dimensional sounds.

For the Spatial condition, participants learned the three word lists on different monitors. However, they did not learn the lists in different projected contexts. They also had no sounds. This condition was designed to assess the individual contribution of spatial cues in the InfoCockpit.

Participants in the Context condition learned the three lists on one monitor. Each list was presented with its corresponding context. Participants learned the lists with the projected context images and sounds, all on the same monitor. This condition tested the importance of contextual place cues on learning in the InfoCockpit.

Testing Phase The testing phase of the experiment took place a day later. All participants returned to the lab and were tested on how many word pairs they remembered from the training phase. Retrieval was done on a laptop in a different room than the training phase. Participants were given cue words from each of three lists, one list at a time, and were asked to type in as many of the targets as they could remember. They received no feedback on their performance.

Results

Out of 30 possible items, the mean recall scores were 8.80 for the Desktop group, 11.65 for the InfoCockpit group, 9.60 for the Spatial group, and 15.05 for the Context group. A one-way analysis of variance (ANOVA) comparing the four conditions revealed significant differences between the groups, $F(3, 76) = 9.065$, $p < 0.000$ (see Figure 2). Post hoc comparisons using the Fisher LSD test showed that participants in the Context condition recalled significantly more word pairs than the Spatial, Desktop and InfoCockpit conditions. In addition, the InfoCockpit condition remembered significantly more word pairs than the Desktop condition. A one-way analysis of variance (ANOVA) comparing number of iterations to learn the word pairs to criteria did not show a significant difference among the conditions, $p = 0.762$.

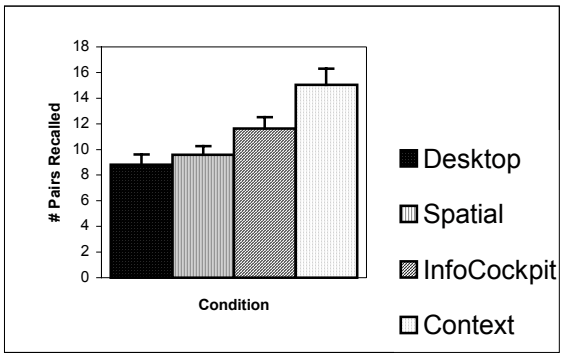


Figure 2. Number of word pairs recalled by each condition.

Discussion

InfoCockpit participants remembered significantly more word pairs than participants using a standard desktop computer. Adding "location" and "place" cues to a computer enhanced participants' memory for information learned on that system. This conclusion replicates previous research from our lab (Tan, Stefanucci, Proffitt, Pausch, 2001).

We assumed the individual components of the InfoCockpit would improve memory with relation to the standard desktop computer, but we did not believe they would be as effective as the ensemble of cues. Our hypothesis that "more is better" was incorrect for this task. Participants receiving only "place" cues at encoding recalled more words than participants in all other conditions.

This finding may be a consequence of the strategy the InfoCockpit participants used when retrieving the word pairs. We believe these participants had two strategies available at recall for remembering the information. One of the strategies involved the location of the list on one of the monitors. Participants could retrieve the appropriate target by recalling the location of the monitor on which it was learned. Likewise, InfoCockpit participants could access environmental place cues to recall the pairs. To remember the target word they could imagine the contextual images and sounds displayed when learning the pair. It is possible that these two recall strategies interfered with each other, compromising the best recall strategy (simply recalling the "place" cue) by the evoking the less effective location recall strategy. In the task we describe, the contextual place information was a more reliable cue for later recall and those participants in the Context condition were able to exploit it to the fullest.

While participants in the InfoCockpit condition performed better than participants in the standard desktop computer, they were not as successful as participants in the Context condition. Providing users with a single cue in isolation (place) was more effective than providing them with two sets of cues. This finding is not surprising given previous research. For instance, Jones (1976) and Smith (1984) found that isolated cues could help their participants retrieve an entire memory, even when combinations of cues were present at encoding. More recently, Parker and Gellatly (1997) showed that an increase in the amount of cues at encoding did not produce a reliable increase in recall. They gave their participants both music and odors while encoding information. At retrieval, either both or only one of the cues was reinstated. In either condition, participants recalled the same amount of information. In contrast, our findings suggest that participants receiving only one of the cues (place) at encoding had an advantage over the other conditions. The type of cue we used may account for the difference. Perhaps the place cue was more distinctive than the location cue and people were more successful in associating it with the words. The "more is better" approach to the design of computer interfaces should be examined closely because there may be situations in which a "less is more" attitude can augment performance to a higher degree.

Conclusions

The InfoCockpit increased memory compared to a standard desktop computer. It utilized both "place" and "location" cues to facilitate memory retrieval. When presented with "place" cues in isolation, participants' memory performance increased significantly in comparison to performance on the InfoCockpit. Providing multiple memory cues at encoding increases recall. However, interference between contextual and spatial cues may have a negative effect on performance. Evaluation of the interactions between cues, as well as the cues themselves, is necessary to ensure a complete understanding of the role that these cues play in memory.

Acknowledgments For their collaboration on this project, we would like to thank Tom Banton, Cedar Riener, and Jessi Witt of the Proffitt Perception Lab at the University of Virginia; Barbara Spellman at the University of Virginia; Adam Fass, Andrew Faulring, Desney Tan and Randy Pausch of the Stage 3 Research Lab at Carnegie Mellon University; Chris Kyriakakis at the University of Southern California. Many thanks to Shawn O'Hargan and Jae Lee for their help in collecting the data.

References

Davis, E. T., Scott, K., Pair, J., Hodges, L. F., & Oliverio, J. (1999). Can audio enhance visual perception and performance in a virtual environment?

Proceedings Of The Human Factors And Ergonomics Society 43rd Annual Meeting, 1197-1201.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325-331.

Gordon, K. (1903). Meaning in Memory and in Attention: Memory as dependent upon the complexity of its content. *Psychological Review, 11*, 267-283.

Jones, G. V. (1976). A fragmentation hypothesis of memory; cued recall of pictures and sequential position. *Journal of Experimental Psychology: General, 105*, 277-293.

Logan, G. D. (1998). What is learned during automatization? II. Obligatory encoding of spatial location. *Journal of Experimental Psychology: Human Perception & Performance, 24*(6), 1720-1736.

Parker, A., & Gellatly, A. (1997). Moveable cues: A practical method for reducing context-dependent forgetting. *Applied Cognitive Psychology, 11*(2), 163-173.

Rothkopf, E. Z. (1971). Incidental memory for location of information in text. *Journal of Verbal Learning and Verbal Behavior, 10*, 608-613.

Smith, S. M., Vela, E., & Williamson, J. E. (1988). Shallow input processing does not induce environmental context-dependent recognition. *Bulletin of the Psychonomic Society, 26*, 537-540.

Smith, S. M. (1984). A comparison of two techniques for reducing context-dependent forgetting. *Memory & Cognition, 12*, 477-482.

Smith, S. M. (1982). Enhancement of recall using multiple environmental contexts during learning. *Memory & Cognition, 10*(5), 405-412.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460-471.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory& Cognition, 6*(4), 342-353.

Tan, D.S., Stefanucci, J.K., Proffitt, D.R., & Pausch, R. (2001). The InfoCockpit: Providing Location and Place to Aid Human Memory. *Workshop on Perceptive User Interfaces 2001*, *Orlando, Florida.*

Zechmeister, E. B., & McKillip, J. (1972). Recall of place on the page. *Journal of Educational Psychology, 63*, 446-453.

Naive Strategic Thinking

Eugenia Steingold (genya@wjh.harvard.edu)
Department of Psychology
Harvard University
33 Kirkland Street
Cambridge, MA 02138

P. N. Johnson-Laird (phil@princeton.edu)
Department of Psychology
Princeton University
Green Hall
Princeton, NJ 08540

Abstract

The mental model theory postulates that reasoners build mental models of the situations described in premises, and that each model represents a possibility. The theory's main principle — the principle of *truth*— is that these representations are incomplete, because mental models represent only what is true, and not what is false. The present paper defends an analogous principle of *self-interest*: when individuals have to think strategically, they tend to represent only their own options and payoffs, not those of their opponents. Four experiments have corroborated this principle.

Introduction

Strategic thinking underlies those decisions that depend on what other individuals are likely to decide. Such decisions are ubiquitous in business, politics, and daily life. To outwit an adversary, or to maximize the benefits of co-operation with a partner, you often need to think about what this other individual is likely to choose to do. The theory of games provides a normative theory of strategic thinking, especially since John Nash's theorem determining those strategies that are optimal for all players — the so-called Nash equilibrium (see, e.g., Dixit and Nalebuff, 1991). However, as Reinhard Selten, co-winner of the Nobel prize with Nash in 1994, has remarked: game theory is rational theology (personal communication). That is, na ve individuals in daily life are unlikely to abide by its canons of rationality. The term na ve here refers to individuals who have not mastered game theory; it does not impugn their intelligence. Thinking about what other people may be thinking is indeed likely to be difficult, because it calls for a representation, not only of your own state of mind, but also of another person's preferences, and for inferences about their likely strategies. If you fail to think through their optimal strategy, then you are unlikely to select a rational strategy for your self.

Consider as an example a simple game in which you have two options (A and B), your opponent has two options (A and B), you both make your choices simultaneously, and you both receive payoffs as a function of your choices. We show your payoff followed by your opponent's payoff in each cell:

		Your opponent's options	
		A	B
Your options	A	**$4/$4**	**$1/$3**
	B	**$2/$4**	**$6/$1**

If your opponent chooses A then you should choose A and receive $4. But, if your opponent chooses B, you should choose B and receive $6. Hence, to make the right choice you should think about what your partner is likely to choose. If she is rational, she should think: If I choose A, then regardless of what my opponent chooses, I get $4; but if I choose B, then regardless of what my opponent chooses, I will get less than $4. Therefore, I should choose A (because B is *dominated* by this strategy). Hence, if you are rational and you know that your opponent is rational, then you should choose A. This sort of thinking exemplifies game theory, which formalizes optimal strategies on the assumption that players are rational.

In contrast, the present paper argues that the task of inferring rational predictions about other players choices is too difficult for naive human players. The paper accordingly proposes an alternative account of how na ve individuals make strategic decisions.

Mental Models and Na ve Strategic Thinking

The mental model theory postulates that

reasoning depends on understanding the meaning of premises and using this meaning and general knowledge to envisage the states of affairs that are possible given the truth of the premises (Johnson-Laird and Byrne, 1991). Each mental model represents a possibility. A conclusion is necessary if it holds in all the models of premises. It is possible if it holds in at least one model of premises (Bell and Johnson-Laird, 1998). And its probability — its likelihood of being true — depends on the proportion of models of the premises in which it is true (Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni, 1999). The theory has also been extended to account for meaning and reasoning with causal relations (Goldvarg and Johnson-Laird, 2001). A fundamental assumption of the theory is the principle of *truth*: in order to minimize the load on working memory, mental models normally represent only what is true. The failure to represent what is false gives rise to illusory inferences of various sorts, i.e. inferences that nearly everyone makes, but that are wrong (see, e.g., Goldvarg and Johnson-Laird, 2000).

The model theory extends naturally to human strategic thinking. It postulates that individuals construct mental models of the options and payoffs. In order to minimize the load on working memory, however, the theory is based on an assumption analogous to the principle of truth. According to the principle of *self-interest*, mental models normally represent only individuals own options and payoffs. The failure to represent other players payoffs should give rise to systematic errors in strategic thinking. The aim of our experiments was to test the principle of self-interest.

Experiment 1: Games as payoff matrices

Our first experiment examined whether or not individuals spontaneously represent their partner s payoffs in games in which their optimal choice depends on their partner s choice. The participants played 25 games, each presented in the form of a payoff matrix. Five of the games were *independent*, that is, the participants optimal choice did not depend on their partners choice. Here is an example of an independent game in which only the participant's payoffs are shown:

		Your partner's options	
		A	B
Your options	A	**$5**	**$6**
	B	**$4**	**$5**

In the absence of information about their partner's payoffs, there are three possible strategies that the participants could adopt:

1) They could maximize their mean payoffs. In this game it is option A, which leads to a mean payoff of $5.5.

2) They could maximize their minimum gain. Option A yields the larger minimum gain (of $5).

3) They could maximize their maximum gain. Option A yields the maximum gain (of $6). Hence, in general, with independent games all three strategies lead to the same choice.

The remaining 20 games were *dependen*t, that is, to make the optimal choice, the participants needed to know their partner's choice, e.g.:

		Your partner's options	
		A	B
Your options	A	**$5**	**$7**
	B	**$6**	**$4**

Here, if the partner chooses A, then the participant should choose B, whereas if the partner chooses B, the participant should choose A. Granted the principle of self-interest, however, the participants should not notice the difference between independent and dependent games. They should be prepared to play both sorts of game without knowledge of their partner's payoffs. In the present game, the three preceding strategies all lead to the choice of option A. The experiment included three other sorts of dependent game that allowed us to identify the participants' strategies.

We tested 20 Princeton undergraduates, and in this and the other experiments, we checked that they were na ve about game theory. The instructions explained that the participants would be presented with simple games, and that their task was to decide which option they would choose in each game. Before the participants responded to each game, they were asked: Do you have all the necessary information to play the game? If *No*, what else do you need to know to make the decision? The participants then made a choice.

Results

The majority of the participants responded that they had the all the necessary information to play the game: only five participants requested their partners' payoffs on more than half the trials. Hence, the bias to play the games without knowing these payoffs was reliable (Sign test, $n = 20$, $p < .025$). There was no reliable difference between dependent games and independent games in the tendency for participants to play in ignorance of their partners' payoffs (15 participants in both cases played on more than half the trials; $F < 1$, $p > .25$). The preferred strategy was to maximize the maximum possible gain (14 out of 20 participants on more than half the trials, $p < .05$).

Experiment 2: Conditional descriptions of games

The previous experiment presented games in the form of payoff matrices. Skeptics could argue that this format is not familiar enough to na ve participants to enable them to do justice to their ability. We therefore carried out a similar experiment, but each game was presented in a set of conditional assertions, which are easy to understand. For example, an independent game was described in the following way:

> If you choose A and your partner chooses A, you will receive \$5.
>
> If you choose A and your partner chooses B, you will receive \$6.
>
> If you choose B and your partner chooses A, you will receive \$4.
>
> If you choose B and your partner chooses B, you will receive \$5.

One group of participants received such descriptions, which are incomplete because they do not say anything about the partner's payoffs. A second group of participants received only partial descriptions of their own payoffs, e.g:

> If you choose A and your partner chooses A, then you will receive \$5.
>
> If you choose B and your partner chooses B, then you will receive \$6.

These descriptions were only for cases in which both players chose the same options (AA and BB). A third group of participants received descriptions of only their partner's payoffs but not their own, e.g.:

> If you choose A and your partner chooses A, your partner will receive \$5.
>
> If you choose A and your partner chooses B, your partner will receive \$6.
>
> If you choose B and your partner chooses A, your partner will receive \$4.
>
> If you choose B and your partner chooses B, your partner will receive \$5.

According to the principle of self-interest, the participants should be more likely to ask for their own payoffs than for the payoffs of their partners. The participants in each group were asked three questions: Would you play the game? Do you have all the necessary information to make a choice? If *No*, what else do you need to know to make your choice? The procedure and the 25 games were the same as those in Experiment 1. We tested ten Princeton undergraduates in each group.

Results

In the group that knew only their own payoffs, 100% of responses were decisions to play both dependent and independent games. In the group that knew only their own partial payoffs, 100% of responses were decisions to play both sorts of games. But, as predicted, in the group that knew only their partners' payoffs only 40% of the responses (4 participants on more than half the trials) were decisions to play the games. The difference between the three groups is significant, Kruskal-Wallis ($\chi 2 = 14.5$, $p < .001$). Hence, the majority of participants who had only their partners payoffs refused to play the games without knowing their own payoffs. The participants' preferred strategy was to maximize their average payoff (80% of strategies, Binomial test, $p < .005$).

Experiment 3: Games with real opponents

The participants in the previous experiments might have played games without knowing their partners' payoffs because the games seemed artificial and unreal. We therefore carried out a replication of Experiment 2 in a way that was closer to real games. The participants played against real opponents and they could win real money. Each experimental session tested two participants at a time, who had not previously met. They were told that they would be playing a set of games against each other. And they then went to different rooms, and each participant received the instructions and carried out the games. The design and procedure were otherwise identical to Experiment 2. We tested 10 Princeton undergraduates in each group. The participants were told that the person who won the most nominal money in the games would enter a lottery with a possibility to win \$20.

Results

The step towards verisimilitude slightly enhanced performance, but the results otherwise replicated the previous experiment. Two participants who knew only their own payoffs requested additional information on more than half the trials. Five participants who knew only their opponents' payoffs requested additional information on more than half the trials. Nine participants who knew only their own partial payoffs requested additional information (but only about the rest of their own payoffs). The difference between the three groups was significant (Kruskal-Wallis $\chi 2 = 11.21$, $p < .005$). A planned comparison between those who knew their own payoffs and those who knew their opponents' payoffs was also significant ($z = 3.2$, $p < .03$). The participants' preferred strategy was again to maximize their average payoff (70% of participants, $p < .005$). Hence, on the whole, individuals who knew their own payoffs were happy to play with real partners and for real money.

Experiment 4: The effect of losses.

In the previous experiments, none of the games threatened the participants with a loss of money depending on the outcome of a game. People are known to think differently about losses than about gains. They are risk seeking when they think about losses but risk averse when they think about gains (Tversky and Kahneman, 1981). Gains and losses are represented asymmetrically. The negative effect of a loss of a certain amount is greater than the positive effect of a gain of the same amount. Hence, people may represent games with losses differently from the way they represent games with only gains. Likewise, losses may also affect individuals' strategic thinking. In particular, they might think more carefully and require more information about losses than they do when they think about gains. This greater care may, in turn, elicit a need for information about their partner's options and payoffs. A different possibility, however, is that potential losses would make individuals anxious and confused. Hence, they might represent less information than usual, and make even fewer requests for their partner's payoffs. To examine these possibilities, Experiment 4 used a set of games similar to those in the previous experiments, but it included payoffs that were losses. One group of participants played games with gains only in the first block of the experiment and games with gains and losses in the second block. Another group received the two blocks in the opposite order. The procedure was similar to Experiment 3, using both real partners and a real monetary reward.

Results

The participants who received only gains in the first block were more likely to play in that block (97% of responses) than the group who received gains and losses in the first block (80% of responses in the first block, Kruskal-Wallis, $p < .005$). Overall, there was no difference between the first and second blocks in the propensity to play. But, the effect of loss in inhibiting participants from playing was greater in the second block than in the first block (Kruskal-Wallis $\chi 2 = 10.4$, $p < .05$).

Both groups were equally likely to request additional information (36% and 35% of responses). There was no significant effect of block. However, there was again a significant interaction between the group and the block, Kruskal-Wallis ($\chi 2 = 10.5$, $p < .05$). Only one participant who received only gains in the first block requested more information on more than half the trials; whereas six of the participants who received gains and losses in the first block requested more information on more than half the trials (Kruskal-Wallis $\chi 2 = 11.2076$, $p < .005$). Thus, the participants did not notice that something was missing when they played games with gains only. But, when they moved on to games with gains and losses, they tended to notice that the games were incomplete. In contrast, those who received games with gains and losses showed no change in their requests for additional information when they moved on to games with only gains (36% requests in the first block, and 36% requests in the second block). In sum, losses had their largest effects when they appeared after the games with gains only. The participants preferred strategy was again to maximize average payoffs (85% of strategies, $p < .005$).

General Discussion

Our everyday experiences often call for representing what other people think, believe and desire. Strategic decision making is one of many situations that call for individuals to think about their partners preferences and choices. Moreover, they often have to take into account their opponents considerations of their own options, and so on and on. This problem may easily become intractable, and, at the very least, highly complex. The present paper has defended

the principle of *self-interest*: when individuals have to think strategically, they tend to focus on their own options and payoffs. The experiments corroborated this principle. Experiment 1 demonstrated the effect with payoff matrices. Experiment 2 replicated it with conditional descriptions of games. Both Experiment 2 and 3 showed that individuals notice when their own payoffs are missing, and that they then request them. When their own payoffs are described only in part, they request information about the rest of their payoffs. Otherwise, they are prepared to play in ignorance of their partner's payoffs both for dependent and independent games. Experiment 4 showed that when participants can lose money depending on their choices, the majority of them were still prepared to play without asking for additional information. It is rational to play independent games in ignorance of your partner's payoffs or likely choice of option; but it is not rational to play dependent games in these circumstances.

Why don't individuals realize that they should ask for their partner's payoffs? The simplest game that elicits the need for strategic thinking is one in which two players each have two options. Such a game, however, calls for reasoners to hold in mind four separate possibilities in order to work out an optimal strategy. Four simple possibilities are known to be at the limit of typical adult competence in reasoning (Johnson-Laird and Byrne, 1991), and so it is natural for reasoners to focus on their own payoffs. This tendency, as we have shown, is influenced by experimental manipulations. We believe that if na ve individuals were given immediate payoffs after each game, and ran the risk of loss with sub-optimal choices, then they would soon realize the need to infer their opponents' likely choices. They should then be able to eliminate clearly dominated strategies in simple games. But, it is unlikely that they would be able to compute an equilibrium for more complex games — the number of possibilities would overwhelm the processing capacity of working memory. As our results imply, na ve individuals do not normally envisage the payoffs of their opponents, and that is why they do not ask for this information. As the adage says, it is hard to put oneself into others' shoes. This difficulty can, in turn, lead to an erroneous choice of option in those cases in which the optimal choice depends on the choices of others.

Acknowledgements

This research was supported in part by a grant from the National Science Foundation (Grant BCS 0076287) to the second author to study strategies in reasoning. We thank the following colleagues for their advice: Victoria Bell, Zachary Estes, Mary Newsome, Vladimir Sloutsky, Jean-Baptiste van der Henst, and Yingrui Yang.

References

Bell,V., and Johnson-Laird, P.N. (1998). A model theory of modal reasoning. Cognitive Science, 22, 25-51.

Dixit, A.K., and Nalebuff, B.J. (1991) Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life. New York: Norton.

Goldvarg, E., and Johnson-Laird, P.N. (2001). Na ve causality: a mental model theory of causal meaning and reasoning. Cognitive Science, 25, 565-610.

Goldvarg, Y., and Johnson-Laird, P.N. (2000). Illusions in modal reasoning. Memory & Cognition, 28(2), 282-294.

Johnson-Laird, P.N., and Byrne, R.M.J. (1991). Deduction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M.S., and Caverni, J-P. (1999). Na ve probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.

Tversky, A., and Kahneman. D. (1981). The framing of decisions and psychology of choice. Science, 211, 453-458

Implicit Learning of Serial Reaction Time Tasks: Connectionist vs. Symbolic Models

Ron Sun (rsun@cecs.missouri.edu)
Department of CECS, University of Missouri, Columbia, MO

Chris Terry (cterry@cs.ua.edu)
Department of Computer Science, University of Alabama, Tuscaloosa, AL

Abstract

This paper describes simulations of implicit learning experiments. It compares simulations using connectionist models with existing simulations using symbolic models. It addresses an interesting issue raised by proponents of symbolic models, namely, the claim that implicit learning is better modeled by symbolic rule learning programs. This paper revisits such an issue by quantitatively comparing connectionist simulations with symbolic ones, in the context of the serial reaction time task of Lewicki et al (1987). This comparison is interesting because it helps to clarify, to some extent, some long standing confusions compounded by many claims and counter-claims. It also points to the idea of hybrid connectionist and symbolic models.

Introduction

There have been a variety of simulations of implicit learning experiments. The majority of them are connectionist, while some are symbolic. Proponents of symbolic models, however, raised some interesting issues. They claim that implicit learning is "better modeled by symbolic rule learning programs" (Ling and Marinov 1994), and symbolic models are better for "not only conscious processing but also unconscious processing", based on some limited success of modeling Lewicki's experiments (Lewicki et al 1987) using C4.5, a decision tree learning algorithm developed by Ross Quinlan. In this paper, we will revisit such claims by quantitatively comparing connectionist simulations with symbolic ones, especially in the context of the serial reaction time (SRT) task of Lewicki et al (1987). This comparison is interesting because it helps to clarify, to some extent, some long standing confusions compounded by many similar claims and counter-claims.

Some background is in order here. With regard to the serial reaction time task specifically, Cleeremans and McClelland (1991) simulated an SRT task involving nondeterministic grammars. They employed a recurrent backpropagation network that saw one position at a time but developed an internal context representation over time that helped to predict next positions. The model succeeded in matching human data in terms of degrees of dependency on preceding segments in a sequence (i.e., conditional probabilities). However, their success was obtained through introducing additional mechanisms for several types of priming (e.g., short-term weight changes and accumulating activations). They did not deal with capturing directly the reaction time data of their subjects.

Ling and Marinov (1994) simulated the SRT data from Lewicki et al (1987), using a symbolic decision tree learning algorithm (i.e., C4.5). Their model produced data on quadrant prediction accuracy and, based on the data, they succeeded in matching the human reaction time data, using a transformation that included a power function (for capturing unspecific learning). However, they did not attempt the match without such a power function.

Similarly, Lebiere et al (1998) simulated data on SRT using ACT-R. The simulation was based on a combination of instance-based learning implemented in ACT-R and a set of hand-coded, symbolic, a priori rules. A fit with data was found.

It has been claimed, on the connectionist side, that a vast majority of human cognitive activities (i.e., implicit processes), including "perception, motor behavior, fluent linguistic behavior, intuition in problem solving and game playing — in short, practically all skilled performance", can only be modeled by subsymbolic computation (connectionist models), and symbolic models can give only an imprecise and approximate explanation to these processes (Smolensky 1988). It has also been claimed, on the symbolicist side, that "one and the same algorithm can be responsible for conscious and nonconscious processes alike", or even that implicit learning "should be better modeled by symbolic rule learning programs" (Ling and Marinov 1994). See also Fodor and Pylyshyn (1988).

This argument is in a way similar to what has been happening in relation to modeling past-tense acquisition in children (including how to capture the U-shaped curves in the process). For arguments and counter-arguments concerning advantages or disadvantages of connectionist and symbolic models in relation to past-tense acquisition, see, for example, Christiansen et al (1999). In this paper, let us look into the simulation of implicit learning specifically.

Simulating Lewicki et al (1987)

The Model. CLARION is a general cognitive architecture capable of simulating a variety of cognitive data (see Sun 1999, Sun et al 2001). The model consists of two levels: an implicit learning level (the bottom level) that learns using trial-and-error processes through a combination of backpropagation and reinforcement learning (i.e., Q-learning) algorithms (Watkins 1989, Sun and Peterson 1998); an explicit learning level (the top level) that learns explicit rules through on-line hypothesis testing based on information from the implicit level (the bottom level), which was termed "bottom-up learning" in Sun et al (2001). Bottom-up learning proceeds by first constructing the most specific rule that corresponds to a "good" decision made by the bottom level, and then refining (generalizing) it through examining the result of applying the rule, mainly through the use of an "information gain" measure that compares the success ratios of various modifications of the current rule.

Note that for this type of task, there is no significant amount of explicit learning in human subjects, as shown by Lewicki et al (1987). (This point can be controversial; more discussions later.) Correspondingly, in the model, the top level is not relevant (practically speaking). A parameter in the model is set in accordance with domain characteristics, which prevents explicit learning from occurring. The parameter concerns the minimum frequency of repetitions of a pattern in order for the afore-mentioned explicit learning to occur (see Sun et al 2001 for details of explicit learning).

In the bottom level, a simplified learning process was employed, again in accordance with domain characteristics, in which the backpropagation algorithm was used but temporal credit assignment (Q-learning) was not. This was because in this task, subjects predicted one position at a time, with immediate feedback, and thus there was no role for temporal credit assignment (Q-learning) to play.

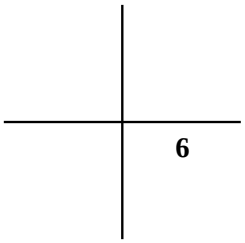
Specifically, in the model, $Q(x, a)$ computes the likelihood of the next position a, given the information concerning the current and past positions x. The actual probability of choosing a as the current prediction (of the next position) is determined based on the Boltzmann distribution, as is common for Q-learning:

$$p(a|x) = \frac{e^{Q(x,a)/\alpha}}{\sum_i e^{Q(x,a_i)/\alpha}}$$

where α controls the degree of randomness (temperature) of the decision-making process. This method is also known as Luce's choice axiom (Watkins 1989).

The error signal used in the learning algorithm is as follows:

$$\Delta Q(x, a) = \alpha(r + \gamma \max_b(y, b) - Q(x, a) = \alpha(r - Q(x, a))$$



5	1	3	8	4	7
4	9	7	1	**6**	0
7	8	2	4	5	2
9	0	3	4	5	1
3	2	5	8	0	9
1	9	8	9	2	7

(1) identification trials (2) matrix scanning trials

Figure 1: Identification trials and matrix scanning trials.

where x is the input, a is one of the outputs (the predictions), $r = 1$ if a is the correct prediction, $r = 0$ if a is not the correct prediction, and $\gamma \max_b(y, b)$ (which represents Q-learning of Watkins 1989) is set to zero because of the fact that only one-step prediction is involved in this task. This process amounts to a variant of the backpropagation algorithm. Based on the error measure, the backpropagation algorithm is applied to adjust internal weights (which are randomly initialized before training) as usual.

The Task. The task was based on matrix scanning: Subjects were to scan a matrix, determine the quadrant of a target digit (the digit 6) and respond by pressing the key corresponding to that quadrant. Each block consisted of six identification trials followed by one matrix scanning trial. In identification trials, the target appeared in one of the quadrants and the subject was to press the corresponding keys. In matrix scanning trials, the target was embedded among 36 digits in a matrix, but the subject's task was the same. See Figure 1. In each block of 7 trials, the actual location of the target in the 7th (matrix scanning) trial was determined by the sequence of the 6 preceding identification trials (out of which 4 were relevant). 24 rules were used to determine the location of the target on the 7th trial. Each of these rules mapped the target quadrants in the 6 identification trials to the target location on the 7th trials in each block. 24 (out of a total of 36) locations were possible for the target to appear. The major dependent variable was the reaction time on the 7th trial in each block.

The whole experiment consisted of 48 segments, each of which consisted of 96 blocks of 7 trials (so there were a total of 4,608 blocks). During the first 42 segments, the afore-mentioned rules were used to determined target locations. However, on the 43rd segment, a switch occurred that reversed the outcomes of the rule set: the upper left was replaced by the lower right, the lower left was replaced by the upper right, and so on. The purpose was to separate unspecific learning (e.g., motor learning) from prediction learning (i.e., learning to predict the target location on the 7th trial).

The Data. The reaction time data of three sub-

jects were obtained by Lewicki et al (1987). See Figure 2. Each curve showed a steady decrease of reaction times up until the switch point. At that point, there was a significant increase of reaction times. After that, the curve gradually lowered again.

The Model Setup. The input to the model contained (a sequence of) 6 elements, with each element having 4 possible values (for 4 different quadrants). [1] The output contained the prediction of the 7th element in a sequence. Thus, 24 input units (representing 6 elements, with 4 values each), 24 output units (one for each possible location of the 7th element of a sequence), and 18 hidden units were used. We tried various parameter settings. The best learning rate was 0.5, with a momentum term of 0.2. The model was trained by presenting the stimulus materials in the same way as the human experiment described by Lewicki et al (1987), without any further embellishments or repetitions of the materials.

Because in this experiment there were a total of 6^4 sequences with each consisting of 7 elements, the setting was too complex for subjects to discern the sequence structures explicitly, as demonstrated in human experiments by various explicit tests done by Lewicki et al (1987). Computationally, no explicit representation of knowledge could be extracted in the model, because the large number of sequences entailed that there were no sufficient repetitions of any particular sequence throughout the experiment, which prevented the model from coming up with any rule. The density parameter was set to be 1/50; that is, at least one repetition (of a sequence) was necessary every 50 blocks in order to maintain an explicit rule (Sun and Peterson 1998). In this task, there were 4,608 presentations of sequences and there were $6^4 = 1296$ different sequences, and thus on average the repetition rate of any sequence was only 0.0007716.

A note concerning the existence of explicit knowledge in implicit learning tasks in general is in order here. It has been hotly debated whether there is a significant amount of explicit knowledge involved in implicit learning tasks and whether explicit knowledge can account for the performance in such tasks (see Sun et al 2001, 2002 for reviews). Without getting into details of such debates, we can reasonably believe that, although explicit knowledge may be present in many implicit learning tasks, the existence of a significant amount of such knowledge is highly unlikely in the task of Lewicki et al (1987), given the complexity of their task setting (Sun et al 2002). This was the assumption made in our simulation, although it would not change our main points even if this assumption was dropped.

The Match. We were able to create an error rate curve going downwards (averaged over 10 runs to ensure its representativeness), resembling Lewicki et al's reaction time curves. The model reached 100% accuracy at a point before the switch. See Figure 3. The question is how we should translate the error rates into reaction times.

[1] A sequence of 6 elements was assumed to be within the capacity of the short-term working memory.

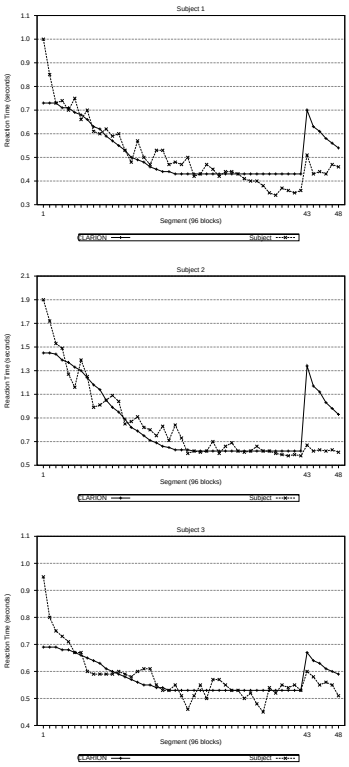


Figure 2: Matching Lewicki's data using linear transformation. See Figure 4 for parameter values.

One way of translation is through a linear transformation from error rates to reaction times (as commonly used in existing simulation work); that is, $RT_i = a * e_i + b$, where RT_i is the reaction time (in ms), e_i is the error rate, and a and b are two free parameters. For each set of human data, we adjust the parameters to match the data. One possible interpretation of linear transformation is that it specifies the time needed by a subject to search for a target item as well as the time needed by a subject to respond to a target item without searching (through correctly predicting its location); that is,

$$RT_i = ae_i + b = b(1 - e_i) + (a + b)e_i$$

where b is interpreted as the time needed to respond to an item without searching (since $1-e_i$ is the probability of successfully predicting the location of a target item), and $a+b$ is interpreted as the time needed to respond to an item by first searching for it and then responding to it (after finding it). So, instead of relying on additional functions (as in e.g. Ling and Marinov 1994), this method relies only on error

rates to account for human performance in terms of reaction times.

Another way of generating reaction time from prediction accuracy is through a formula used by Ling and Marinov (1994):

$$RT_i = t_1(1 - e_i) + t_2 e_i + B\alpha^{-t}$$

where t_1 is the time needed to respond when there is no search (using correct predictions), t_2 is the time needed to respond when search is necessary, B is the initial motor response time, and α is the rate at which the motor response time decreases. The third term is meant to capture unspecific practice effects (mostly resulting from motor response learning). In other words, in this formula, we separate the motor response time from the search time and the prediction time (as represented by t_1 and t_2 respectively). Note that, if we set $B = 0$, we have $t_1 = b$ and $t_2 = a + b$ and thus this equation becomes the same as the previous one. This formula takes into account the independent nature of motor learning, as separate from the learning of prediction of target locations. However, it involves two more free parameters.

Using the linear transformation (without the power function), we generated three sets of data from the error rate curve reported earlier, [2] one for matching each human subject in Lewicki et al (1987), using different a and b values for each subject. [3] As shown in Figure 2, the model outcome fit the human data well up to the point of switching (segment 42). When the switch to a random sequence happened, the model's reaction times became much worsened whereas the subject's reaction times suffered only slightly (although in a statistically significant way). [4]

[2]When curve fitting, we used Microsoft Excel Solver to find the best parameter values (e.g., a and b in $a * x + b$) such that the difference between the model data and the subjects data was minimized. Microsoft Excel Solver uses the Generalized Reduced Gradient nonlinear optimization algorithm.

[3]The error rate curve reported earlier was the best curve and happened to match all three subjects approximately equally well after the transformation (with different parameters for each subject). Note that Ling and Marinov (1994) also used a single error rate curve to match different subjects with different parameters for transformation.

[4]We tried many different parameters but discovered that the size of the jump tended to vary little (unless the match as a whole was bad). It is clear, from our experiments with different settings of the parameters, that, if the model learns the sequences perfectly before the switch (as is the case with our model), the model data inevitably have huge jumps. However, the more of the sequences it does not learn, the flatter the curve and the less the jump. Although this may model Subjects 1 and 3 satisfactorily, Subject 2 has a large drop in reaction time early on which is best matched by having the model increase its accuracy in a rapid manner.

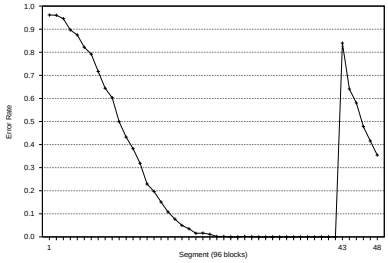


Figure 3: The model prediction errors.

Subject 1:	$a = 320$	$b = 430$
Subject 2:	$a = 860$	$b = 620$
Subject 3:	$a = 170$	$b = 530$

Figure 4: Parameters for matching Lewicki's data.

Next, we added the power function, in order to compare our simulation with Ling and Marinov (1994). After adding the power function, we re-fit the parameters. The effect of adding the power function was that we reduced the contribution from the model prediction (i.e., the error rate e_i) while we took into consideration the contribution from the power function. In this way, we obtained a much better match after the switch and a good match before the switch at the same time. [5] See Figure 6. Note that in the figure, we used exactly the same parameter settings (for a, b, B, and α) as in Ling and Marinov (1994). These parameters might be further optimized, which led to a slightly better fit but the difference was not significant.

The match between our model and the human data was excellent as measured by the sum-squared errors. Compared with Ling and Marinov (1994), CLARION (with the power function) did better on two of the three subjects, using the same parameters for transformation as Ling and Marinov did. See Figure 7 for a comparison.

So, what conclusion can we draw concerning the relative merits of the two models? The next section attempts to answer this question in a more theoretically oriented way.

Connectionist vs. Symbolic Models

Revisiting the argument of whether connectionist or symbolic models are better models (see Introduction), what do the above simulations have to say about it? To put it simply, we believe that this issue is a red herring. Being able to simulate some limited

[5]By adding the power function, we were able to reduce the total difference between the jumps in our curves and the corresponding jumps in the subject data by half. This comparison suggested that the amount of benefit the human subjects got from their predictions (i.e., by lowering e_i) was, although significant, relatively small. Significant benefit was gained through the improvement of motor responses as represented by the power function.

Subj.1:	$t_1 = 150$	$t_2 = 350$	$B = 700$	$\alpha = 0.33$
Subj.2:	$t_1 = 150$	$t_2 = 350$	$B = 1600$	$\alpha = 0.33$
Subj.3:	$t_1 = 100$	$t_2 = 210$	$B = 800$	$\alpha = 0.19$

Figure 5: Parameters for matching Lewicki's data with power functions added (as in Ling and Marinov 1994).

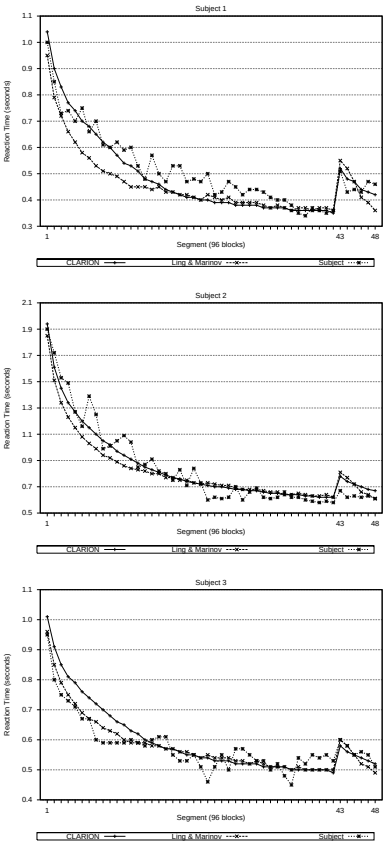


Figure 6: Matching Lewicki's data with power functions added. See Figure 5 for parameter values.

data of implicit learning amounts to very little, in that any Turing equivalent computational process, that is, any generic computational model, should be able to simulate these data. Thus, the simulation of data by itself does not prove whether a particular model is a suitable one or not (cf. Roberts and Pashler 2000). Other considerations need to be brought in to justify a model.

We would suggest that one such issue is accessibility (Sun 1995, 1999, Cleeremans et al 1998). While symbolic models of implicit learning lead to explicit symbolic representation of implicit knowledge (e.g., Ling and Marinov 1994, Lebiere et al 1998, Anderson 1993) that is evidently accessible (without using any add-on auxiliary assumptions), connectionist models of implicit learning lead to implicit (subsymbolic, distributed) representation of knowledge that is inherently inaccessible (such as in the bottom level of CLARION).

model	subj.1	subj.2	subj.3
CLARION w/o power f.	0.30	1.85	0.14
CLARION w/ power f.	0.07	0.25	0.05
Ling & Marinov (1994)	0.14	0.43	0.04

Figure 7: Comparing the goodness of fit in terms of SSEs.

Note that it is generally not the case that distributed representations (as in the bottom level of CLARION) are absolutely inaccessible, but they are not as immediate as localist representations. Distributed representations may be accessed through indirect, transformational processes. As Kirsh (1990) put it, "explicitness [of representation] really concerns how quickly information can be accessed..... It has more to do with what is present in a process sense, than with what is present in a structural sense". The accessibility difference between the two levels should be understood in this way.

Thus, connectionist models have a clear advantage: Being able to match human implicit learning data (at least) as well as symbolic models, they also account for the inaccessibility of implicit knowledge better and more naturally than symbolic models (Cleeremans et al 1998, Sun 1999). In this sense, they are better models.

On the other hand, it is generally agreed upon that symbolic/localist models have their roles to play too. They are better at capturing explicit processes, including their accessibility (Smolensky 1988, Sun 1995, 1999).

This contrast lends support to the belief that, since connectionist models are good for implicit processes and symbolic models for explicit processes, the combination of the two types should be adopted in modeling cognition (Sun 1995, 1999). There have been many philosophical and psychological theories related to this point (Sun et al 2001): See, e.g., James (1890), Schacter (1987), Reber (1989), Stanley et al (1989), Clark and Karmiloff-Smith (1993), and Sun (1999). This combination is exemplified by the CLARION model (Sun et al 2001).

This combination may also shed some light on the issue of consciousness, because the implicit/explicit difference involves, in its core, the issue of awareness, which is the key to consciousness (Cleeremans et al 1998). The representational distinction provides a plausible grounding for the notion of awareness (see Sun 1999 for details of theoretical arguments).

Simulating Other Tasks

Beside modeling the data from Lewicki et al (1987), CLARION has also been applied to model a variety of other SRT experiments, including Curran and Keele (1993) and Willingham et al (1989). Notably, in

these tasks, due to sufficient repetitions of sequential patterns, explicit learning of these patterns was involved, although implicit learning was dominant. Therefore, the top level of CLARION was utilized. Together, the model demonstrates the interaction between implicit and explicit learning (Sun et al 2002).

Beside SRT simulations, CLARION can capture data from many other implicit learning tasks. These tasks include artificial grammar learning (Reber 1989) and process control (Stanley et al 1989) (see Sun et al 2002). In addition, CLARION has also simulated extremely complex skill learning tasks as well, such as the minefield navigation task (see Sun et al 2001). The generality of CLARION has been amply demonstrated, on top of its cognitive validity.

Acknowledgments

This work was supported in part by Office of Naval Research grant N00014-95-1-0440 and by Army Research Institute grant DASW01-00-K-0012.

References

J. R. Anderson, (1993). *Rules of the Mind.* Lawrence Erlbaum Associates. Hillsdale, NJ.

M. Christiansen, N. Chater, and M. Seidenberg, (eds.) (1999). Connectionist models of human language processing: progress and prospects, special issue of *Cognitive Science*, 23(4), 415-634.

A. Clark and A. Karmiloff-Smith, (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language.* 8 (4), 487-519.

A. Cleeremans, A. Destrebecqz and M. Boyer, (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, 2, 10, 406-416.

A. Cleeremans and J. McClelland, (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General.* 120. 235-253.

T. Curran and S. Keele, (1993). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 19, 189-202.

J. Fodor and Z. Pylyshyn, (1988). Connectionism and cognitive architecture: A critical analysis. In: Pinker and Mehler (eds.) *Connections and Symbols*, MIT Press, Cambridge, MA. 1988

W. James, (1890). *The Principles of Psychology.* Dover, NY.

D. Kirsh, (1990). When is information explicitly represented. In: P. Hanson, (ed.) *Information, Language, and Cognition.* University of British Columbia Press, Vancouver, Canada.

C. Lebiere, D. Wallach, and N. Taatgen, (1998). Implicit and explicit learning in ACT-R. *Proc. of ECCM'98*, pp.183-189. Nottingham University Press.

P. Lewicki, M. Czyzewska, and H. Hoffman, (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition.* 13 (4), 523-530.

C.X. Ling and M. Marinov, (1994). A symbolic model of the nonconscious acquisition of information. *Cognitive Science*, 18(4), 595-621.

A. Reber, (1989). Implicit learning and tacit knowledge. *Journal of Exp Psychology: General.* 118 (3), 219-235.

S. Roberts and H. Pashler, (2000). How persuasive is a good fit? *Psychological Review*, 107 (2), 358-367.

D. Schacter, (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 501-518.

P. Smolensky, (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–74.

W. Stanley, et al, (1989). Insight without awareness *Quarterly Journal of Experimental Psychology.* 41A (3), 553-577.

R. Sun, (1995). Robust reasoning. *Artificial Intelligence.* 75, 2. 241-296.

R. Sun, (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, Vol.8, 529-565.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skill to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, Vol.25, No.2, pp.203-244.

R. Sun, P. Slusarz, and C. Terry , (2002). The interaction between the implicit and explicit processes: a dual process approach. Submitted.

C. Watkins, (1989). *Learning with Delayed Rewards.* Ph.D Thesis, Cambridge University, UK.

D. Willingham, M. Nissen, and P. Bullemer, (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 15, 1047-1060.

Detecting the Local Maximum: A Satisficing Heuristic

Yanlong Sun (suny@bgnet.bgsu.edu)
Ryan D. Tweney (tweney@bgnet.bgsu.edu)

Department of Psychology, Bowling Green State University
Bowling Green, OH 43403 USA

Abstract

In a simulated yard sale task, participants were asked to sell a series of objects, each of which would attract three customers making a randomly determined offer. Participants were told to maximize the total "take" from the sale. The analysis of the data revealed that high-performing naive participants were using a strategy that made them relate the current event to the seemingly irrelevant preceding events. We argue that this strategy is consistent with Simon's (1982) notion of "satisficing heuristic", which accounted for both participants' limited computation capacity and the task environment.

Introduction

Intuitive predictions and probabilistic judgments are often used as tasks to evaluate people's performance in judgment and decision-making research, and a common scheme is to collect incorrect predictions and misjudgments by "setting up a 'trap' that subjects would fall into if they were using a particular heuristic" (W. Goldstein & Hogarth, 1996, p.26). In this type of research, predictions derived from probability theory are often used as an objective criterion, and violations of the normative models are labeled as biased or irrational. Tversky and Kahneman's "heuristics and biases" program has been the most influential in this field. They suggested that intuitive predictions and judgments are often mediated by a small number of distinctive mental operations, which they called *judgmental heuristics*. "These heuristics … are often useful but they sometimes lead to characteristic errors or biases" (Kahneman & Tversky, 1996, p.582). For example, people's tendency to use a small sample of preceding events to evaluate an overall process was attributed to a "representativeness" bias (Tversky & Kahneman, 1973). This bias has been used to account for many cognitive behaviors, such as the tendency to see streaks in random sequences (Gilovich, Vallone & Tversky, 1985), and failure to "acquire a proper notion of regression" (Tversky & Kahneman, 1973). In a recent study on gambling behaviors, Thaler & Johnson (1990) concluded that "current choices are often evaluated with the knowledge of the outcomes which have preceded them, (but) such knowledge can often be a handicap" (p.643).

However, the heuristics and biases research program has recently been controversial, partly because "biases" sometimes appear highly adaptive. Thus, Tweney & Doherty (1983) argued that confirmatory tendencies ("confirmation biases") can be adaptive when hypotheses are relatively new and untested. Further, in an extensive series of studies, Gigerenzer and his colleagues (e.g. 1991, 1994, Gigerenzer & Todd, 1999) found evidence which led them to strongly disagree with Kahneman and Tversky. They argued that many seemingly naïve "fast and frugal heuristics" are adaptive in an uncertain environment. Similarly, Kareev, et al., suggested that the limited capacity of working memory (hence the use of small samples) could actually help the early detection of covariation since small samples of correlated variables are highly skewed (Kareev, 1995; Kareev, Lieberman & Lev, 1997).

The present study followed Simon's (1982) notion of "bounded rationality", which takes into account both people's limited computation capacity, and the structure of task environments. Our findings suggest that under circumstances when the precise prediction derived from statistics or probability theory is not the only criterion, heuristics based on a small sample size can be valuable. With a ***satisficing*** strategy that only needs to "look for a satisfactory alternative" (Simon, 1982, p.295), naïve participants were able to effectively accomplish the goal of the task, based on the evaluation of a few preceding events.

Recognizing the Maximum of a Sequence

The statistical properties of sequential lists of evidence have long been of interest to mathematicians. The dowry problem (or the secretary problem) is a classic example in the dynamic programming literature, one analyzed by Cayley in 1875 (see Ferguson, 1989). As a mathematical problem, the dowry problem is difficult to solve, requiring advanced mathematical knowledge and problem solving ability. Obviously, few, if any, people are likely to work out the exact stopping point mathematically in an everyday life situation when a similar problem is encountered. Instead, without complicated calculations, a player might need to use "common sense" to make decisions. The present study adopted a simplified version of the problem – a simulated "yard-sale" task – to test how naïve people evaluate preceding events and make decisions when facing sequential events generated by an unknown process.

Participants were asked to sell a series of objects in a simulated yard sale. Each object attracted three potential buyers, each of whom came at a different time and made a

different offer. It was explained that offers that were rejected would not return, so that the task was to guess which was the best offer, and to take it when available.

Imagine that a person is selling a used car, and that visitors with different offers come up in a random order. After 5 offers have been declined in a week, a visitor comes in with a price higher than any of the previous ones. Another 5 offers will probably take another week and by then this car must be sold. Whether to stop waiting and grab the currently available offer then depends on how satisfied the car owner feels about the current offer. The only information available to evaluate the current situation is the previous encounters. Probably, "common sense" would tell this car owner to take the offer now, because future offers might not get better.

This is in effect a ***satisficing heuristic*** (Simon, 1990), which is a strategy that only needs to "look for a satisfactory alternative" (Simon, 1982, p.295), as suggested by the notion of bounded rationality. The strategy also fits the category of fast and frugal heuristics suggested by ecological rationality, because it makes "a choice from a set of alternatives encountered sequentially when one does not know much about the possibilities ahead of time" (Gigerenzer & Todd, 1999, p.13).

We show that in at least one situation – when the random process that generates offers is independently and identically distributed – this satisficing strategy is ***optimal***. Let R_i denote the offer at time i, where $i = 0$ is the current offer, -1 is the previous one, +1 is the next one, and so on. Assume the car owner has encountered m R's (from R_{-1} to R-m) and found that R_0 is the best one so far. If he actually chooses it, because R_0 now is the biggest number in a local sequence of $(m + 1)$ numbers, in the long run, the value of such R_0 has a good chance to be higher than the population mean. For a continuous distribution from 0 to 1, the expected value of such R_0 is $(m+1)/(m+2)$. Further, R_0 might just be a good stopping point because the potential gain from the following n offers after R_0 might not have a good chance to get better. To see this, let A denote the event that R_0 is higher than its previous m offers, and B denote the event that R_0 is higher than its following n offers. Then two prior probabilities can be described as

$$p(A) = p\,(R_0 > R_{-1}, \ldots, R_{-m}) = 1/(m+1)$$
$$p(B) = p\,(R_0 > R_1, \ldots, R_n) = 1/(n+1)$$

And the conditional probability can be calculated as

$$p(B|A) = p(AB)/p(A) = (m+1)/(m+n+1)$$

Note that, with a fixed n, $p(B|A)$ approaches to an asymptote of 1 as m increases. That is, with an appropriate m (after considering a certain number of offers), the car owner can make a better decision than a random guess. For example, when $m = 5$, $n = 5$, $p(B|A)$ is 6/11, and this favors selling. To take the message of $p(B|A)$ in another way, it has suggested a ***stopping point***, because the coming n offers do not have a good chance to get better.

Two Optimal Strategies for the Yard Sale Task

With the development above, we can easily determine the optimal strategy for the yard sale task. Suppose there is only one trial in the task (only one object for sale). Let P_1 denote the first offer, P_2 the second and P_3 the third. Before knowing any of the three offers, the prior probability for each offer to be the best is equal:

$$p(P_1 \text{ is best}) = p(P_2 \text{ is best}) = p(P_3 \text{ is best}) = 1/3$$

Note that knowing the exact value of P_1 does not change this probability. With a random guess, the chance of hitting any of the three possible prices is 1/3. However, if we skip P_1 and consider P_2, the conditional probability is no longer equal. If P_2 is higher than P_1, we should take it immediately because $p(P_2 > P_3 \mid P_2 > P_1) = 2/3$. Otherwise, we should take P_3. A pay-off matrix (Table 1) shows that the optimal strategy (Option B*) is to always skip P_1. If P_2 is better than P_1, accept P_2; if P_2 is worse, choose P_3. This strategy increases the chance of hitting the best offer to 1/2, with a 1/3 chance of hitting the middle price, and a 1/6 chance of hitting the lowest one. For convenience, we will refer to this strategy as the "**one deal strategy**".

Table 1: The pay-off matrix for the seller

Option	Rank orders of offers						Total
	LMH	*MHL*	*LHM*	*MLH*	*HLM*	*HML*	
A	-1	0	-1	0	1	1	0
B*	0	1	1	1	0	-1	2
C	1	-1	0	-1	-1	0	-2

Note: *L* is the lowest price, *M* the middle, *H* the highest. "*LMH*" means that the lowest price comes first, and so on.
Option A: always choose P_1 (random guess).
Option B*: choose P_2 if $P_2 > P_1$, otherwise choose P_3.
Option C: choose P_2 if $P_2 < P_1$, otherwise choose P_3
Gains: the seller gains –1 when hitting the lowest offer; 0 for the middle offer; 1 for the highest offer.

However, in a real-life situation, decisions are rarely made in temporal isolation. Thus, as in a common scheme in laboratory experimental settings, our yard sale task used repeated trials to collect multiple data points from each individual participant. This fact had a significant impact on the optimal strategy. Recall that the single deal strategy assumes that in each deal, the order in which three offers appear is completely independent from any other events, and requires that the first offer always be ignored. What if the first offer actually is the best one? With the information from the preceding trials, we can actually evaluate how good the first offer is. Calculating an optimal strategy for deals in a sequence is very complicated because it needs to specify a distribution of three offers for each deal. However, when distributions of offers in several deals within a local sequence are similar, as an approximation, the principles we presented above can be generalized. In our experiment, we set the basic price for each object to range from $50 to $100, with a maximum random fluctuation of ± $16. Figure 1 shows the overall distribution of these offers.

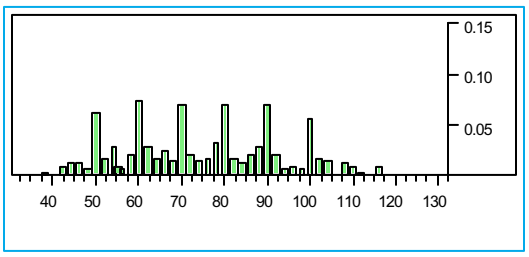


Figure 1 The distribution of offers. N = 5760, x-axis is price, y-axis is proportion.

The satisficing principle suggests that the offers for previous objects (in previous trials) can be used to predict whether you are getting a good offer for the next item. In other words, when you are considering a first offer for a table, if you can recall that the last several visitors, who were seeking other items, were not as generous as the current customer, you may want to sell the table right now. It seems quite against a researcher's intuition that a normative strategy would predict that previous offers for an umbrella will help to predict the current offers for a table, especially when one thinks that the umbrella deal is "over", and the two deals should be independent. The answer to this counter-intuitive puzzle is that the independence is only partial. While the order in which different offers come out for each deal is independent, the values of these offers, if they are in the same or similar distributions, regress to the population mean.

There are two ways to evaluate the first offer for a given trial. The '**local count strategy**" is based upon a count of the number of previous low offers. That is, if the current first offer is higher than a certain number of previous offers (for other items), take it. The optimal strategy depends on the specific distribution of the random offers and the payoff matrix. In our specific experimental setting, we used computer simulations and found that the best number of comparisons was 6. A modification of the local count strategy, the "**moving average strategy**" compares the current offer with the average of previous several offers. We reasoned that participants might not remember the exact values of the previous offer but might still have a vague memory of the overall average in a short local sequence. A logistic regression over simulated data showed that in our experiment setting, the value difference of the first offer for a given item from the mean of the previous 6 offers (for the other two items), is significant as a predictor of whether the first offer is the best among all three offers: $\chi 2$ (1, N = 1888) = 191.03, p<.01. That is, as this difference increases, this first offer is more likely to be the best of the three.

With this background, we were ready to find out whether participants are good at detecting good offers when they actually appear, and whether they use information from previous encounters to help their current decision-making.

Method

Participants were 15 undergraduate students from an introduction to psychology class at Bowling Green State University, none of whom had taken a course on game theory or probability theory. We refer to them as novice participants. One graduate student with extensive experience in judgment and decision-making and related research also participated, and will be referred as the expert participant.

The task was conducted using a self-paced computer program. Each participant completed 120 trials (the number of objects to be sold). One object was to be sold in each trial. Participants could take any of the three offers at the time it was available, but could not go back to an earlier declined offer. Once an offer was taken, offers thereafter were not presented. The third offer was forced if the first two were rejected by the participant, and this was the only case when participants knew exactly if they had hit the best out of three offers. After each trial, participants were given a confirmation that the object was sold at the price they selected. Participants' choices and their total earnings were recorded. An average experiment session lasted about 25 minutes.

Results

Overall Performance To evaluate participants' overall performance, we ran a simulation 5000 times using each of the three strategies: a random guess (randomly choosing one of the three offers), the "single deal strategy" and the "local count strategy". Each time the simulation sold 120 items using the actual selling list that was used in the experiment. In the local count strategy, the first offer for each item was compared with 6 previous offers (which were for the preceding items [1]). It was accepted when it was the highest in the comparison. Otherwise, it was declined and the single deal strategy was applied. Table 2 shows the simulation results and the actual participant data.

Table 2 Comparisons between human participants and 3 simulations

Group	N	Mean Score (95% confidence interval)	Std Dev.
Random Guess	5000	8889.8 ± 2.0	72.74
Single Deal Strategy	5000	9160.7 ± 2.1	75.54
Local count Strategy	5000	9277.7 ± 3.4	121.22
Human Participants	16	9196.0 ± 31.0	58.15

Note that all 16 participants received a score that was at least 1.5 SD above the mean of the random guess simulation.

[1] When an offer was taken before all offers were presented, the number of items whose offers were being compared may exceed 2.

Each participant's score was then compared to the result as if the single deal strategy had been applied to his/her actual selling list. Ten participants' scores were higher than the result of the single deal strategy. Using the standard deviation resulting from the single deal strategy simulation (75.54), four participants' scores were at least 1.5 SD above the score resulting from the single deal strategy. We will refer to these four as the "outstanding participants".

Strategy Use We looked at participants' choice patterns in regard to their consistency with the optimal strategies, at three steps when each offer was being considered. The following three choices are consistent with the optimal strategies (single deal or multiple deals):

C1. Accept the 1st offer if it is better than several previous offers (for other items).

C2. Decline the 1st offer, and accept the 2nd offer if it is better than the 1st one.

C3. Accept the 3rd offer if the 2nd is worse than the 1st.

C2 and C3 are equivalent to the single deal strategy, now separated into two parts. All three choices above are consistent with the local sequence strategy. Since choices at the 3rd offer were forced, whether participants' actual choices were consistent with the optimal strategies could be looked at whether they had met or violated the conditions at C1 and C2. Note that the single deal strategy actually forbids C1. Specifically, C1 can result from considering the count of the previous low offers (the local count strategy) or the value difference of the first offer compared to the mean of the previous offers (the local average strategy), and we tested them separately.

Of all 16 participants, only the expert participant found the single deal strategy, and followed C2 and C3 consistently. The 15 novice participants, by contrast, often violated either C2 or C3 or both. However, to a significant extent, their choices did follow C1. For each individual novice participant, we ran a logistic regression, using the value difference of the first offer from the mean of the previous 6 offers, to predict the participant's acceptances of the first offers. Of the 15 participants, 11 showed significant results at a 0.01 level. On the group level, the result is also significant: $\chi 2$ (1, N=1770) = 304.69, $p<.01$. This indicates that the novices were at least partly using the moving average strategy.

Since the one deal strategy is a subset of the local count strategy, we combined the 16 participants' reactions on all three offers to see if their behaviors were consistent with the local count strategy. Table 3.1 and Table 3.2 show that they did show such consistency when the previous 1 or 6 offers were compared to the current offer. That is, if the offer being considered was better than all of the previous 1 or 6 offers, participants were more likely to accept it. Otherwise, they were more likely to decline it. This finding was consistent with the local count strategy.

Table 3.1 Compared to previous 1 offer

	Worse	Better	Total
Decline	1773	952	2725
Accept	616	1424	2040
Total	2389	2376	4765

$\chi^2 = 580.177$, $p<.01$

Table 3.2 Compared to previous 6 offers

	Worse	Better	Total
Decline	2397	328	2725
Accept	1553	487	2040
Total	3950	815	4765

$\chi^2 = 114.140$, $p<.01$

All of the 4 "outstanding participants" were novice participants. However, they actually outperformed the expert participant and the one deal strategy. They were different from the other 11 non-expert participants in that their behaviors were consistent with one of the requirements of the one deal strategy (C2 and C3), although not both. Their gains on the first offers when these offers were the best had offset the losses from violations of the condition of C2.

Learning across Trials In a study of the Monty Hall dilemma, Granberg & Dorr found that participants showed signs of learning across trials under certain conditions. In our study, we also looked at whether there were systematic changes in participants' choices across trials. Specifically, we suspected that participants might have learned the specific distribution of random offers in earlier trials, so that, in later trials, they only needed to recognize "globally big numbers" instead of applying their heuristics independently and locally. For example, an offer of $116 might have been the best one for an item sold in an early trial. If participants had this number memorized, they might just pick an offer of $116 or higher in a later trial, no matter when this offer was presented (whether it was the 1st, 2nd, or 3rd offer). If this were the case, "big wins" might have been over-represented in terms of participants' uses of simple heuristics.

However, in our experimental setting, each item's 3 offers varied around its own basic price. Although these basic prices could be close, there was no way to tell that $116 was the best offer for item A only because it had been the best offer for item B. In other words, recognizing "big numbers" alone would not help in optimizing the total performance. In fact, when we partitioned each participant's 120 trials into 3 blocks with 40 trials each, we did not find any significant differences across blocks, indicating that learning was probably not important across trials.

Discussion

None of the 15 novice participants found and consistently used the one deal strategy. We reasoned that this was because finding and consistently applying this strategy required participants to use background knowledge in probability theory, and they simply did not have this training. Their consistency with the local sequence strategy explained why they had good performances. This does not suggest that they actually did the calculation and found the correct mathematical solution, because this would require even more knowledge and computational capacity, not to mention that it was within a short experiment session. However, as we suggested before, it is not necessary for a person to work out the correct mathematical proof to use the local sequence strategy. Such a strategy could arise from participants' everyday life experiences, from which they had learned a simple satisficing heuristic: "grab any good chance when you can".

Surprisingly, the outstanding performers actually outperformed the expert participant who found and consistently applied the one deal strategy. This was because the one deal strategy has to give up all opportunities of accepting the first offers when they were the best. One reason that prevented the expert participant from finding the local sequence strategy might have been that the everyday life heuristic had been "blocked" by his knowledge of probabilistic judgment research. This finding is very similar to Goldstein and Gigerenzer's (1999) "less-is-more" effect, that relative ignorance can actually benefit a decision maker. By isolating the previous encounters from the current decision-making situation, the expert participant had to search the infinite probability space again, and previous experience, either beyond or within the experiment task, could not help.

In their 1973 paper, Kahneman and Tversky suggested that "people do not acquire a proper notion of regression, … they do not expect regression in many situations where it is bound to occur", because "regression effects typically violate the intuition that the predicted outcome should be maximally representative of the input information". On the contrary, the finding in this study that participants' behaviors were consistent with the local sequence strategy, indicated that people do have good intuitions about such regression, and can also take advantage of it.

We argue that to evaluate naïve people's probabilistic judgment and decision-making, one has to take into account both people's limited computation capacity and the task environment. One obvious message of the task was that, if we had used the single deal strategy as the ***only*** criterion, we might have concluded that participants were being irrational, and would then have to face the puzzling evidence that they actually performed very well. Instead, the results suggest that the advantages of the satisficing principle are important and cannot be ignored. By using these strategies, people can benefit from their own experiences, even from a small sample of preceding events.

References

Ferguson, T. S. (1989). Who solved the secretary problem? Statistical Science, 4(3), 282-296.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." In W. Stroebe & M. Hewstone (Eds.), European review of social psychology. Chichester, England: Wiley.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), Subjective probability. New York: Wiley.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Gigerenzer, G., Todd, P. M. & the ABC Research Group (Eds*.),* Simple heuristics that make us smart. New York: Oxford University Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17, 295-314.

Goldstein, D., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In Gigerenzer, G., Todd, P. M. & the ABC Research Group (Eds.), Simple heuristics that make us smart. New York: Oxford University Press.

Goldstein, W. M, & Hogarth, R. M. (1996). Judgment and decision research: Some historical context. In Goldstein, W. M & Hogarth, R. M. (Eds.), Research on judgment and decision making: Currents, connections, and controversies. New York: Cambridge University Press.

Granberg, D., & Dorr, N. (1998). Further exploration of two-stage decision making in the Monty Hall dilemma. American Journal of Psychology, 111(4), 561-579.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction, Psychological Review, 80, 582-591.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions, Psychological Review, 3, 582-591.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. Cognition, 56, 263-269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. Journal of Experimental Psychology: General, 126(3), 278-287.

Simon, H. A. (1982). Models of bounded rationality, Vol.3, Cambridge, MA: MIT Press.

Simon, H. A. (1990). Invariants of human behavior, Annual Review of Psychology, 41, 1-19.

Thaler, R. H., & Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. Management Science, 6, 643-660.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Tweney, R. D. & Doherty, M. E. (1983). Rationality and the psychology of inference. Synthese, 57*,* 139-162.

Top-Down versus Bottom-Up Learning in Skill Acquisition

Ron Sun (rsun@cecs.missouri.edu)
Xi Zhang (xzf73@mizzou.edu)
Department of CECS, University of Missouri, Columbia, MO 65211, USA

Abstract

This paper studies the interaction between implicit and explicit processes in skill learning, in terms of top-down learning (that is, learning that goes from explicit to implicit knowledge) vs. bottom-up learning (that is, learning that goes from implicit to explicit knowledge). Instead of studying each type of knowledge (implicit or explicit) in isolation, we highlight the interaction between the two types of processes, especially in terms of one type giving rise to another. The work presents an integrated model of skill learning that takes into account both implicit and explicit processes and both top-down and bottom-up learning. We examine and simulate human data in the Tower of Hanoi task. The paper shows how the quantitative data in this task may be captured using either top-down or bottom-up approaches, although top-down learning is a more apt explanation of the human data currently available. The results demonstrate the difference between the two different directions of learning (top-down vs. bottom-up), and also provide a new perspective on skill learning in the Tower of Hanoi task.

Introduction

This paper studies the interaction between the implicit and explicit processes in skill learning. It explores two directions of skill learning: top-down learning and bottom-up learning. Top-down learning goes from explicit knowledge to implicit knowledge, while bottom-up learning goes from implicit knowledge to explicit knowledge. Instead of studying each type of knowledge (implicit or explicit) in isolation, we want to highlight the interaction between the two types of processes, especially in terms of one type giving rise to another.

In this work, we want to test possibilities of bottom-up learning vs. top-down learning. We do so by using the task of Tower of Hanoi, which is arguably a typical benchmark problem in high-level cognitive skill acquisition and has been used in many previous studies of skill acquisition, cognitive modeling, and cognitive architectures (see, e.g., Proctor and Dutta 1995, Anderson 1993, Anderson and Lebiere 1998).

To explore bottom-up and top-down learning, the work presents an integrated model of skill learning that takes into account both implicit and explicit processes and both top-down and bottom-up learning, although the model was initially designed as a purely bottom-up learning model. We examine and simulate human data in the Tower of Hanoi task. The work shows how the quantitative data in this task may be captured using either top-down or bottom-up approaches, although we will show that top-down learning is a more apt explanation of the human data currently available in this task.

Overall, the result of our simulations suggests that both directions are possible in human cognitive skill acquisition, and the actual direction may be either bottom-up or top-down (or a mix of both), depending on task settings, instructions, and other variables. These results demonstrate the two different directions of learning (top-down vs. bottom-up), and also provide a new perspective on skill learning.

Top-Down vs. Bottom-Up: The CLARION Model

The role of implicit learning in skill acquisition and the distinction between implicit and explicit learning have been widely recognized in recent years (see, e.g., Reber 1989, Stanley et al 1989, Willingham et al 1989, Anderson 1993, Seger 1994, Proctor and Dutta 1995, Stadler and Frensch 1998). However, although implicit learning has been actively investigated, complex and multifaceted interaction between the implicit and the explicit and the importance of this interaction have not been universally recognized. To a large extent, such interaction has been downplayed or ignored, with only a few notable exceptions (e.g., Mathews et al 1989, Sun et al 2001). Similar oversight is also evident in computational simulation models of implicit learning (with few exceptions such as Cleeremans 1994 and Sun et al 2001).

Despite the lack of studies of interaction, it has been gaining recognition that it is difficult, if not impossible, to find a situation in which only one type of learning is engaged (Reber 1989, Seger 1994, Sun et al 2001). Our review of existing data has indicated that, while one can manipulate conditions to emphasize one or the other type, in most situations, both types of learning are involved, with varying amounts

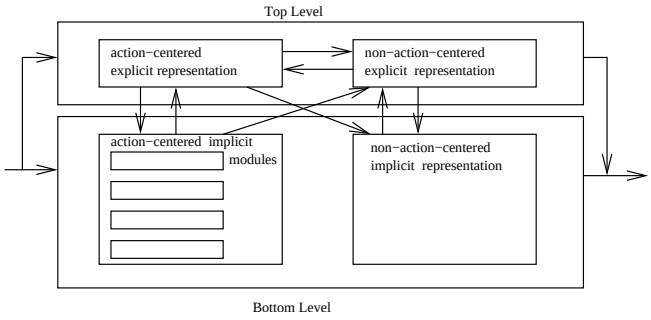


Figure 1: The CLARION architecture.

of contributions from each.

Empirical demonstrations of interaction can be found in Stanley et al (1989), Willingham et al (1989), Bower et al (1990), Wisniewski and Medin (1994), and Sun et al (2001). These demonstrations used a variety of means, including experimental manipulations such as verbalization, explicit instructions, and dual tasks.

Likewise, in the development of cognitive architectures (e.g., Rosenbloom et al 1993, Anderson 1993, Anderson and Lebiere 1998), focus has been mostly on "top-down" models (that is, learning first explicit knowledge and then implicit knowledge on the basis of the former). The bottom-up direction (that is, learning first implicit knowledge and then explicit knowledge, or learning both in parallel) has been largely ignored, paralleling and reflecting the related neglect of the interaction of explicit and implicit processes in the implicit learning literature.

However, there are a few pieces of work that did demonstrate the parallel development of the two types of knowledge or the extraction of explicit knowledge from implicit knowledge (e.g, Willingham et al 1989, Stanley et al 1989; see also Karmiloff-Smith 1986, Mandler 1992), contrary to usual top-down approaches in developing cognitive architectures.

To tackle these issues, we developed the model CLARION (Sun and Peterson 1998, Sun et al 2001). CLARION is an integrative model with a dual representational structure. It consists of two levels: the top level encodes explicit knowledge and the bottom level encodes implicit knowledge. See Figure 1. In this paper, we will focus only on action-centered components of the model.

Overall Action Decision Making

1. Observe the current state x.
2. Compute in the bottom level the "value" of each of the possible actions (a_i's) in the state x: $Q(x, a_1)$, $Q(x, a_2)$, ..., $Q(x, a_n)$.
3. Find out all the possible actions (b_1, b_2, ..., b_m) at the top level, based on the the current state information x (which goes up from the bottom level) and the existing rules in place at the top level.
4. Choose an appropriate action a, by combining (in some way) the values of a_i's (at the bottom level) and b_j's (which are sent down from the top level).
5. Perform the action a, and observe the next state y and (possibly) the reinforcement r.
6. Update the bottom level in accordance with an appropriate algorithm (to be detailed later), based on the feedback information.
7. Update the top level using an appropriate algorithm (for constructing, refining, and deleting rules, to be detailed later).
8. Go back to Step 1.

The Bottom Level

Representation The input to the bottom level consists of three groups: (1) sensory input, (2) working memory items, (3) the top item of the goal stack. The output of the bottom level is the action choice. It consists of three groups of actions: working memory set/reset actions, goal push/pop actions, and external actions. These three groups are computed by separate networks.

Learning The *Q-learning* algorithm (Watkins 1989) is a reinforcement learning algorithm. In the algorithm, $Q(x, a)$ estimates the maximum (discounted) cumulative reinforcement that can be received from the current state x on. The updating of $Q(x, a)$ is based on:

$$\Delta Q(x, a) = \alpha(r + \gamma e(y) - Q(x, a)) \qquad (1)$$

where γ is a discount factor, y is the new state resulting from action a in state x, and $e(y) = \max_b Q(y, b)$. Note that x and y include sensory inputs (internal and external), working memory items (if any activated), and the current goal (if exists).

Q-learning can be implemented in backpropagation networks (Sun and Peterson 1998). Applying Q-learning, the training of the backpropagation network is based on minimizing the following error at each step:

$$err_i = \begin{cases} r + \gamma e(y) - Q(x, a_i) & \text{if } a_i = a \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where i is the index for an output node representing the action a_i, and a is the action performed. Based on the above error measures, the backpropagation algorithm is applied to adjust internal weights (which are randomly initialized before training).

The Top Level

Representation At the top level, in contrast to the bottom level (which involves distributed representation due to the use of backpropagation networks), explicit knowledge may be captured in computational modeling by a symbolic or localist representation, in which each unit is easily interpretable and has a clear conceptual meaning, i.e., a semantic label. This characteristic captures the property of

explicit knowledge being accessible and manipulable (Sun 1995). Explicit knowledge is expressed in the form of rules.

The condition of a rule, similar to the input to the bottom level, consists of three groups: sensory input, working memory items, and the current goal. The output of a rule, similar to the output from the bottom level, is an action choice. It may be one of the three types: working memory actions, goal actions, and external actions.

Bottom-Up Learning The *Rule-Extraction-Refinement* algorithm (RER) learns explicit rules using information in the bottom level (to capture the bottom-up learning process). The basic idea of this algorithm is as follows: If an action decided by the bottom level is successful (i.e., if it satisfies a certain criterion), then the agent extracts a rule (with its action corresponding to that selected by the bottom level and with its condition specifying the current state), and adds the rule to the top-level rule network. Then, in subsequent interactions with the world, the agent refines the extracted rule by considering the outcome of applying the rule: If the outcome is successful, the agent may try to generalize the condition of the rule to make it more universal; if the outcome is not successful, then the condition of the rule should be made more specific and exclusive of the current state.

The way we measure the successfulness of an outcome (which is based on an information gain measure) and the way generalization/specialization is carried out (which is based on adding/removing allowable input values) have been fully described in Sun and Peterson (1998) and Sun et al (2001). Due to lengths, we will not repeat the details here.

Fixed Rules Some of the rules at the top level may be fixed. This type of rule (FR) represents genetic pre-endowment of an agent presumably acquired through evolutionary processes, or prior knowledge acquired from prior experience.

FRs enable top-down learning. With these rules in place, the bottom level learns under the guidance of the FRs. Initially, the agent relies mostly on the FRs in its action decision making. But gradually, when more and more knowledge is acquired by the bottom level through observing actions directed by FRs, the agent becomes more and more reliant on the bottom level (given that the cross-level combination is adaptable). Hence, top-down learning takes place.

Combining the Two Levels

In P_{RER} percent of the steps, if there is at least one RER rule indicating a proper action in the current state, we use the outcome from that rule set; in P_{FR} percent of the steps, if there is at least one fixed rule indicating a proper action in the current state, we use the outcome from that rule set; otherwise, we use the outcome of the bottom level. These probabilities are adaptable based on "probability matching" (with two parameters; Sun and Peterson 1998).

When we use the outcome from the top level, we randomly select an action suggested by the matching rules. When we use the outcome from the bottom level, we use the stochastic decision process for selecting an action: $p(a|x) = \frac{e^{Q(x,a)/\alpha}}{\sum_i e^{Q(x,a_i)/\alpha}}$, where x is the current state, a is an action, and α controls the degree of randomness (temperature) of the decision-making process.

Condition/No. of disks	2	3	4	5
No verbalization	0.0	2.1	4.3	21.2
Verbalization	0.0	0.0	0.9	1.3

Figure 2: The RT data of Gagne and Smith (1962).

Experiments

Tower of Hanoi

In the Tower of Hanoi task of Gagne and Smith (1962), there were three pegs. At the beginning, a stack of disks was stored on one of the pegs. The goal was to move these disks to another (target) peg. Only one disk can be moved at a time from one peg to another. These disks were of different sizes, and larger disks could not be placed on top of smaller disks. Initially, the stack of disks was arranged according to size so that the smallest disk was at the top and the largest was at the bottom.

Subjects were given 3-disk, 4-disk, and 5-disk versions of the task in succession, each version running until a final stable solution was found, and their mean numbers of moves (and excess moves) were recorded. Some subjects were instructed to verbalize: They were asked to explain why each move was made. The performance of the two groups of subjects (verbalization vs. no verbalization) was compared. In this task, we intend to capture the verbalization effect on performance.

Figure 2 shows the performance of the two groups in terms of mean number of excess moves (in excess of the minimum required number of moves in each version). Comparing the verbalization group and the no verbalization group in the figure, the advantage of verbalization is apparent. ANOVA indicated that there was a significant difference between verbalization and no verbalization ($p < 0.01$).

There have also been data concerning the response time of each move made by human subjects in this task. For example, the RT data from Anderson (1993) were obtained under the special instructions to subjects that encouraged a goal recursion approach (Anderson 1993). Data were available for the cases of 2, 3, 4, and 5 disks (Anderson 1993).

Bottom-Up Simulation

The Model Setup. To implement bottom-up sim-

ulation, we set up the following: (1) For deciding on each type of action (external, goal stack, or working memory actions), there is a corresponding network and a set of RER rules, respectively. (2) The input to each network is the same, including sensory input, the top goal stack (GS) item, and working memory (WM) items. (3) The outputs of the networks are external actions, GS actions and WM actions, respectively. (4) At each step, if the actions are decided by the top level, we use the existent RER rule set to get three actions—external, GS or WM actions; if the actions are decided by the bottom level, we use Boltzmann distribution to select an action from the output of each network. (5) The chosen action are coordinated and performed, and the top level and the bottom level are updated then.

During the simulation of the verbalization group, we changed the parameters for probability matching in cross-level combination to reflect the heavier reliance on the top level by the verbalization group.

Strictly speaking, GS is not necessary. But we include GS, because of generality, and because it may help learning sometimes (but it may also hamper learning sometimes). The format of GS is not important. For our simulation, each GS item includes both a subtower and a focal disc:

DSIZE: Size of SUBTOWER
FROM: Current peg of SUBTOWER
TO: Target peg of SUBTOWER
DSIZE1: Size of FOCAL-DISK
FROM1: Current peg of FOCAL-DISK
TO1: Target peg of FOCAL-DISK

A subtower is a set of disks at the top of a peg. The focal-disk is the disk beneath a subtower. Note that this set of information is redundant.

Multiple goal items could be stored in the GS one on top of another. Whenever a goal item is achieved, it will be popped.

A simple set of possible goal recursion rules is as follows (Anderson 1993):

If DSIZE $>$ 0, then push a new goal for moving a subtower of size DSIZE-1 to the spare peg and for moving the disk of size DSIZE to its target peg.

If DSIZE = 0, then make a move of FOCAL-DISK to its target peg.

If LOC(SUBTOWER)=TO and LOC(FOCAL-DISK) $\neq$ TO1, then move FOCAL-DISK to its target peg.

If LOC(SUBTOWER)=TO and LOC(FOCAL-DISK)=TO1, then pop the current goal.

Such a set of rules was hand-coded into the model in the ACT-R simulation of Anderson (1993). However, in this simulation, we did not use such hand-coded, a priori rules in the model. We want the model itself to learn something that has essentially the same effect (in both the bottom level and the top level through bottom-up learning).

The Match. The result of our simulation is shown in Figure 3. 20 runs (simulated subjects) were included in each group. Analogous to the analysis of the human data, ANOVA (number of disks $\times$ verbalization vs. no verbalization) indicated that

Condition/No. of disks	2	3	4	5
No verbalization	0.0	1.6	3.2	10.5
Verbalization	0.0	0.4	0.9	2.5

Figure 3: The bottom-up simulation of Gagne and Smith (1962).

in the model data, there was likewise a significant main effect between verbalization and no verbalization ($p < 0.01$), confirming the verbalization effect we discussed.

We compared this bottom-up simulation with a *bottom-only* simulation. We noticed that the bottom-only simulations consistently failed to learn, even when given 10 times as much training trials. This contrast suggests the importance of top-level explicit knowledge and bottom-up learning. Without them, the task was hard to learn. This fact is consistent with our synergy hypothesis (see Sun and Peterson 1998, Sun et al 2001): The reason why there are these two distinct levels (implicit and explicit) is because of the synergy that may be generated from the interaction of the two levels. The interaction of the two levels helps to improve learning, and facilitate performance and transfer (Sun et al 2001).

However, both the bottom-up and the bottom-only simulation failed to capture the RT data reported earlier.

Top-Down Simulation

The top-down simulation of the Tower of Hanoi task involves the use of fixed rules, along the line of Anderson's (1993) model, but adds the involvement of the bottom level (implicit processes), which may interfere with the top-level fixed rules. Therefore, compared with Anderson's, this is a far more complex simulation, using a more complete model that involves both explicit and implicit knowledge.

The Model Setup. Specifically, in this simulation, fixed rules were used, which implemented Anderson's (1993) analysis of subjects' performance of this task as a subset. That is, we first implemented the previous set of rules (Anderson 1993), as fixed rules at the top level of CLARION. However, this simulation was a lot more complex than top-level only (rule-based only) simulations because we had to deal with the interference from the bottom level, as the bottom level was running in parallel with the top-level rules but might recommend different actions and thus interfere with the top-level goal recursion process. The main change lied in the process of popping a sequence of goals from the GS, when a move made by the bottom level was not consistent with the top goal in the GS. In that case, we kept popping goals until reaching a goal on the GS that was consistent with the move or until the GS was empty. The structure of the GS was the same as before. The

implemented set of fixed rules was an extension of the previous set. Due to their lengths, we will not show them here.

In the bottom level, Q-learning was used. Due to the use of fixed rules, Q-learning was under the "guidance" of the top level in this case. Therefore, top-down learning was involved in this case.

For capturing the performance of the verbalization subjects, the parameters for probability matching in cross-level combination were adjusted to reflect their tendencies to rely more heavily on the top level.

The Match. The result, comparing verbalization vs. no verbalization, is shown in Figure 4. 20 runs (simulated subjects) were included in each group. Analogous to the analysis of the human data, ANOVA (number of disks × verbalization vs. no verbalization) indicated that in the model data, there was likewise a significant main effect between verbalization and no verbalization ($p < 0.01$), confirming again the verbalization effect we discussed.

In this simulation, we further tackled the capturing of the RT data from Anderson (1993), which incidentally included only a portion of the total moves in each case. The data were obtained under the special instructions to subjects that encouraged the goal recursion approach (as embodied by the fixed rules used in the top level of CLARION).

Figure 5 shows the data. The comparisons between the human and the simulation data were presented for the cases of 2, 3, 4, and 5 disks. In the data, there is a regular pattern of RT peaks, which arguably reflect planning periods during which goal recursion (establishing a sequence of subgoals to be accomplished) happens (Anderson 1993).

As demonstrated by Figure 6, it is clear that the response times of the two simulated groups were reasonably close to the human data (where there was no distinction between verbalization and no verbalization). Although the match of both groups were excellent, the match between the simulated verbalization group and the human data were closer.

This particular simulation shows that the CLARION framework can accommodate traditional accounts of human performance in this task (such as Anderson 1993, Anderson and Lebiere 1998). Moreover, it extends such accounts by incorporating implicit processes (at the bottom level) as well as explicit processes (at the top level). The role of the bottom level in this task (and other high-level cognitive skill tasks) is that of "quick-and-dirty" reactions that may lead to bad performance initially due to interference with top-level rule-guided actions, but may also lead to faster and better performance given sufficient training.

The account of human RT data is important, because such an account has been viewed as the hallmark of a successful simulation. We succeeded in showing that the two-level framework of CLARION can capture the essential patterns of the human RT

Condition/No. of disks	2	3	4	5
No verbalization	0.00	1.50	4.90	12.55
Verbalization	0.00	0.25	0.90	2.65

Figure 4: The top-down simulation of Gagne and Smith (1962).

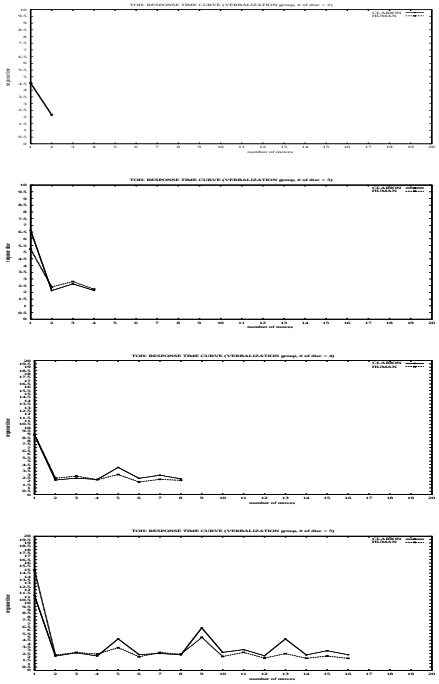


Figure 5: Simulation of the response time data of Tower of Hanoi from Anderson (1993). Four cases are included. The verbalization group is used.

data, which further testifies to the cognitive validity of the model.

Discussions

Along with the simulations of other tasks (see Sun et al 2001), we fully demonstrated that CLARION is capable of both bottom-up and top-down learning, although it was initially developed as a purely bottom-up learning model. The original reason for developing a bottom-up learning model was that in the existing literature, bottom-up learning has been very much ignored as argued by Sun and Peterson (1998) and Sun et al (2001), and therefore, there is a real need to counter-balance this bias. Our bottom-up learning model, since then, has been successful in accounting for a wide variety of skill learning tasks in a bottom-up fashion, ranging from serial reaction time tasks (sequence learning tasks), to minefield navigation tasks (Sun et al 2001). But one lingering question has been: Can this same model account for top-down learning? The present work answers this question clearly in the affirmative: CLARION can not only account for bottom-up learning data, but also

the verbalization group:

	MSE	relative MSE
2-disk	0.002	0.001
3-disk	0.529	0.107
4-disk	0.252	0.098
5-disk	1.555	0.299
overall	0.967	0.200

the non-verbalization group:

	MSE	relative MSE
2-disk	0.222	0.049
3-disk	0.086	0.024
4-disk	0.579	0.109
5-disk	3.271	0.375
overall	1.925	0.236

Figure 6: The MSEs and the relative MSEs of the RT simulations of Tower of Hanoi.

top-down learning ones. And it accounts for top-down learning equally well.

Our experiments in the TOH task showed that top-down learning is a more plausible way of accounting for the existing human data in this task. This does not come as a surprise. The task structure of TOH is highly structured, with inherent recursive embedding, and involves a small number of input/output dimensions. These characteristics naturally lead to explicit processing. This tendency is even further exacerbated by the instructions that explicitly encourage goal recursion strategies.

Acknowledgments

This work was supported in part by Army Research Institute grant DASW01-00-K-0012.

References

E. Altmann and J. Trafton, (2002). Memory for goals: An activation-based model. *Cognitive Science*, 26, 39-83.

J. R. Anderson, (1993). *Rules of the Mind.* Lawrence Erlbaum Associates. Hillsdale, NJ.

J. Anderson and C. Lebiere, (1998). *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

A. Cleeremans, (1994). Attention and awareness in sequence learning. *Proc. of Cogntiive Science Society Annual Conference*, 330-335.

R. Gagne and E. Smith, (1962). A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology*, 63, 12-18.

A. Karmiloff-Smith, (1986). From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition*. 23. 95-147.

J. Mandler, (1992). How to build a baby. *Psychology Review*. 99, 4. 587-604.

R. Mathews, R. Buss, W. Stanley, F. Blanchard-Fields, J. Cho, and B. Druhan, (1989). Role of implicit and explicit processes in learning from examples: a synergistic effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 15, 1083-1100.

R. Proctor and A. Dutta, (1995). *Skill Acquisition and Human Performance*. Sage Publications, Thousand Oks, CA.

A. Reber, (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*. 118 (3), 219-235.

P. Rosenbloom, J. Laird, and A. Newell, (1993). *The SOAR papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA.

C. Seger, (1994). Implicit learning. *Psychological Bulletin*. 115 (2), 163-196.

M. Stadler and P. Frensch, (1998). *Handbook of Implicit Learning*. Sage Publication, Thousand Oaks, CA.

W. Stanley, R. Mathews, R. Buss, and S. Kotler-Cope, (1989). Insight without awareness: on the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*. 41A (3), 553-577.

R. Sun, (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*. 75, 2. 241-296.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, 2001.

C. Watkins, (1989). *Learning with Delayed Rewards*. Ph.D Thesis, Cambridge University, Cambridge, UK.

D. Willingham, M. Nissen, and P. Bullemer, (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15, 1047-1060.

E. Wisniewski and D. Medin, (1994). On the interaction of data and theory in concept learning. *Cognitive Science*. Vol.18, 221-281.

Incremental Referential Domain Circumscription during Processing of Natural and Synthesized Speech

Mary D. Swift (mswift@ling.rochester.edu)
Department of Linguistics, University of Rochester
Rochester, NY 14627

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

James F. Allen (james@cs.rochester.edu)
Department of Computer Sciences, University of Rochester
Rochester, NY 14627

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627

Abstract

We present experimental evidence from a study in which we monitor eye movements as people respond to pre-recorded instructions generated by a human speaker and by two text-to-speech synthesizers. We replicate findings demonstrating that people process spoken language incrementally, making partial commitments as the instruction unfolds. Specifically, they establish different referential domains on the fly depending on whether a definite or indefinite article is used. Importantly, incremental understanding is observed for **both** natural speech instructions and synthesized text-to-speech instructions. These results, including some suggestive differences in responses with the two text-to-speech systems, establish the potential for using eye-tracking as a new method for fine-grained evaluation of dialogue systems and for using dialogue systems as a theoretical and experimental tool for psycholinguistic experimentation.

Background

Rapid increases in the accuracy and speed of automatic speech recognition and the increased availability of off-the-shelf text-to-speech systems has fueled great interest in spoken dialogue systems (e.g., Allen, Byron, Dzikovska, Ferguson, Galescu & Stent, 2001; Zue, Seneff, Glass, Polifroni, Pao, Hazen & Hetherington, 2000). As the sophistication of such systems increases, we can expect applications to more open-ended domains with larger vocabularies and more varied utterance types. The feasibility of such systems raises both applied and theoretical issues for work on natural language processing that crosses disciplinary boundaries. We focus on two issues here. The first, a computational issue, addresses the need for developing better evaluation tools for dialogue systems, especially tools that can evaluate comprehension on an utterance-by-utterance and within-utterance basis. The second, a psycholinguistic issue, is the possibility that in the near future implemented dialogue systems could serve as a powerful tool for developing and testing psycholinguistic models by allowing stimuli to be generated 'on the fly,' conditioned on the current state of the discourse.

A necessary prerequisite for enabling both of these goals is that people respond to synthesized speech in much the same way as they do to natural speech. We present experimental evidence from a study in which we monitor eye movements as people respond to pre-recorded instructions generated by a human speaker and by two text-to-speech synthesizers. We replicate findings demonstrating that people process spoken language incrementally, making partial commitments as the instruction unfolds. More specifically, listeners establish referential domains on the fly depending on whether a definite or indefinite article is used.

Eye movements as an evaluation tool

Spoken utterances unfold over time, resulting in a stream of temporary ambiguities. For example, as the instruction *Click on the beaker* unfolds, the word *beaker* is briefly consistent with multiple candidates, including *beetle*, *beeper*, and *speaker*. Numerous psycholinguistic studies demonstrate that people comprehend utterances continuously, entertaining multiple lexical candidates (e.g., Marslen-Wilson, 1987), making provisional commitments at points of syntactic ambiguity, and resolving reference incrementally (e.g., Altmann, 1998; Tanenhaus & Trueswell, 1995). Recent studies using eye movements to a task-relevant object in a visual workspace as people follow spoken instructions provide striking evidence for both incremental understanding and rapid integration of multiple constraints (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995; 1996; Tanenhaus, Magnuson & Chambers, forthcoming). For example, if the instruction *Click on a beaker* is presented in a context in which there are two icons of beakers and two icons of beetles, then reference will be delayed until the word *beaker* is disambiguated phonetically

(Allopenna, Magnuson & Tanenhaus, 1998). However, if there is only a single beetle, then reference will be speeded because the indefinite article *a* implies that there should be multiple referents – a condition met by the beakers but not by the beetle. Similarly, reference resolution for *Click on the beetle* will be facilitated because the beetle is the only icon that is unique. However, if the alternatives do not satisfy the uniqueness conditions associated with the article, e.g., if there is only one beaker and two beetles, and the instruction is *Click on a beaker*, then listeners are temporarily confused, looking first at the beetles before clicking on the beaker (Hanna, 2001).

If such incremental behavior carries over into recognition of synthesized utterances, then it should be possible to develop evaluation measures that can track the temporary commitments listeners make as they are processing utterances. This could establish a new evaluation methodology for speech synthesis and dialogue systems that could provide much more fine-grained information than is possible with existing techniques. This is particularly important for evaluating the quality of speech synthesis because crucial information about potential reference resolution, such as the form of an article, is carried by monosyllabic unstressed words that exhibit considerable variability with local phonetic context, as well as the overall prosodic environment of an utterance. Eye-tracking measures are good potential candidates for such an evaluation methodology because they can be incorporated into natural tasks and are well suited for any application in which the user is working within a visual workspace. The present study explores the feasibility of using eye tracking for this purpose by examining (a) whether processing is incremental when instructions are generated from a text-to-speech system and (b) whether such investigations might reveal subtle problems with synthesized speech that could impair real-time performance in natural tasks involving reference resolution.

Dialogue systems as a psycholinguistic tool

There is a growing awareness in the psycholinguistic community of the importance of examining real-time processing in natural tasks using conversational language. The advent of head-mounted eye tracking has begun to make such investigations possible. However, the field is currently facing both a theoretical and a methodological challenge. The long-term theoretical challenge is that we need theories of discourse processing that can incorporate the notion of a rich, dynamic context that characterizes the type of knowledge that listeners and speakers bring to bear on real-time interactive conversation. We suggest that practical dialogue systems, that is, dialogue systems in which participants focus on a specific task such as tutoring or problem solving, are the right grain to provide such models, if they are modified to address real-time generation and understanding. The shorter-term methodological challenge is that we need methods of generating utterances on the fly based on the current state of the discourse, in order to allow testing of alternative hypotheses, by presenting trials on which, for example, an inappropriate referential expression is used. Such trials cannot be plausibly generated by a confederate speaker, nor is it feasible to use pre-recorded instructions in any but the simplest experiments. We believe that it will soon be possible to use practical dialogue systems for this purpose. However, a crucial precondition is to determine whether listeners do indeed process synthesized utterances incrementally.

Experiment

The current experiment was intended as an initial investigation of the utility of using eye movements to evaluate spoken dialogue systems and using text-to-speech utterances in psycholinguistic experiments. We addressed the following question: Would listeners use the presence of an indefinite article compared to a definite article to differentially circumscribe potential referents as an expression unfolds? We addressed this question by examining eye movements within displays containing a pair of identical shapes and two unique shapes, using instructions such as *Click on the/a square*. Previous research with experimenter-generated instructions demonstrates that listeners assume that a definite article introduces a uniquely describable referent, whereas an indefinite article assumes that more than one referent meets the referential description (Chambers, Tanenhaus, Eberhard, Filip & Carlson, in press; Hanna, 2001).

Method

Fifteen members of the University of Rochester community were paid for their participation in this study. All participants were native speakers of English and had normal or corrected-to-normal vision. In the experimental trials, participants saw a visual display (described below) and heard an auditory instruction (one of three voice conditions) directing them to click on one of the objects on the screen. We used a within design, so each participant heard all three voice conditions (synthesizer 1, synthesizer 2 and the human voice), which were counterbalanced across experimental trials. The auditory stimuli were generated using two commercially available text-to-speech synthesizers and a digitally recorded human voice. For the human voice auditory stimuli, each instruction sentence was read aloud by an adult male volunteer and recorded with a TASCAM portable DAT recorder. The recorded voice instructions were then digitized using the SoundEdit 16 program. All auditory stimuli were minimally adjusted digitally so that the critical noun phrases were comparable in length for all three voices.

Eye movements were monitored using a lightweight head-mounted pupil/corneal reflection tracking system (ISCAN, model RK-726PCI). Calibration was monitored throughout each trial, and adjustments were made between trials if necessary. The experimental materials were presented with the PsyScope 1.0 program on a Power

Macintosh 7100/66 with a 15" color monitor.

During the experimental session, participants were seated at a comfortable distance from the computer monitor. For each trial a grid (Figure 1) appeared on the screen. The participant then clicked on the bull's eye in the center of the grid to hear the auditory instruction, e.g., *Click on the heart*, which began playing 2000 ms after the mouse click. When the participant clicked on the target object the grid was replaced by a white screen with the printed instruction *Click here for the next trial* in a random location on the screen.

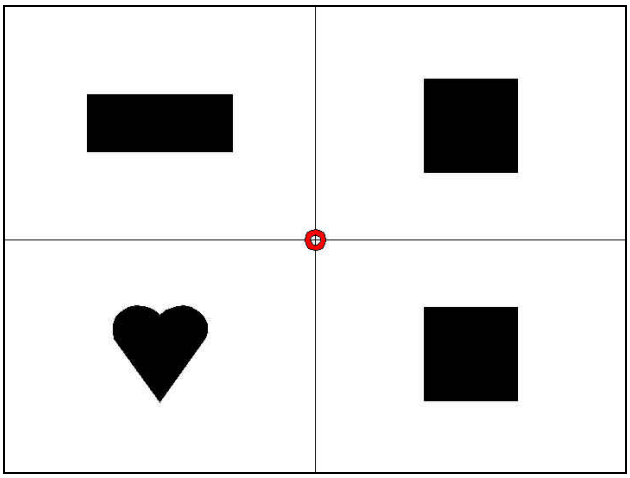


Figure 1: Sample trial screen for the experiment

Each person participated in 24 such trials, 8 in each speech condition. The voice conditions were counterbalanced across experimental trials in three lists. The order of the voice condition was pseudo-randomized so that the same voice would occur in no more than two consecutive trials. Half of the trials involved a definite noun phrase, e.g., *Click on the heart*, and half involved an indefinite noun phrase, e.g., *Click on a square.* The definite and indefinite articles were never used infelicitously – shapes referred to with an indefinite article were always duplicated and shapes referred to with a definite article were always unique.

Results

Participants clearly made use of information about definiteness when comprehending both human and synthesized speech. Recall that the display contained four objects: two were identical (duplicated) and two were unique. For a definite instruction one of the unique objects was the target. For an indefinite instruction, the participant could select either of the duplicated shapes. For instructions with definite articles (e.g., *Click on the heart*) participants were more likely to look at the unique distractor than either of the duplicated distractors ($F(2)=310.38$, $MSE=7.34$, $p<.01$), and there was no interaction with voice type. For instructions with indefinite articles (e.g., *Click on a square*) participants were more likely to look at either of the duplicated items than at the definite distractors ($F(1)=117.52$, $MSE=10.29$, $p<.01$). Again, there was no interaction with voice type.

Let us first consider the trials with instructions containing definite articles. Figure 2 shows the proportion of looks over time to the target, the unique unrelated item, and the two duplicate unrelated items in the trials with a definite article for the human voice condition. The zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *the heart*. Participants clearly use the definiteness information carried by the article because looks to the duplicate unrelated items subside approximately 100 milliseconds before the target is distinguished from the unique unrelated item. Thus the items that are consistent with the definite article are first disambiguated from the items that are not consistent with the definite article, and then the target is disambiguated from the unrelated item.

The data from the two synthesized voice conditions follow the same general pattern. Specifically, the disambiguation between unique and duplicated items (i.e., definite vs. indefinite) occurs approximately 100 milliseconds before the two unique items (i.e., definite target vs. definite unrelated) are disambiguated in each of the synthesized voice conditions.

There is, however, an important difference between the human and the synthesized voice conditions – the disambiguation points occur later in the synthesized voice conditions than in the human voice condition (Figure 3). This difference is not due to differences in the length of the articles in the three voice conditions – we have verified that these did not differ.

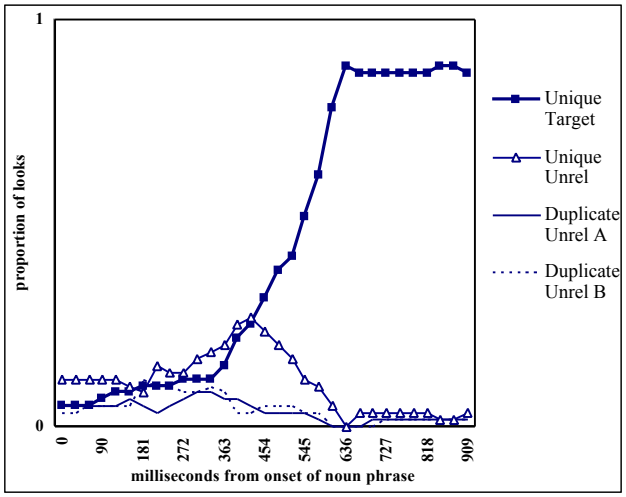


Figure 2: Results from the human voice condition: Proportion of looks to all items for definite instructions

One explanation for this difference could be that participants have greater difficulty understanding the synthesized voices. During debriefing, all participants reported that they had heard at least two distinct voices during the experiment, and at least one of these voices was

readily identifiable as synthetic.

We evaluated this hypothesis more formally by conducting a simple voice judgment survey. A new group of 16 participants listened to the auditory stimuli without the visual context and wrote down what they thought they heard for each trial. We compared their responses to the intended speech and found no differences in accuracy between the voice conditions for the definite instructions – in fact performance was at ceiling for all definite instruction auditory stimuli except one. In contrast, we observed large differences in accuracy for the indefinite instructions ($F = 6.43$, $p < .01$). The average accuracy for the human voice was 80%, while the accuracy scores for the synthesized voices were 65% for synthesizer 1 and 31% for synthesizer 2. This suggests that the delay in reference resolution for synthesized definite instructions may be due to distributional characteristics of the voices over the course of the interaction. We will return to this issue after examining the results for the indefinite instruction trials.

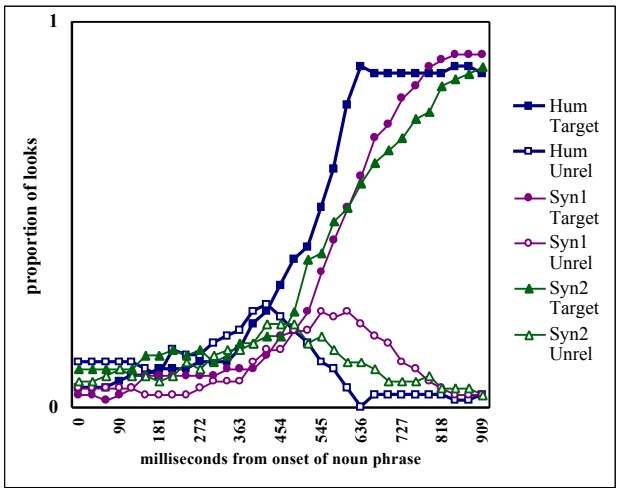


Figure 3: Proportion of looks to target (unique) and unique unrelated items for definite instructions

Now let us consider the trials with instructions containing indefinite articles. Figure 4 shows the average proportion of looks over time to the target (duplicate) items and the unrelated (unique) items for each of the three voice conditions during the trials with indefinite instructions, e.g., *Click on a square*. Note that in the indefinite condition, either of the duplicated items is an appropriate target in response to the spoken instruction. For clarity of presentation, looks to either of the indefinite targets are summed together and represented as a single line for each of the voice conditions in Figure 3. Similarly, looks to either of the unrelated (unique) items are summed together in a single line for each voice. Again, the zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *a square*.

For all voice conditions, looks to the duplicated items diverge from looks to the unique items at roughly the same point. We cannot tell from this data whether these eye movements are due to processing of the indefinite article or whether they are due to processing of the noun. It is surprising that we do not see differences in the time course of looks between the voice conditions, given the differences in accuracy for the voice judgement survey, but an examination of looks to the two duplicated items may provide an explanation.

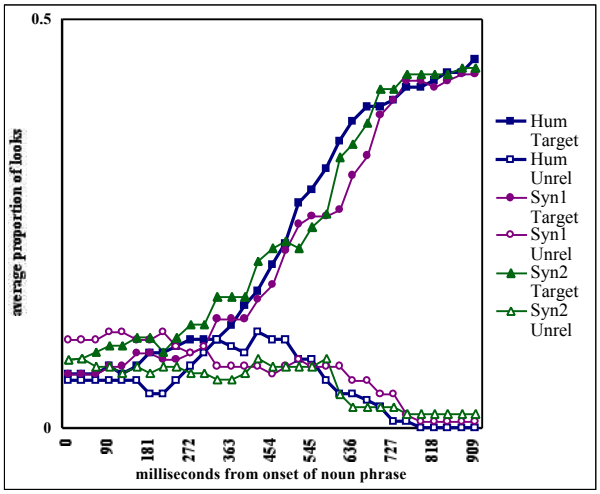


Figure 4: Average proportion of looks to target (duplicate) and unrelated (unique) items for indefinite instructions

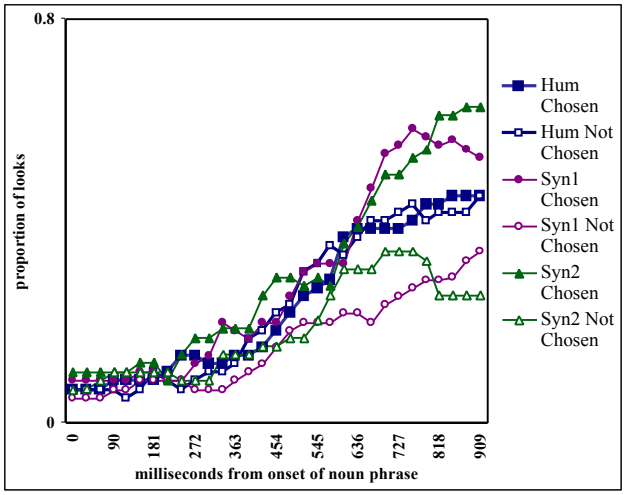


Figure 5: Proportion of looks to the target (duplicate) items chosen and not chosen for indefinite instructions

Figure 5 shows the proportion of looks over time to the two possible target items. For each voice condition the item identified as "chosen" is the duplicated item that the participant eventually clicked on and the item identified as "not chosen" is the other duplicated item.

For instructions in the human voice condition, participants considered each of the duplicated items before clicking on one, reflecting the expected circumscription of referential

domain in the indefinite condition. For instructions in the synthesized voice conditions, participants tended to click on the first of the duplicated items that came to their attention, reflecting a more restricted referential domain than expected – one more appropriate to a definite article interpretation.

These results demonstrate that participants make different assumptions about the felicity of the use of the article for the synthesized speech instructions than for the natural speech instructions, due to global differences in how well the indefinite articles could be understood in the three voice conditions. This could also explain the delays in disambiguation for the definite article instructions.

Implications

People process spoken language continuously, even though continuous recognition entails resolving numerous temporary ambiguities on the fly. We have shown that this mode of recognition carries over to speech that is clearly identifiable as computer-generated artificial speech. The results suggest that this paradigm can be used to provide a fine-grained evaluation of comprehension during human-computer dialogue. Specifically, during reference resolution, listeners use cues such as definiteness that are often carried by short unstressed words that are difficult to synthesize. Lack of clarity in synthesizing these words may interfere with reference resolution. While perhaps not a problem in such simple tasks as these, we can expect it to be more problematic in more complex applications, and as the global characteristics of the speech cause interactions with additional error sources, such as unnatural prosodic cues. The important point here, however, is that the eye-tracking technique can reveal even subtle comprehension problems at a fine degree of temporal resolution. This suggests that the same technique could be used to evaluate components such as those affecting lexical choice, sentence structure, intonation and even higher-level discourse intentions. In addition, the eye-tracking paradigm may provide a valuable new method of comprehension evaluation in multimodal language applications using visual displays.

Moreover, our finding that people naturally process synthesized speech incrementally means that computational dialogue-based systems have the potential to be a psycholinguistic tool, especially for experimental questions where it is important to be able to generate utterances on the fly. By using such systems, we could generate more complex stimuli than is possible using a confederate or pre-recorded speech. While there remains much to be done to make this a reality, the range of experiments it would enable is great.

Acknowledgments

This research was supported by NIH HD-27206 to MKT.

References

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*, *22*, 27-35.

Chambers, C.G., Tanenhaus, M.K, Eberhard, K.M., Filip, H. & Carlson, G.N. (in press). Circumscribing referential domains in real-time sentence comprehension. *Journal of Memory and Language*.

Allopenna, P., Magnuson, J., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439.

Altmann, G. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, *2(4)*, 146-152.

Hanna, J. E. (2001). *The effects of linguistic form, common ground, and perspective on domains of referential interpretation*. Doctoral Dissertation, Department of Brain and Cognitive Sciences, University of Rochester.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71-102.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*, 109-147.

Tanenhaus, M. K., Magnuson, J., & Chambers, C. (forthcoming). Eye-movements and spoken language comprehension: Bridging the language as action and language as product tradition. *Trends in Cognitive Science*.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, *268*, 1632-1634.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1996). Using eye-movements to study spoken language comprehension: Evidence for visually-mediated incremental interpretation. In T. Inui & J. McClelland (Eds.), *Attention & Performance XVI: Integration in Perception and Communication*. Cambridge: MIT Press.

Tanenhaus, M. K., & Trueswell, J. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *Handbook of Perception and Cognition*. San Diego: Academic Press.

Zue, V., Seneff, J., Glass, J., Polifroni, J., Pao, C., Hazen, T., & Hetherington, L. (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, *8(1)*.

The Role of Consciousness in Second Language Acquisition

Edina Torlaković (edina_@scs.carleton.ca)
Institute of Interdisciplinary Studies; Cognitive Science, Carleton University
2214 Dunton Tower, 1125 Colonel By Drive
Ottawa, ON
K1S 5B6

Andrew Brook (abrook@ccs.carleton.ca)
Institute of Interdisciplinary Studies; Cognitive Science, Carleton University
2216 Dunton Tower, 1125 Colonel By Drive
Ottawa, ON
K1S 5B6

Abstract

In this paper we argue that in order to resolve the controversy in Second Language Acquisition research concerning whether or not direct instruction is needed for second language acquisition, we need to use a broader sense of 'consciousness' than is used by second language researchers. Block's classification of consciousness into Access and Phenomenal consciousness seems promising. We associate Phenomenal consciousness with explicit knowledge and suggest that explicit instruction is useful. It enhances linguistic competence.

Introduction

This paper addresses a question that is of great importance for Second Language Acquisition (SLA) research. The question is 'what should the role of consciousness in second language (L2) acquisition be?' It is important to answer this question in order to resolve one of the biggest debates in the field of SLA, namely whether or not direct instruction is necessary or even valuable in L2 acquisition.

SLA researchers interested in consciousness should start by considering what others have to say about it. This is necessary to develop a comprehensive picture of consciousness. The debate in SLA needs to be informed by an adequate notion of what consciousness is. Only in this way can we reach an adequate view about its role.

One place to start is to consider what is said about consciousness in philosophy. We will start by comparing the different definitions of consciousness used by SLA researchers and by philosophers. Next we will introduce the controversy over whether L2 learners need to be conscious of grammar rules to learn the target language. Then we will examine Block's well-known distinction between access (A) consciousness and phenomenal (P) consciousness and where language, or more specifically second language, fits into this categorization. With this, we might be one step closer to understanding the role of consciousness in L2 learning/acquisition.

Issues and Positions

Definition(s) of Consciousness

How do SLA theorists and philosophers think about consciousness? As it turns out, quite differently. Let us look at some of the similarities and differences.

When SLA theorists talk about consciousness, they use the term in a quite narrow sense. Schmidt (1995), for example, points out that there are three different senses of the term 'consciousness' as it is used in SLA theory: levels of perception, noticing, and understanding. By contrast, philosophers have a broader understanding of the term. According to Clark (2001), the possibilities include wakefulness, self-awareness, availability for verbal report, availability for control of intentional action, and qualia.

To determine if all these terms are discussing the same, complex entity, they need to be further defined. If one desires to apply concepts of one discipline to another (philosophy to SLA in this case), this is something that we need to know.

According to Schmidt, 'levels of perception' could be defined as levels of a process of obtaining and perhaps processing information. Schmidt defines 'noticing' as rehearsal in short-term memory, while by 'understanding' he refers to rule understanding, i.e., grasping the meanings of rules and becoming thoroughly familiar with them.

Definitions of the terms from Clark's list of possibilities might go as follows: wakefulness is defined as a state in which we are sensitive to our surroundings and in which we can process incoming information and respond to it appropriately. Self-awareness he defines as a capacity to represent ourselves and to be conscious of ourselves 'as distinct agents'. Availability for verbal report is the capacity to access our own inner states and to describe them using natural language, while qualia concerns how things feel to us.

From the above, one can conclude that SLA theorists take consciousness to be something narrower than philosophers

believe. Perception and wakefulness may refer to (or be contained in) the same aspect of consciousness, while noticing and understanding could be seen as part of availability for verbal report. However, self-awareness and qualia are missing from the SLA picture of consciousness. Yet in second language learning (SLL) and acquisition, self-awareness and qualia may play an important role. It is well known that language is closely associated with consciousness in the broader understanding that we find in Clark and other philosophers. If so, this broader notion of consciousness needs to be considered by SLA theorists. We will return to this topic. For now, let us simply note that SLA researchers use a narrow notion of consciousness.[1]

Consciousness and SLA

Next we want to consider a group of related issues: the role of consciousness in various SLA theories the debate in SLA and L2 pedagogy about its proper role, the role of Universal Grammar (UG) in L2 acquisition , and the respective roles of implicit and explicit learning in SLA.

According to Robinson (1996), current debate in SLA is centred on the role of consciousness in L2 development. This controversy is centred in turn on the question of whether or not grammatical instruction is effective for L2 acquisition and if so what kind of grammatical instruction is best. There are researchers who argue that grammatical instruction has only minimal effect on L2 acquisition, Krashen (1981) for example. According to him, L2 development is largely an unconscious process. Krashen does allow that there are two processes involved in L2 development, a conscious process of learning and an unconscious process of acquisition. The conscious process of learning is a system based on rules and their application, while the unconscious process of acquisition is a system responsible for language production. According to Krashen, conscious learning is limited to a small set of simple rule-governed domains. By contrast, development of the much more substantial acquired system is fostered by avoiding instruction and the provision of L2 rules. In his view, learners only have to be exposed to comprehensible language input in order to acquire grammar.

On the other side, there are researchers who argue that comprehensible input alone is not enough for optimal acquisition of the different aspects of grammar and that conscious grammatical instruction is necessary if learners are to have the data they need to acquire grammar (Strozer, 1994). In particular, Schmidt (1994) argues that consciousness of input at the level of noticing is a necessary condition for L2 development. Many other researchers support this view. They use terms such as focus-on-form (Long, 1994), consciousness-raising (Ellis 1993, Fotos and Ellis 1991, Rutherford, 1987), and input-enhancement (Sharwood Smith, 1991). In one way or another, all of these terms are about directing learners' attention to grammatical form in order to help them internalize the L2 system. According to these researchers, teaching should include opportunities for learners to focus on form and consciously notice features of the L2 they are learning.

Universal Grammar and L2 Acquisition

It might appear that because there is little or no need for conscious instruction in L1 acquisition, there is little or no need for it in L2 acquisition either. However, it is well known that the two processes are quite different from one another. Let us compare the two in terms of the theory of Universal Grammar (UG).

According to Chomsky (1980), all of us have an innate capacity for language and we cannot chose not to learn language. We have a mental faculty for language that simply 'grows' as any other organ of our body grows. All that we need is a triggering cause, namely, a language environment. For L1 acquisition, little or no direct teaching is needed.

According to Flynn (1996), the theory of UG does not make any direct claims about L2 acquisition. However, it is important to know whether or not L2 learners in the process of L2 acquisition have access to UG. Ellis (1997) points out that different theories deal with this issue differently. However, there is some good evidence for the a Partial Access Hypothesis, which holds that only the parameters of UG that are common to L1 and L2 are accessible to an L2 learner. According to this view, an L2 learner needs to learn everything else by using general problem-solving strategies. If this is so, there is clearly room for direct conscious instruction in L2 learning.

Explicit and Implicit Knowledge and SLA

It is obvious that in L2 acquisition, both explicit and implicit learning are present. But that is not the same thing as direct conscious instruction being *necessary* for L2 learning. Nor does it say anything about the effects of each type of learning. So let us look these two kinds of learning and their connection(s) to consciousness. Before we enter this inquiry, let us define the two types of knowledge that these two types of learning yield.

According to Ellis, explicit knowledge is "the L2 knowledge of which a learner is aware and can verbalize on request", while the implicit knowledge is "the L2 knowledge of which a learner is unaware and therefore cannot verbalize." (Ellis, 1997,139). We can report explicit knowledge, while we are not aware of implicit knowledge.

Hulstijn and Graaff (1994) attempt to determine to what

[1] It is important to mention that the above are not the only definitions of consciousness in either discipline. However, they are the most common ones.

extent SLA and acquisition of implicit knowledge can be assisted by explicit learning (instruction). According to them, learning varies from spontaneous discovery by a learner to explicit instructions by a teacher. They argue that in fluent speakers, knowledge of L2 is mostly implicit. That, however, does not settle the question of whether, before native-like fluency in L2 is reached, there is a need for explicit instruction.

There are two positions concerning the question, 'how fluent can a speaker become without explicit knowledge?'. They are the Noninterface Position, which argues that implicit knowledge is not influenced by explicit knowledge, and the Interface position, which urges that the acquisition of implicit knowledge may be influenced by explicit knowledge. The Interface Position is divided into a Strong-Interface Position and a Weak-Interface Position. According to the Strong Position, explicit knowledge becomes implicit knowledge through practice. This position is derived from skill acquisition theory; L2 acquisition is seen as the automatization of the application of explicit grammar rules. According to the Weak Position, explicit knowledge only aids the acquisition of implicit knowledge. If a learner is ready for the new knowledge, his conscious knowledge will become implicit. Application of implicitly knowledge can merely be improved through explicit instruction (Ellis, 1993).

Let us now turn to philosophical definition(s) of consciousness.

Consciousness in Philosophy

The initial task of this paper was to consider what philosophers have to say about consciousness in order to search for conceptual issues that may ease and perhaps resolve the current debate in SLA related to the role of consciousness in L2 learning and/or acquisition. Let us return to the issue of what philosophers have to say about consciousness.

Block (1999) introduced an interesting categorization of consciousness into A-consciousness and phenomenal P-consciousness. Block argues that A-consciousness is informational processing and control of thought and action. According to him "a state is A-conscious if it is poised for direct control of thought and action, .. for free use in reasoning and for direct 'rational' control of action and speech." (Block, 1999). By contrast, he defines P-consciousness as what we see, smell, taste, and feel. According to Block, P-consciousness is what it is like to have sensations, feelings, perceptions, thoughts, wants, and emotions: "what makes a state phenomenally conscious is that there is something it is like to be in that state" (Block, 1999). P-consciousness is what we ordinarily call experience.

Block points out that there are three main differences between A-consciousness and P-consciousness. The first difference concerns content. P-consciousness content is phenomenal (it is like something to have it) while the content of A-consciousness is representational. The latter enters into reasoning, behavioural control, etc. The second difference is that A-consciousness is defined in terms of functions in a cognitive system while P-consciousness is not. The third difference is in the paradigms of each type of consciousness. The paradigmatic cases of P-consciousness are sensations, while those of A-consciousness are propositional attitudes.

As for the relationship between A-consciousness and P-consciousness, Block argues that even though A-consciousness and P-consciousness are separate entities, they do interact, influence one another and might even be the product of one another. A P-consciousness change in what is figure and what is ground, for example, might have functional effects on what one comes to believe or do. However, lack of one type of consciousness does not guarantee lack of the other. We will return to the issue of whether it is possible to have A-consciousness without P-consciousness or vice-versa.

Let us now examine how L2 acquisition and language in general fit into the Block's distinction.

Philosophical Views of Consciousness and the Issue of Conscious Instruction

So far we have laid out the controversy in SLA on the question of whether L2 learners benefit from direct grammar instruction and we had looked at a philosophical view of consciousness. Let us now try to connect the two. The hope is that philosophy can help us to ease this controversy. How does L2 acquisition fit into the distinction between A-consciousness and P-consciousness?

A-consciousness plays an important role in reasoning and information processing. It is closely related to 'knowing how to do something'. P-consciousness, by contrast, is 'it being like something to be in some state'. And, as we saw, however, different these two notions are, the two kinds of consciousness interact. In particular, one can be P-conscious of knowing how to do something. Similarly, the way something feels to you can make a difference to cognitive functioning. What we now need to consider is how the two types of consciousness relate to our knowledge of language, in particular our knowledge of the syntactic structure of the language.

Let us introduce Chomsky's distinction between competence and performance. Competence refers to a speaker's knowledge of the language while performance is the actual use of language and reflects not only competence but also such other factors as ability to utilize competence, time constraints, and so on. Performance is the actual use of language in different situations, how we actually speak, use, or manipulate language. Linguistic performance is part of a lot of A-consciousness. Competence, ' speaker's/hearer's knowledge of language', is not tied in the same way to P-consciousness. Most of our competence is in this sense

*un*conscious. Still, there is a relationship between competence and P-consciousness, as we will see. To bring out this relationship, let us connect A-consciousness and P-consciousness to implicit and explicit knowledge. We will argue that P-consciousness is similar, if not identical, to explicit knowledge and that explicit instruction which eases the acquisition of implicit knowledge enhances competence.

Start with UG. If UG is innate, it is not dependent on P-consciousness. What is developed after the triggering effect of the language environment is at least A-consciousness, 'poised for control of thought and action'. A-consciousness of language is not present at birth. A-consciousness of language is indirectly influenced by P-consciousness of language with which it interacts even during the developmental phase because children not only come to use language, it is like something for them to have language (in the usual, P-conscious sense of the term, they are conscious of the language they know), and this consciousness has effects on how they use language. The two types of consciousness of language develop roughly simultaneously. However, that does not show yet that P-consciousness of language enhances any of acquisition, competence, or performance.

Time to bring L2 acquisition back onto the stage. Does P-consciousness have a special role to play in it? We think it does. The Partial Access Hypothesis introduced earlier in this paper shows why. If an L2 learner needs to learn all the parameters of UG that are not common to his L1 and L2, that means that the parameters peculiar to his L2 are not included in his current linguistic competence. If all UG parameters are present at birth, then the UG parameters peculiar to the L2 were lost at some point during or after the process of L1 acquisition. These missing parameters is a major difference between L1 and L2 acquisition. Put in terms of the language of consciousness, the parameters that have dropped out are in neither the subject's A-consciousness or P-consciousness of language. (An implication of this is that, not surprisingly, consciousness of language is language specific.)

We should agree with Krashen when he points out that there are two different processes in L2 development, namely learning and acquisition. However, it does not follow that acquisition is fostered by avoiding explicit instruction. Recall that in L2 learning, both implicit and explicit knowledge are present. It is plausible to suggest that if a required piece of linguistic competence is no longer part of the current competence of a learner, then he will need to learn it explicitly in order to (re)gain an implicit, automatized ability to use it again.

Earlier in this paper Ellis' definition of explicit and implicit knowledge with regard to L2 learners was accepted. However, the definition of the two types of knowledge needs to be modified in terms of P-consciousness and A-consciousness of language. Explicit knowledge is something that a learner is P-conscious of and can verbalize on request. By contrast, though some implicit knowledge is A-conscious, it is by definition never P-conscious. Even though we are not P-conscious of our implicit knowledge, the latter can be influenced by explicit, P-conscious knowledge, just because A-consciousness can be influenced by P-consciousness.

Now that we have presented arguments that linguistic performance consists (at least often, maybe always) in A-consciousness of language and P-consciousness of language can enhance implicit, A-conscious competence and performance, let us now look at their interaction in a bit more detail, to try to see where explicit instruction fits.

We presented the Interface Positions earlier in this paper and said that according to the Weak Position explicit instruction in L2 directly influences explicit knowledge which aids the acquisition of implicit knowledge. Given the connection between A-consciousness of language and implicit knowledge and P-consciousness and explicit knowledge, the Weak Position and Block's (1999) view that P-consciousness can influence A-consciousness are in line with one another.

In L1 acquisition, there is no need for direct instruction. During this process implicit knowledge of language, including A-consciousness of it and P-consciousness or explicit knowledge of language are present. On the one hand, native speakers can always judge whether on not a sentence is grammatically acceptable. However, in most cases they cannot explain why. They have explicit knowledge of the sentence's grammaticality, only implicit knowledge of why. In the case of L2 learners, the situation is quite the opposite. If they can judge the grammaticality of a sentence, they can also cite the relevant rules. Implicit knowledge of rules plays little role since the necessary competence is not innate and yet the acquisition process has not yet rendered it automatic and implicit. Indeed, it is not rare that an L2 learner is capable of explicitly spelling out a grammatical rule of his L2 and yet cannot apply it in his spoken or written L2 productions. To sum up, explicit grammar instruction, essential for acquisition of explicit knowledge (P-consciousness), can also enhance implicit linguistic competence and performance. For that reason, it should be used in L2 teaching (see Figure 1).

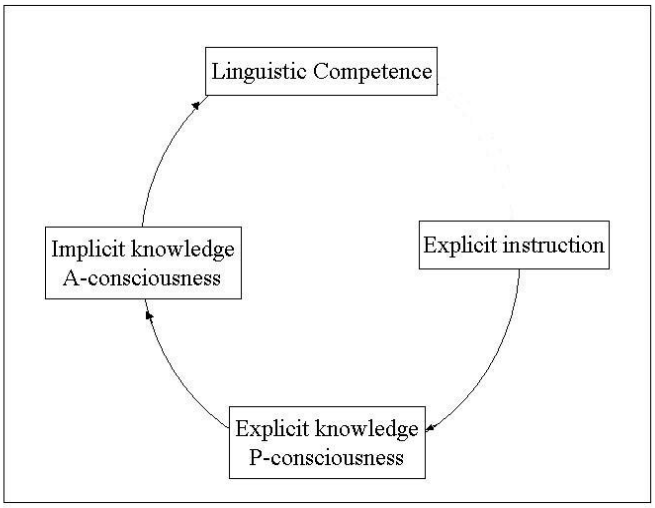


Figure 1: Indirect effect of explicit instruction in L2 on linguistic competence

Conclusion

In this paper we examined an important question for SLA theory, namely, the role of consciousness in L2 acquisition. This question was examined in order to determine whether or not explicit instruction in grammar is advantageous during this process.

We first compared the definitions of consciousness accepted in the two disciplines and concluded that perhaps SLA theorists need to consider consciousness in a broader sense than they do.

We next looked at how different SLA theories view the role of consciousness and we examined the controversy in SLA theory concerning the value of explicit instruction in grammar. In this connection, we paid special attention to the role of UG in L2 acquisition and role of implicit and explicit knowledge in use of UG.

This led us to Block's distinction between A-consciousness and P-consciousness. We examined how L2 acquisition and language in general fit this categorization. We suggested the following:

- linguistic performance is a form of A-consciousness, often at least,
- subjects can be both A-conscious and P-conscious of language, including UG,
- A-consciousness of language can be language specific,
- A-consciousness of language is a form of implicit knowledge, while P-consciousness of language is explicit knowledge,
- P-consciousness can influence A-consciousness, and,
- P-consciousness of language can enhance linguistic competence by improving implicit knowledge.

The relationship of all this to explicit instruction in L2 teaching is as follows. Explicit instruction influences explicit knowledge, obviously. Having explicit knowledge means being P-conscious. But having explicit knowledge or P-consciousness of language can influence one's A-consciousness of it. In this way, P-consciousness can enhance L2 competence and implicit knowledge (see figure 1).

This paper tries to provide a theoretical base for the value of explicit grammar instruction in L2 training. What needs to be examined next is the form of the instruction. One of us has pointed out elsewhere that instruction in grammar should take the form of task-based, form-focussed instruction that contains both positive and negative evidence (Torlaković, 2001).

References

Block N. (1999). On a Confusion About a Function of Consciousness. In N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press, 375-416.

Burge , T. (1999). Two Kinds of Consciousness. In N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press, 427-434.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford, Oxford University Press.

Chomsky, N. 1980. Rules and Representations. *The Behavioral and Brain Sciences*. 3. 1-61.

Ellis, R. The Structural Syllabus and Second Language Acquisition. *TESOL Quarterly*. 27. 91-113.

Ellis. R. 1997. *Second Language Acquisition*. Oxford: Oxford University Press.

Flynn, S. 1996. Parameter Setting Approach. In Ritchie, W. and T. Bhatia. *Handbook of Second Language Acquisition*. San Diego: Academic Press.

Fotos, S., & Ellis, R. (1991). Communicating About Grammar; A Task-based Approach. *TESOL Quarterly*, 25, 87-112.

Hulstijn, J. H. and R. Graaff. 1994. Under What Conditions Does Explicit Knowledge of a Second Language Facilitate the Acquisition of Implicit Knowledge. *Aila Review*. 11, 97-113.

Jackson, F. (1999). What Mary Didn't Know. In N. Block, O. Flanagan, and G. Guzeldere (eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press, 567-570.

Krashen, S. (1981). *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press.

Robinson, P. (1996). *Consciousness, Rules, and Instructed Second Language Acquisiton: Theoretical Studies in Second Language Acquisition*. New York: Peter Lang Publishing Inc.

Rutherford, W. (1987). *Second Language Grammar: Learning and Teaching*. London, New York: Longman.

Schmidt, R. (1995). Consciousness and Foreign Language: A Tutorial on the Role of Attention and Awareness in Learning. In R. Schmidt (ed*.), Attention and Awareness in Foreign Language Teaching and Learning*, pp. 12-55. University of Hawaii at Manoa: Second Language Teaching and Curriculum Center Technical Report # 9.

Schmidt, R. 1994. Deconstruction Consciousness in Search of Useful Definition for Applied Linguistics. *Aila Review*. 11.11-26.

Sharwood Smith, M. (1991). Speaking to Many Minds: On the Relevance of Different Types of Language Information for the L2 Learner. *Second Language Research*, 7. 118-132.

Strozer, J. 1994. *Language Acquisition After Puberty*. Washington: Georgetown University Press.

Torlaković, E. (2001). *Application of a CALL System in the Acquisition of Adverbs in English*. Carleton University, MA thesis.

The Instantiation and Use of Conceptual Simulations in Evaluating Hypotheses: Movies-in-the-Mind in Scientific Reasoning

Susan B. Trickett (stricket@gmu.edu)
Department of Psychology, George Mason University
Fairfax, VA 22030-4444 USA

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory, NRL Code 5513
Washington, DC 20375 USA

Abstract

This study investigates the strategies used by expert scientists to evaluate hypotheses when they analyze data. We used an *in vivo* methodology to observe experts' on-line thinking. In contrast to the results of laboratory studies of scientific reasoning, we found that the scientists rarely used experimentation but relied on a variety of other strategies, including conceptual simulation. This strategy was most prevalent in evaluating a hypothesis about a phenomenon that violated the scientists' expectations.

Introduction

How do scientists test and evaluate hypotheses? One obvious answer is that they design and conduct experiments. The canonical method of scientific inquiry is represented by a cycle of hypothesis generation, experimentation, data analysis and hypothesis refinement that has its roots in the philosophy of science (Popper, 1956) and is frequently taught explicitly to students (Okada & Shimokido, 2001).

Psychologists investigating the processes of scientific reasoning have also been influenced by the "scientific method" and so have focused on experimentation in investigating hypothesis-evaluation strategies,. There have been numerous laboratory studies of scientific reasoning in which participants are asked to find the cause of a given effect (e.g. Dunbar, 1993; Schunn & Anderson, 1999), or to identify the role of a causal mechanism (e.g., Klahr & Dunbar, 1988; Trafton & Trickett, 2001a; Trickett, Trafton, & Raymond, 1998; Vollmeyer, Burns, & Holyoak, 1996). In these studies, participants propose hypotheses, then design and run experiments to test them

There are several reasons why participants in laboratory studies of science use experimentation to evaluate hypotheses. The instructions in these studies explicitly tell participants to run experiments. Participants have little choice—they are provided with limited time, equipment, and materials. Moreover, they are frequently asked to reason in a domain about which they have no relevant knowledge. Running an experiment is also "cheap"—the variables are already identified, it involves a few mouse-clicks, and the results are almost instantaneous and easy to interpret.

However, practicing scientists have a wider array of options. They can select their own methods and equipment, and, as experts, they have domain knowledge to guide their problem-solving. Experimentation may *not* be the best strategy, as it is expensive in terms of planning, paperwork, personnel, the need for special equipment, the complexity of data interpretation, and the high cost of errors.

What strategies besides experimentation might scientists use to evaluate hypotheses? Prior research on scientific thinking suggests several possibilities. One likely strategy is extracting information from data, whether by reading off information, transforming data, replotting data), or looking at data that is not currently on view but that is available. Trafton found that expert meteorologists spent considerable time on information extraction (Trafton et al., 2000).

Given the cost of experimentation, it is also likely that scientists use different strategies to reason about hypotheses before committing to an experiment. Analogical reasoning has been shown to be a powerful strategy in science (Clement, 1988; Dunbar, 1997; Gentner et al., 1997). It allows people to make inferences about an unknown entity based upon their knowledge of a different, known entity (Gentner, 1983) and has been proposed as a mechanism of conceptual change in numerous historic scientific advances (Gentner et al., 1997; Nersessian, 1992; Thagard, 1992). It is also a strategy used by successful contemporary scientists in scientific problem-solving, such as hypothesis generation (Clement, 1988), experimental design (Dunbar, 1997), and discovery itself (Ueda, 1997). Given its widespread use in other aspects of scientific reasoning, it seems plausible that analogy may be used as a hypothesis-testing strategy; however, whether this is the case remains an open question.

Conceptual simulation has also been shown to be a means of successful scientific reasoning (Nersessian, 1999; Qin & Simon, 1990; Schraagen, 1993). A conceptual simulation is a mentally constructed model of a phenomenon or data representation that is manipulated in such a way that there is a resulting change of state (a formal definition is provided below). As with analogy, conceptual simulations have been proposed as a strategy used by both historical and practicing scientists. In historical reconstructions, Ippolito and Tweney have developed a model of insight that involves the construction of a dynamic, "runnable" mental model (Ippolito & Tweney, 1995), and Nersessian proposes that scientists construct and conduct mental experiments that yield usable data, in a process that mirrors an empirical experiment (Nersessian, 1999). In contemporary scientific problem-solving, Hegarty has found that people develop sequences of mental animations (Hegarty, 1992). Qin and Simon (1990) found that people used a series of mental processes of manipulation, control, and inspection in order to extract information that was only implicit in their initial mental image. Similarly, participants in Schraagen's study of experimental design used a strategy of mental simulation to project what experimental procedures would look like under particular circumstances (Schraagen, 1993). As with

analogy, how much scientists use conceptual simulations in evaluating hypotheses remains an open question.

One can imagine several other means whereby scientists might evaluate a hypothesis. For example, a scientist might consult a colleague or other expert or attempt to tie the hypothesis into current theoretical understanding of the domain. A scientist might also defer evaluation until some later time or even abandon a hypothesis altogether.

The purpose of this research is to investigate the means by which scientists evaluate hypotheses. In order to investigate this issue, we adapted Dunbar's *in vivo* methodology (Dunbar, 1997), an observational technique developed to study creative and complex thinking in a real-world context. The main advantage of Dunbar's method is that it allows the collection of on-line measures of thinking by experts engaged in authentic scientific tasks.

Method

We chose to investigate scientists at work during the data analysis phase of their research because it is a stage at which a great deal of scientific reasoning takes place. Scientists must integrate their expectations about the data with the actual data; it is thus likely to be rich in hypotheses.

We analyzed 8 different datasets from 9 scientists working in one of 4 domains—neuroscience, astronomy, computational fluid dynamics (CFD), and psychology. Each dataset consists of a recorded session in which one or more scientists analyzed their data.

Participants were all working scientists recruited through personal connection of the experimenters. Either they were expert scientists who had earned their PhDs more than 6 years previously, or they were graduate students working alongside one of these experts. Only experts with a Ph.D. worked alone; in the group sessions involving graduate students, the scientist in charge always had a Ph.D.

Participants agreed to contact a member of the research team when they were ready to conduct some analysis of recently acquired data, and an experimenter visited the scientists at their regular work location. Participants working alone were trained to give talk-aloud verbal protocols. For scientists working in groups, we recorded their conversation as they engaged in scientific discussion about their data. All participants were instructed to carry out their work without explanation to the experimenter (Ericsson & Simon, 1993). It is important to emphasize that all participants were performing their usual tasks in the manner in which they typically did so. At the beginning of the session, some participants gave the experimenter an explanatory overview of the data and the questions to be resolved, and after the session, the experimenter interviewed the participants to gain clarification about any uncertainties. During the analysis session itself, however, the experimenter did not interrupt the participants.

All utterances were later transcribed and segmented according to complete thought. All segments were coded by 2 coders as on-task (data analysis) or off-task (e.g., software management, phone interruptions, jokes, etc.). Inter-rater reliability for this coding was more than 95%. Introductory comments from the scientists to the experimenter and post-session interviews of the scientists were excluded from analysis. The number and percentage of on-task utterances, the number of participating scientists, and the duration of the relevant portion of each individual session are reported in Table 1. Finally, a coding scheme (described below) was developed to examine how the scientists evaluated hypotheses they developed in the course of analyzing data.

Table 1: Characteristics of datasets

Domain	Utterances: On-Task	Total	Time (mins)	# scientists
Astronomy	649	859	49	2
CFD sub	430	954	39	1
CFD laser 1	172	400	15	1
CFD laser 2	184	249	13	1
fMRI	317	373	55	2
Neuroscience	219	343	54	2
Psychology 1	482	541	31	3
Psychology 2	914	1426	75	2

Although each scientist or group used different tools, their tasks shared several characteristics. All the scientists were analyzing data that they themselves had collected, from observations, from a controlled experiment, or from running a computational model. They displayed this data using their regular tools, whether custom-built visualization programs, while others used widely available commercial products, such as Microsoft's Excel. Figure 1 shows an example of the type of data examined by the astronomers. Visualizations used in other domains were similarly complex.

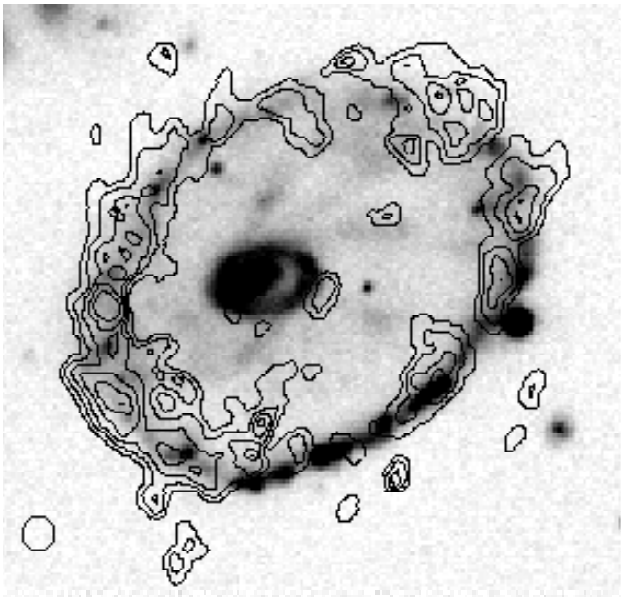


Figure 1: Example of data examined by astronomers. Radio data (contour lines) are laid over optical data.

Almost all sessions represented the initial investigation of this data (the exception was the second CFD session, which was a follow-up to the first session). Although in some sessions the scientists did not have strong *a priori* beliefs about the data (these sessions were thus exploratory), in others, the scientists did approach the task with particular hypotheses that they expected to be supported by the data. It is interesting to note, however, that none of the scientists performed any statistical analyses

Coding Scheme

In addition to coding all segments as on- or off-task, we coded the following (see Table 2 for examples):

Table 2: Examples of coding scheme
(Coded utterance in italics)

Code	Utterance
Hypothesis	You'd think [the number of reclassifications] would go up for condition C, but it didn't…*So maybe the subjects are having a better memory of the ones they've already done* (Psychology 2)
Data collection	*Do you think it's worth getting some more time, just to do an offset plane, or offset velocity?* (Astronomy)
Information extraction	*Well, that's a really clean neuron,, uh it goes down and up and away from the edges* (Neuroscience)
Consult colleague	*I'm gonna have to discuss it with ah, Robbie when he gets back.* (CFD submarine)
Tie-in with theory	*OK, so how do these Fourier modes work?* (CFD laser 1)
Analogy (general)	*Think of this as a spiral arm* (Astronomy)
Analogy (alignment)	And, if I've got a scaling problem, then it should show up here too, *but it doesn't show up here* (CFD submarine)
Conceptual simulation	*In a perfect sort of spider diagram, if you looked at the velocity contours without any sort of streaming motions, no, what I'm trying to say is, um, in the absence of streaming motions, you'd probably expect these lines here to go all the way across, you know, the ring* (Astronomy)

Hypotheses All statements that attempted to explain or account for a phenomenon identified in the data were coded as hypotheses. After a hypothesis, utterances that pertain to (elaborate) that hypothesis were identified. Such utterances constitute further investigation of the hypothesis and may be support or oppose the hypothesis. All subsequent utterances pertaining to a hypothesis were coded as follows:

Data collection Utterances in which the scientist proposed to collect more data were coded as data collection strategies. These include statements that propose an experiment, plans to run such an experiment, or plans to collect additional data for an experiment that has already been run (e.g., increasing the sample size or making some other adjustment) or to collect more observational data. Data collection strategies also include plans to build and run computational models.

Information extraction Statements that "read off" data from the visible display (i.e., extract information) were coded as information extraction (Trafton et al, in press). In addition, we coded as information extraction strategies statements that refer to looking at data in a different way (e.g., replotting the data or displaying it in a different visualization), to "tweaking" data (by transformation, removing outliers, etc.), or to looking at data that is not currently on view but that is available.

Consult a colleague Utterances that refer to showing the data to or asking the opinion of a co-worker or other expert were coded as consulting a colleague.

Tie-in with theory We expected that expert scientists with a vast array of domain knowledge stored in memory were likely to apply that theoretical domain knowledge to their hypotheses. We coded as "tie-in with theory" utterances that refer to theoretical underpinnings of the data.

Analogy/Alignment Although different theories of analogy specify different processes by which the mapping between source and target occurs (Gentner, 1983; Holyoak, 1985), all theories share these elements: source, target, and a process of mapping or alignment. During alignment, the relevant parts of the source are "applied" to the target. It is thus during this phase that inferencing occurs, and hence we expected that scientific reasoning would occur during this part of the analogical process.

We coded analogies using the definition and coding scheme developed by Dunbar (1997). According to this scheme, analogy is coded when a scientist either refers to another base of knowledge to explain a concept or uses another base of knowledge to modify a concept. Analogies were coded at both a "general" level (e.g., "The atom is like the solar system") and at the level of the actual mapping or alignment. Statements of similarity (i.e., "X is like Y") were *not* considered analogies; they do not provide explanations nor result in mapping features from the source to the target.

Conceptual Simulations Recall that a conceptual simulation is a mentally constructed model of a phenomenon or data representation. The initial representation may be grounded in memory (e.g., theoretical knowledge of the phenomenon) or in a mental modification of the displayed image. The key feature of a conceptual simulation is that it involves a simulation "run" that alters the representation, such that there is a change of state.

To code conceptual simulations, we adapted Trafton's spatial transformation framework (Trafton & Trickett, 2001b; Trafton, Trickett, & Mintz, in press;). We conducted

a spatial transformation analysis to determine for each on-task utterance whether the speaker was extracting information from the display ("read-off") and which mental operations, if any, were applied to a representation. Some possibilities include rotation, modification, moving an image, creating a mental representation, animating features, and comparison. Conceptual simulations may be defined formally as a specific sequence of spatial transformations:

1. Create representation: The scientist creates a mental representation that is not the same as the currently displayed representation. This representation creation may occur via the display (it modifies the display), via theory, (a theoretical construct); or via memory (the scientist recalls a previously viewed representation).
2. Simulation Run: The scientist builds on the created representation by spatial transformation (e.g., extend, add, delete) such that its state is changed.

Note that these codes are not mutually exclusive, and that the created representation and explicit run can occur in the same utterance. Approximately 20% of the data has been coded for conceptual simulations by 2 independent coders, and initial inter-rater reliability was greater than 90%.

Results

Eight *in vivo* datasets, comprising 330 minutes of relevant protocol and 3508 on-task utterances were analyzed. We coded 68 hypotheses, an average of approximately 1 hypothesis every 5 minutes. 57 hypotheses (84%) were elaborated; that is, the scientist made some follow-up utterance(s) that further explored the hypothesis.

How did the scientists evaluate the hypotheses?

We identified and counted the type of utterance following each hypothesis. Table 3 summarizes this count. Counts were performed in the following manner: Each individual instance of information extraction was included in the count. For example, the sequence "If I look at the average of that, it's a nice clean spike" (utterance 1) "and I can look at the standard deviation around that and it's pretty tight right in the middle where it needs to be" (utterance 2) was coded as two instances of information extraction. Each utterance identifies a different piece of information extracted (average, standard deviation). In all other cases, the count was based on the number of instances of the coded phenomenon. For example, the sequence "In a perfect sort of spider diagram" (utterance 1) "if you looked at the velocity contours without any sort of streaming motions, (utterance 2) "no, what I'm trying to say is, um, in the absence of streaming motions," (utterance 3) "you probably would expect these lines here [gestures] to go straight across, you know, the ring" (utterance 4) was coded as one conceptual simulation because each utterance contributed to, but did not constitute, one conceptual simulation.

As Table 3 shows, the most frequent strategy used for evaluating hypotheses was information extraction. This result is unsurprising, in that the scientists' task was to examine and analyze the data; one would therefore expect them to devote a significant amount of time to extracting information directly from the data itself. Similarly, the second most frequent strategy, tie-in with theory, might also be predicted from an understanding of the general procedures of science. These scientists have significant expertise and knowledge of the theories relevant to their domains, and one would expect them to consider new data in the context of current theoretical understanding of the domain. One might also expect data collection strategies (which include plans to design or conduct experiments) to occur frequently; however, these were one of the *least* frequent strategies used by these scientists.

Table 3: Frequency of hypothesis-evaluation strategies

Strategy	Frequency
Information extraction	268
Tie-in with theory	36
Conceptual simulation	34
Analogy/Alignment	30
Data collection	3
Consult a colleague	1

The use of analogy is also of interest. Of the 30 uses of the analogy/alignment strategy, only one consisted of a "general" analogy. The remaining 29 were alignments in which the mapping between source and target actually took place. This result is consistent with findings of other studies in which analogy use has been found to be more "local" than "global" (Dunbar, 1997; Saner & Schunn, 1999). The use of alignment is discussed in more detail below.

Of particular interest is the relative frequency of the conceptual simulation strategy. Specifically, this strategy was linked with the alignment strategy in a sequence that took the form of conceptual simulation followed by alignment. There were 34 conceptual simulations and 29 alignments; out of these, there were 27 Conceptual Simulation —> Alignment sequences. Thus most (79%) of the conceptual simulations were immediately followed by an alignment, and most (93%) of the alignments immediately followed a conceptual simulation.

The frequency of the Conceptual Simulation—> Alignment sequence suggests a tight coupling between the two strategies. It appears that the scientists used conceptual simulation to build a "mental model" of the data, based on assumption that the hypothesis under evaluation was true. The scientists used the data on display and their domain knowledge to investigate the implications of the hypothesis, by dynamically constructing a mental simulation of a series of processes. The result of this conceptual simulation was an inspectable mental model that was used as the source of comparison with the actual data in the alignment process. To the extent that the two models aligned, the hypothesis was supported; if there were relevant differences between the models, the hypothesis would be rejected. Figure 2 illustrates this process of model-building and alignment.

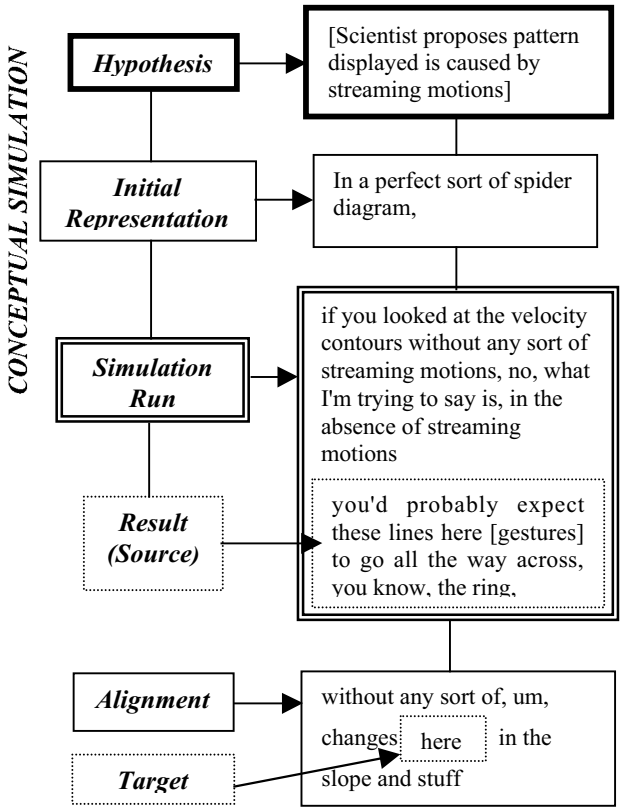


Figure 2: Conceptual simulation as source of comparison in alignment process

Why were conceptual simulations used?

There were 57 elaborated hypotheses in these datasets, and 34 conceptual simulations. The high frequency with which conceptual simulation was used as an evaluation strategy indicates that its use is important and significant. Under what circumstances did the scientists use this strategy? Conceptual simulations were used across a variety of criteria: in both group and individual settings, when the data consisted of either images or numerical tables, in exploratory and confirmatory analysis sessions, and across a variety of domains. It seems, therefore, less likely that conceptual simulations were motivated by characteristics of the data than by some characteristic of the task.

An examination of the structure of a conceptual simulation reveals that its dynamic nature allows an understanding of the *processes* involved in constructing the revised mental representation of the relevant phenomenon. Understanding process may be particularly important when there is significant uncertainty. For example, a poorly understood phenomenon is likely to evoke more investigation than one that is well understood (Trickett, Trafton, & Schunn, 2000). Thus we conjectured that the use of conceptual simulation, with its associated construction of underlying process, was associated with attempts to account for a phenomenon that violated the scientists' expectations.

In order to investigate this possibility, the hypotheses in this dataset were categorized into those that attempted to account for some expectation that wasn't met, and those that pertained to some expected phenomenon. The coding criteria for this categorization were adapted from Trickett et al., 2000. In some cases, the scientists made explicit verbal reference to the fact that something was expected or unexpected. If there was no explicit reference, domain knowledge was used to determine whether a phenomenon was expected or not. A phenomenon might be associated with (i.e., identified as similar or dissimilar to) another phenomenon that had already been established as expected or not, or the scientist might question a phenomenon, thus implying that it was not what was expected. This coding scheme was applied by two independent coders to a subset of the data (the entire astronomy protocol), and agreement between those coders was 87%. Table 4 provides examples.

Table 4: Examples of expectation-violation hypotheses (hypotheses in italics)

Domain	Utterance
CFD (submarine)	Computational model does not agree with the experiments in the least…*It could be that the turbulence is all screwed up too.*
Astronomy	That, that's odd…Why isn't there star formation going on there?…*It may be because of the large velocity dispersion.*

After we coded the hypotheses as associated with expectation violation or confirmation, we counted the use of conceptual simulation and information extraction strategies to evaluate each type of hypothesis. Note that the purpose of the analysis was to determine the circumstances under which each strategy was used, not the frequency with which the strategy followed a hypothesis; thus, only the first instance of each strategy use was counted. We performed a *phi* coefficient association measure. The correlation between hypothesis type and conceptual simulation was significant, $r_\phi = .487$, $p < .01$. There was no correlation between hypothesis type and information extraction, $r = .006$. Table 5 summarizes the results of this analysis.

Table 5: Strategy use and hypothesis type

	Violate Expectation	Confirm Expectation
Conceptual Simulation	22	3
Information Extraction	27	20

General Discussion and Conclusion

The protocol data discussed above have provided a rich dataset by which to investigate the on-line thinking of working scientists analyzing data. The scientists develop hypotheses to account for the data and then evaluate those hypotheses in light of theoretical knowledge and the data itself. In contrast to results of laboratory studies of scientific reasoning, the analyses presented above reveal that the scientists *rarely* chose to evaluate hypotheses by

experimentation (including planning experiments). They frequently used a strategy of conceptual simulation followed by alignment. In particular, they used the conceptual simulation-alignment strategy most often to evaluate a hypothesis about something that violated their expectations.

Conceptual simulation is a process of mental model-building and manipulation that results in a revised mental model, or "Qualitative Mental Model" (QMM) (Trafton et al., 2000). This QMM serves as the source of an analogy that allowed the scientists to compare the QMM with the observed data and from there to evaluate the scientist's current hypothesis. Insofar as the QMM matched the data, the scientist found evidence for the hypothesis; in the absence of a match, the scientist needed to revise the hypothesis. The alignment between source (QMM) and target (data) occurred as a series of mental processes, which amount to a recreation of the *processes* that underlie the external manifestation of the phenomenon of interest.

Acknowledgments

This research was supported in part by grants N00014-00-WX-20844 and N00014-00-WX-4002 to the 2nd author. We thank Christian D. Schunn for comments on this research.

References

Clement, J. (1988). Observed methods for generating analogies in scientific problem solving. *Cognitive Science, 12*(4), 563-586.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17*(3), 397-434.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward & S. M. Smith (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 461-493). Washington, DC, USA: APA

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* (2nd ed.). Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure Mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, Pl.& Forbus, K. D. (1997). Analogical resoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences, 6*(1), 3-40.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: LMP, 18*(5), 1084-1102.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation (Vol. 19)* (pp. 59-87). New York: Academic Press.

Ippolito, M. F., & Tweney, R. D. (1995). The inception of insight. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 433-462). Cambridge, MA, USA: MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. Giere, (Ed.), *Cognitive models of science* (pp. 3-44). Minneapolis, MN: University of Minneapolis Press.

Nersessian, N. J. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 5 - 22). New York: Kluwer Academic/Plenum Publishers.

Okada, T., & Shimokido, T. (2001). The role of hypothesis formation in a community of psychology. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for Science: Implications from everyday, classroom, and professional settings* . Mahwah, NJ: Erlbaum.

Popper, K. R. (1956). *The logic of scientific discovery (rev. ed)*. New York: Basic Books.

Qin, Y, & Simon, H. A. (1990). Imagery and problem-solving. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society.* Pp. 646-653.

Saner, L., & Schunn, C. D. (1999). Analogies out of the blue: When history seems to retell itself. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society.*

Schraagen, J. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17*(2), 285-309.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), 337-370.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: Extracting and generating information from complex visualizations. *International Journal of Human Computer Studies, 53*(5), 827-850.

Trafton, J. G. & Trickett, S. B. (2001a). Note-taking for self-explanation and problem-solving. *Human-Computer Interaction, 16*(1), 1-38.

Trafton, J. G. & Trickett, S. B. (2001b). A new model of graph and visualization use. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (in press). Connecting internal and external images: Spatial transformations of scientific visualizations. *Foundations of Science.*

Trickett, S. B., Trafton, J. G., & Raymond, P. D. (1998). *Exploration in the experiment space: The relationship between systematicity and performance.* In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*.

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2000). Blobs, dipsy-doodles and other funky things: Framework anomalies in exploratory data analysis, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

Ueda, K. (1997). Actual use of analogy in remarkable scientific discovery. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society.*

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*(1), 75-100.

Goal Specificity and the Generality of Schema Acquisition

David L. Trumpower (dtrumpow@unm.edu)
Timothy E. Goldsmith (gold@unm.edu)
Maureen Below (mollybe37@msn.com)
University of New Mexico
Department of Psychology, Logan Hall
Albuquerque, NM 87131 USA

Abstract

Fourteen statistics novices were asked to solve three statistics word problems under standard (SGS) or reduced (RGS) goal specificity. Later, they were asked to solve both structurally identical and structurally different transfer problems, and their structural knowledge of the domain was assessed. Results indicate that participants in the RGS condition performed better on the structurally different transfer problems and had acquired structural knowledge more similar to that of a domain expert. These results extend previous work in showing that the schematic knowledge acquired under reduced goal specificity training is more general than previously realized. The goal specificity effect is discussed in terms of the attentional focus required to solve RGS and SGS problems.

Introduction

Most theorists agree that schemas form the basis for problem solving expertise. Schemas are typically described as knowledge structures that represent generalized concepts, and are comprised of facts and procedures as well as the interrelationships among those facts and procedures. With respect to problem solving, it is generally accepted that schemas allow: (1) problems to be classified according to the general principles required for their solution (Chi, Feltovich, & Glaser, 1981), (2) solution planning (Priest, & Lindsay, 1992), and (3) use of forward-chained solutions (Koedinger, & Anderson, 1990), all of which are hallmarks of expertise. Thus, an important issue for cognitive scientists and educators alike is to understand how schemas are learned.

Cognitive Load Theory (CLT) has been advanced to describe the relationship between problem solving and learning (Sweller, & Levine, 1982; Sweller, 1988). CLT posits that acquisition of schematic knowledge during problem solving is not automatic; rather, it requires a certain amount of cognitive resources. Therefore, if a problem solving task or strategy demands a great deal of cognitive resources then learning will be impaired relative to a task or strategy that carries a low cognitive load.

CLT has been used to explain the finding that reducing the specificity of goals enhances problem solving performance, otherwise know as the goal specificity effect. The goal specificity effect has been shown in maze learning (Sweller, & Levine, 1982), kinematics (Sweller, Mawer, & Ward, 1983), geometry (Ayres, 1993; Sweller, Mawer, & Ward, 1983), trigonometry (Owen, & Sweller, 1985; Sweller, 1988), and several more complex, dynamic tasks (Miller, Lehman, & Koedinger, 1999; Vollmeyer, Burns, & Holyoak, 1996). According to CLT, problems with standard goal specificity (SGS), in which problem solvers are given values for several variables and asked to solve for the value of a specific unknown variable, encourages use of a means-ends strategy. Under a means-ends strategy, problem solvers' attention is focused on reducing the difference between the current problem state and the goal. Moves are guided by the goal state, which requires solvers to keep in memory the goal, any subgoals, and the current problem state. Because this task is cognitively demanding, it detracts from the learning of relations that are relevant for schema acquisition. Reduced goal specificity (RGS) problems, in which problem solvers are asked to solve for the value of as many unknown variables as possible rather than the value of a *specific* unknown variable, eliminate the possibility of a means-ends strategy. Instead, they require a forward-working strategy where moves are generated solely by the current problem state. Because this strategy is less cognitively demanding (see Sweller, 1988), resources are available for learning the relations relevant to schema acquisition, namely, relations between the appropriate operators and problem states.

According to CLT, training with RGS problems is more likely to lead to schema acquisition than training with SGS problems, where schemas are defined as knowledge of problem states and their associated operators. However, this definition of a schema is limited in that it is only applicable to problems with similar structure as those encountered during training (i.e., problems that share, at least some of, the same problem states as the training problems). We will call this the *limited schema view*. Actually, it is difficult to distinguish this view from one that simply postulates the storage of exemplar solutions. If one remembers previous problem solutions, they then have knowledge of problem states and their associated moves/operators…the same information contained in limited schemas. Under this *exemplar view*, the goal specificity effect can be explained by the notion that RGS solutions are easier to remember than SGS solutions (since they require less cognitive load to perform, more resources are available to store them), and they are forward-working. A third alternative is that schemas are acquired under RGS training and that they are

more general than previously believed. We will refer to this possibility as the *general schema view*.

Most of the previous studies investigating the goal specificity effect cannot distinguish among these views, because they have predominantly looked at transfer performance on problems that were structural identical to training problems. For example, Sweller, et al. (1983) showed that novices who practiced with RGS kinematics and geometry problems were more likely to work forward on structurally identical test problems than those who practiced with SGS problems. Although consistent with the idea that RGS participants had acquired schemas (either limited or more general), this result is also compatible with the exemplar view. Since novices tend to use means-ends analysis on standard problems, the solutions to SGS practice problems will be backward-chained, whereas since RGS problems eliminate the possibility of using a means-ends strategy, the solutions to RGS practice problems will be forward-chained. Applying these stored exemplar solutions to test problems would result in forward solutions for RGS participants and backward solutions for SGS participants. Schematic knowledge is not required to account for this finding.

If we assume that the greater cognitive load associated with SGS problems interferes with storage of exemplar solutions, then an exemplar view can also account for the findings that SGS training leads to more errors on isomorphic transfer test problems (Owen, & Sweller, 1985), fewer practice problems accurately recalled (Sweller, 1988), and other related findings.

Furthermore, none of the results mentioned above can distinguish between the limited and general schema views, because both limited and general schemas would apply equally well to problems that are structurally the same as the problems from which the schemas were generated. Structurally different transfer problems, though, would help make the distinction. Limited schemas, comprised of relations between previously encountered problem states and associated operators, would not apply to structurally different problems that have different problem states and different solutions. Exemplar solutions of training problems would not apply either. General schemas that are based on abstract principles, though, would apply to structurally different problems, so long as they could be solved with the same general principle. One finding that may favor the general schema view comes from Owen and Sweller (1985). They trained participants to solve trigonometry problems under either SGS or RGS conditions. Training problems gave values for one side and one angle in a right triangle, and participants were asked to solve for either a specific side of an adjacent triangle, or to solve for the values of as many sides as possible, using the trigonometric ratios sine, cosine, and tangent. Later, performance was tested on structurally identical transfer problems, for which RGS participants showed an advantage. They also tested performance on a diagram construction task in which participants were given values for two sides of a right triangle and were asked to draw the triangle, labeling the values for all three sides. Due to the fact that RGS participants fared better on this diagram construction task as well, the authors concluded that mathematical schema acquisition involves learning mathematical principles, where mathematical principles seem more akin to general than limited schemas. Unfortunately, Owen and Sweller (1985) did not control for the number of sides solved for during training. Because the RGS condition tended to solve significantly more sides during training, any differences upon testing could be attributed to amount of practice rather than goal specificity per se. In the present study, we will use transfer problems that are structurally different from training problems, while also controlling for amount of practice.

Another issue with CLT that remains largely untested is the description of the processes used to account for the effect of goal specificity on schema acquisition. Although Sweller (1988) constructed computational models of SGS and RGS problem solving to show that solving SGS problems do indeed require more cognitive resources, it is nonetheless possible that the functional difference between SGS and RSG problems is solely where attention is focused when solving such problems, and does not depend on the amount of resources available to encode problem information during that time. That is, SGS training might produce just as much learning as RGS training, but if attention is focused in the wrong places, then SGS training will result in erroneous learning. In order to examine this idea, we will employ a structural knowledge measure that is measured independent of problem solving performance, and that allows relatively specific questions about the process of schema acquisition to be tested.

Structural knowledge refers to knowledge of the interrelationships among domain concepts, and is well correlated with domain expertise (for a review, see Goldsmith, Johnson, & Acton, 1991). As such, it is likely to be at least a subset of the knowledge contained in schemas. Trumpower (2000) has shown how network representations of structural knowledge can be used to assess schema acquisition. Briefly, the process involves comparing participants' knowledge networks with those of domain experts. Figure 1 displays a knowledge network of the statistics concepts used in the present study, derived from two statistics experts.

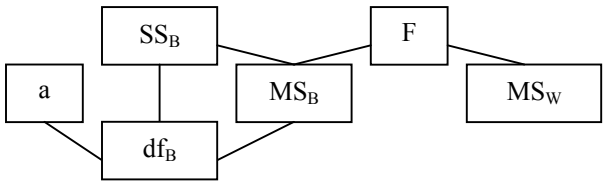


Figure 1: Expert knowledge network

According to an attentional focus explanation of the goal specificity effect, RGS training allows learning of the relationships between problem states (i.e., the subset of

variables that are known at a given time) and appropriate operators (i.e., the equation that can be used at that same time to solve for an unknown). Thus, we might predict that RGS training will result in learning of the relations among concepts contained in the equations used to solve RGS problems, since the equations are the operators and contain the currently known variables. By inspecting the three equations needed to solve training problems in the current study (listed below in the Problem Domain & Materials section), we see that relations among concepts in those equations correspond almost perfectly with the pattern of links in the expert network shown above. Therefore, we predict that participants undergoing RGS training will acquire knowledge structures that look very similar to the expert network shown above.

For SGS training, an attentional focus explanation says that attention is directed toward the goal and reducing differences between current states and the goal, at the expense of noticing the local relationships described above. Therefore, we predict that SGS training will result in making associations between all problem states or known variables and the goal (e.g., links between a-SS_B, df_B-SS_B, MS_B-SS_B, MS_W-SS_B, F-SS_B), but a failure to notice the relevant relations between non-goal concepts (e.g., links between a-df_B, df_B -MS_B, F- MS_B, F- MS_W).

To summarize, the current study addresses three questions: (1) Does goal specificity have its effects primarily on storage of exemplar solutions or schema acquisition?, (2) If the effects are on schema acquisition, then how general are the acquired schemas?, and (3) Can the observed effects be better accounted for by the processes proposed in CLT or an attentional focus explanation? In order to examine these questions, we assessed problem solving performance on transfer problems that were structurally different than training problems, and used a measure of schematic knowledge that is independent of problem solving performance.

Method

Participants

Fourteen undergraduate students enrolled in an Introductory Psychology course at the University of New Mexico participated in this study for partial course credit. None of them had previously completed a college-level statistics course. Half of the participants were randomly assigned to receive training with standard goal specificity problems (SGS), while the other half received training with reduced goal specificity problems (RGS).

Problem Solving Domain & Materials

The problem solving domain used in the present study was one-way analysis of variance (ANOVA). All problems used were relatively simple word problems that could be solved with the following three equations: $df_B=a-1$, $MS_B=SS_B/df_B$, and $F=MS_B/MS_W$, where a = number of groups, df_B = between groups degrees of freedom, MS_B = between groups mean square, SS_B = between groups sum of squares, F = F-ratio, and MS_W = within groups mean square.

All training problems gave values for a, MS_W, and F. Those used in the SGS condition asked to solve for SS_B, while those used in the RGS condition asked to solve for as many unknown values as possible. Notice that in both conditions successful solutions required participants to first solve for df_B and MS_B (in either order), and then solve for SS_B.

Structurally identical transfer problems for both conditions were identical in structure to the training problems received in the SGS condition during training in that they gave values for a, MS_W, and F, and asked to solve for SS_B. Structurally different transfer problems were different in structure from the training problems in that they gave values for different variables, and asked to solve for a different variable. Thus, structurally different transfer problems still required use of the same three equations to solve, but they required that the equations be used in a different order and that the equations be manipulated in a different way than was done during training.

A relatedness rating task was also used in which participants were asked to rate the relatedness of all pairwise combinations of the six statistics terms contained in the equations listed above on a 5-point scale (1="Not at all related", 5="Very related").

Procedure

All participants were tested individually in the presence of an experimenter. Participants were first asked to solve three training problems. During this training period, they were given a Rolodex containing separate note cards containing each of the three equations necessary for solution of the problems, as well as a calculator to perform computations. Participants were allowed five minutes to solve each problem. Within this time, the experimenter would immediately notify the participant if they made a mistake, but would not tell them the nature of the mistake. If the problem was not solved within five minutes, the experimenter would guide them to the solution. After solving a problem, participants went on to the next problem and could not refer back to previous problems.

Upon completion of the third training problem, participants were asked to complete the relatedness rating task, which took approximately five minutes. The equations were not made available to participants during completion of this task.

Next, participants were asked to solve four transfer problems (2 structurally identical, 2 structurally different). Approximately half of the participants in each condition were given the two structurally identical transfer problems first, while the other half were given the two structurally different transfer problems first. Participants were again given the necessary equations, and problem solving proceeded as during training.

Results

Separate one-way ANOVAs were used to compare the SGS and RGS conditions on time to solve each of the training problems, and on time to solve structurally identical and structurally different transfer problems. Additionally, separate one-way ANOVAs were used to compare the number of various kinds of links found in the structural knowledge representations of participants in the SGS and RGS conditions. A .05 significance level was used for all tests.

Training

Participants in the RGS condition solved the first two training problems significantly faster than those in the SGS condition, $F(1,12)=5.03$, $p=.045$ and $F(1,12)=6.89$, $p=.022$, respectively for the first and second training problem. This is consistent with the idea that SGS problems require greater cognitive load, and should therefore require more time to solve. There was no significant difference between the SGS and RGS conditions on time to solve the final training problem, $F(1,12)=1.52$, $p>.10$, suggesting that participants in both conditions had acquired similarly efficient solution procedures by the end of training (see Table 1).

Table 1: Time (in seconds) to solve training, structurally identical transfer (S-I), and structurally different transfer (S-D) problems as a function of training condition.

Problem	SGS	RGS
	Mean (SD)	Mean (SD)
First training	300.00 (0.00)	238.14* (72.94)
Second training	219.00 (58.25)	145.00* (46.59)
Third training	139.00 (50.39)	106.29 (48.85)
S-I transfer	108.29 (31.30)	99.64 (55.44)
S-D transfer	254.93 (45.27)	164.93 (78.52)*

*$p<.05$

Structurally Identical Transfer

There was no difference between the SGS and RGS conditions on average time to solve structurally identical transfer problems, $F<1$ (see Table 1). Apparently, both conditions learned to solve problems of the structure that they were trained on equally well. Although CLT (both the limited schema and general schema views) had predicted better performance from the RGS condition, it is possible that the task was too easy to disrupt learning in the SGS condition. If so, then we would expect no difference on the structurally different transfer problems.

Structurally Different Transfer

Participants in the RGS condition solved the structurally different transfer problems significantly faster than those in the SGS condition, $F(1,12)=6.90$, $p=.022$ (see Table 1). This suggests that although both conditions learned to solve problems structured like the training problems equally well, those in the RGS condition gained qualitatively different knowledge that allowed superior transfer to structurally different problems. This is in contrast to both the limited schema and exemplar views. Schemas comprised of knowledge of problem states encountered during training and associated operators would not apply to the structurally different transfer problems, since these problems involved different problem states. Neither would exemplar solutions acquired during training apply, since the structurally different transfer problems required different solutions. Based on these results, it appears that the schematic knowledge acquired during RGS training is more general than previously thought.

Structural Knowledge

Participant's relatedness ratings were submitted to the Pathfinder scaling algorithm to generate a knowledge network for each (for a review of Pathfinder, see Schvaneveldt, 1990). These networks were then analyzed for the number of: (1) critical links with the training goal, (2) irrelevant links with the training goal, and (3) critical links with non-goal concepts (see Table 2).

There are two critical links with the training goal (SS_B) found in the expert network, one with each of the subgoals, (df_B and MS_B). There was no difference in the mean number of these links possessed by participants in the SGS and RGS conditions, $F<1$, as predicted by an attentional focus explanation.

Four other irrelevant links (i.e., those not found in the expert network) with the training goal are possible. As predicted by the attentional focus explanation, participants in the SGS condition possessed significantly more of these irrelevant links than participants in the RGS condition, $F(1,12)=7.59$, $p=.017$.

Four other critical links, not involving the training goal, are present in the expert network. Of these links, participants in the SGS condition possessed significantly fewer than participants in the RGS condition, $F(1,12)=7.36$, $p=.019$, again consistent with predictions made by the attentional focus explanation.

Taken together, these structural knowledge results are consistent with an attentional focus explanation. Under SGS training, attention is focused on the goal, resulting in both relevant and irrelevant associations being made with the goal, at the expense of other critical schematic associations. RGS training, on the other hand, focuses attention precisely where it is needed for schema acquisition, on the local relations described by the equations.

Table 2: Number of links as a function of training condition.

Link type	SGS	RGS
	Mean (SD)	Mean (SD)
Critical, with goal	1.14 (.69)	1.43 (.53)
Irrelevant, with goal	1.57 (.98)	.29 (.98)*
Critical, with non-goals	1.71 (.76)	3.00 (1.00)*

*$p<.05$

Discussion

The results of the current study support and extend previous studies of the goal specificity effect. Reducing the specificity of training goals led to problem solving advantages. However, the advantage was found on transfer problems that were structurally different than training problems. Thus, it is argued that the schematic knowledge that is more readily acquired under RGS than SGS training is more general than previously considered.

These results are consistent with Owen and Sweller's (1985) contention that schema acquisition involves learning abstract principles. It appears that these principles are not tied to problems of a specific form. With respect to the structural knowledge measure employed in the current study, results suggest that the acquired relational information is not unidirectional. Such findings are important for theories of expertise, since we expect expert-like schemas to be applicable to a wide range of novel problems, not just those encountered in the past. If schemas were limited, then experts would gain no advantage at solving novel problems. The very basis for schema theory is that experts possess not only more knowledge through experience, but also better structured knowledge.

The finding that RGS training leads to the acquisition of general knowledge that can be transferred to structurally different problems than encountered during training is pedagogically important as well. Despite being one of the foremost goals of educators, the difficulty in obtaining transfer to non-isomorphic problems has been well documented (e.g., Gick, & Holyoak, 1983).

The structural knowledge results obtained in the current study are consistent with an attentional focus explanation for the goal specificity effect. It should be pointed out that although the attentional processes invoked by this explanation are described in CLT, they are not dependent upon greater cognitive load being present in the SGS condition. Instead, the present results can be explained by SGS training focusing attention on pedagogically irrelevant relations, and RGS focusing attention towards pedagogically relevant ones. This explanation is similar to one advanced by Miller, et al. (1999). They had participants learn about electrical fields by interacting with a microworld called Electronic Field Hockey (EFH). Participants who practiced moving a puck around the EFH workspace in a no-goal condition performed better on a subsequent test of declarative and procedural knowledge of electrical fields than those who practiced by directing the puck around obstacles and into a specific goal. However, participants who practiced by trying to make the puck follow a well-specifed path denoted by a line leading around obstacles and into a goal, performed almost as well as those in the no-goal condition. Miller, et al. (1999) posit that eliminating the goal worked by requiring interaction with the pedagogically relevant aspects of EFH, just like the specific-path condition. In other words, the specific-path condition directs attention away from the ultimate goal toward a series of more immediate subgoals. By directing attention from more distant goals, it can be focused on local relations involved in solving current subgoals. Similarly, eliminating distant goals altogether allows attention to be focused on immediate local relations, which turn out to be the pedagogically relevant ones.

The results of Miller, et al. (1999) may also be explained by a cognitive load interpretation. If following a specific path shifts attention completely away from the ultimate goal, then the task becomes one of meeting a continuous series of smaller goals. If attention is directed at solving each of the immediate goals (i.e., staying on the path), and if each of these small goals can be solved without use of a means-ends strategy, then the specific-path condition would require no more cognitive resources than the no-goal condition. It may be argued that problem solvers solving no-goal problems do adopt a strategy of setting a series of small goals for themselves that can be solved in a forward-chained manner.

Thus, although neither the present results nor the results of Miller, et al. (1999) require an explanation based on cognitive load, they do not rule it out as a possibility. One way to resolve the issue concerning whether RGS training works due to reduced cognitive load or to a pedagogically relevant focus of attention would be through a dual task paradigm. Problem solvers could be asked to solve RGS problems either while concurrently performing another resource demanding task or not. If the concurrent task interferes with learning in a manner consistent with SGS training, then the cognitive load explanation would be justified.

Overall, this study indicates that eliminating specific goals during training can benefit schema acquisition, and that this advantage is more general than previously considered. Training on problems with non-specific goals allowed better transfer to structurally different problems. It is concluded that non-specific goals allow learning pedagogically relevant, local relations, as opposed to standard problems which interfere with such learning. It is suggested that problems with non-specific goals provide this advantage by focusing attention on relations necessary for schema acquisition.

References

Ayres, P.L. (1993). Why goal-free problems can facilitate learning. *Contemporary Educational Psychology, 18,* 376-381.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Gick, M.L. & Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1-38.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83,* 88-96.

Koedinger, K.R. & Anderson, J.R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*, 511-550.

Miller, C.S., Lehman, J.F., & Koedinger, K.R. (1999). Goals and learning in microworlds. *Cognitive Science, 23*, 305-336.

Owen, E., & Sweller, J. (1985). What do students learn while solving mathematical problems? *Journal of Educational Psychology, 77*, 272-284.

Priest, A.J. & Lindsay, R.O. (1992). New light on novice-expert differences in physics problem solving. *British Journal of Psychology, 83*, 389-405.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257-285.

Sweller, J., & Levine, M. (1982). Effects of goal specificity on means-ends analysis and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 463-474.

Sweller, J., Mawer, R.F., & Ward, M.R. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 112*, 639-661.

Trumpower, D.L. (2000). Schema acquisition and solution strategy in statistics problem solving. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 1061). Maywah, NJ: Lawrence Erlbaum Associates.

Vollmeyer, R., Burns, B.D., & Holyoak, K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75-100.

Precipitate Replications: The Cognitive Analysis of Michael Faraday's Exploration of Gold Precipitates and Colloids

Ryan D. Tweney (tweney@bgnet.bgsu.edu)
Ryan P. Mears (rmears@bgnet.bgsu.edu)
Robert E. Gibby (gibbyro@bgnet.bgsu.edu)
Christiane Spitzmüller (spit@bgnet.bgsu.edu)
Yanlong Sun (suny@bgnet.bgsu.edu)

Department of Psychology, Bowling Green State University
Bowling Green, OH 43403 USA

Abstract

In 1856, Faraday conducted a long series of experiments on the color of gold. We report replications of some of his experiments, permitting an understanding of his response to an important anomaly and the resulting conceptual reorganization of his ideas.

Introduction

In 1856, Michael Faraday (1791-1867) carried out nearly a year's worth of research on the optical properties of thin gold films (Faraday, 1857). In the course of this work, he discovered colloidal gold (the first metallic colloid) and what is now known as the "Faraday-Tyndall Effect," the fact that colloids scatter light. Colloids were an unexpected consequence of his attempts to understand the unusual color properties of thin films of gold. These, in turn, were an important extension of his earlier attempts to understand the interaction of light and matter, and his speculations about the force-centered character of matter (James, 1985; Tweney, 2002).

The present project was initiated by the discovery of over 600 surviving microscope slides and other specimens made by Faraday as part of his research, and now held at the Royal Institution in London (Tweney, 2002). The slides are numbered and indexed in Faraday's Diary and represent nearly the complete set of metallic film specimens used by Faraday in 1856. However, only a few of his colloidal specimens (and none of his precipitates) survived; thus, one goal of the present effort is to restore lost specimens for analysis. Replication is also important even for specimens that still survive -- Faraday often subjected the specimens to destructive and/or damaging manipulations, and these manipulations also need replication. Four general procedures used by Faraday are currently being replicated by our group: (1) Precipitation of gold from solution, (2) "Deflagration" of gold wire, that is, exploding gold wire using sudden surges of current, (3) Producing colloids using reduction by phosphorous, and (4) Producing thin metallic films of gold using reduction by phosphorous.

Here we present our replications of Faraday's precipitates. Besides restoring for analysis certain lost specimens, these allow insight into the "tacit knowledge" implicated in their preparation. More importantly, preparation of our own precipitates allowed analysis of aspects of Faraday's research previously hidden from view, helping to account for an important conceptual change.

In earlier accounts of Faraday's research, our group examined the way in which Faraday experimentally traversed a problem space of hypothesized and real results during his discovery of electromagnetic induction in 1831 (Tweney & Hoffner, 1987). Our analysis suggested that Faraday used a relatively narrow search strategy in the 1831 experiments, one in which potentially disconfirming evidence was initially ignored and only evidence which supported expectations was pursued. In later stages of the research, he made explicit attempts to disconfirm. This "confirm early-disconfirm late" strategy resembled heuristics observed by Klahr in the "Big Trak" studies (Klahr, 2000) and by Dunbar (1995) in an "in vivo" study of laboratory molecular biologists.

Not all aspects of scientific thinking can be characterized as search through multiple problem spaces (Kurz & Tweney, 1998), and this is especially true of Faraday's work. For example, Gooding (1990) replicated Faraday's 1821 discovery of electromagnetic rotations and argued that identification of circular rotatory motions could only have come about by means of an "eye-hand-brain" interaction of a very dynamic character. Rather than "testing hypotheses," Faraday instead had to make the meaning of the otherwise chaotic appearances presented by the experimental apparatus. Similarly, Cavicchi (1997), partly by conducting replications, showed that Faraday's experimentation during his 1845 discovery of diamagnetism proceeded "not by progressively refining

explanations, but by exposing previously unnoticed ambiguities in the phenomena, and uncertainties in interpretation. This exposing deepens the space of [his] confusions" (1997, p. 876). For Cavicchi, such "confusions" (perhaps in response to a surprising result) are a crucial aspect of the pattern-finding involved in discovery. She was further able to show that Faraday's "confusions" resembled those of a student exploring the relationships between bar magnets and iron needles.

Our earlier examination of two of Faraday's papers, one on acoustic vibrations and one on optical illusions of motion, suggested that Faraday's constructive perceptual processes imply a continuum of developing representative explicitness. This continuum began with the perceptual rehearsal of remembered events, proceeded through the construction of "inceptual" representations (that is, representations that abstract away potentially irrelevant features, with an effort to "see" what the results would look like), and finally resulted in a mental model that even included non-perceivable features of phenomena (Tweney, 1992; Ippolito & Tweney, 1995). Again, Faraday appeared to be using an "eye-hand-mind" dynamic in constructing new spaces. Similarly, Nersessian (1999) argued that Faraday and Maxwell used analogies and imagery in a process of generic abstraction, itself important in conceptual generation and change.

Andersen (2002) argued that conceptual reorganization in science often requires a semantic shift, in which exemplars change category. She showed that such change was an important part of the resolution of anomalies in particle physics in the 1930s. The present paper extends the scope of Andersen's argument, by showing that Faraday's work on precipitates was an active source of a crucial conceptual change near the beginning of his 1856 research on gold, perhaps dependent upon the "confusions" engendered by some of the appearances of gold (described below). The dynamics of the reorganization depended upon "epistemic artifacts" constructed by Faraday to serve as active agents in exploration of a new domain.[1]

A distinguishing feature of all of Faraday's research was his determination to produce phenomena of such clarity that his explanations of the phenomena would be transparent to his audiences. "Seeing was believing" in a deep sense for him (see, e.g., Fisher, 2001; Gooding, 1990), and his attention to anomalies in the present case is especially important in explaining the cognitive dynamics of the research. Thus, our replications can potentially contribute to the understanding of the cognitive dynamics of visual representation in scientific research generally (see, e.g., Kulkarni & Simon, 1988; Trickett, et al., 2000).

Why Gold?

Gold films interested Faraday because thin transparent gold films manifest a different color by transmitted light than by reflected light; green, blue, and purple are the most frequent transmitted colors for gold leaf. Faraday thought that gold was therefore a good place to look for insight into the interactions of matter and light. For him, the profoundly interesting question concerned the manner in which such very thin (and apparently continuous) films could so alter light. Although he failed to achieve a definitive answer to this question, he successfully showed that many of the optical properties of metals in general could be produced by the interaction of discrete particles with light. His discovery of gold colloids was an integral part of this argument.

A colloid is a suspension of finely divided particles of a substance held in suspension in a fluid. Colloids differ from solutions in that solutions represent ionized particles of atomic size, carrying an electrical charge. Ions, the particles that form a solution, are much smaller than colloidal particles and affect light in different ways. Faraday's discovery that metals could form colloids was a breakthrough, especially since he also showed that the particles were pure gold, chemically identical to the metal films. Note also that colloids differ from precipitates, which are formed of even larger particles than colloids. If a reducing agent is added to a solution of a gold salt, then metallic gold (Au) is precipitated as a solid. In modern notation;

$$Au^{+++} + 3\,e^{-} \rightarrow Au^{0}\,(solid)$$

Because the particles in a precipitate are far larger than those in a colloid (sometimes being visible to the unaided eye), they settle quite quickly. Colloidal particles (which are far too small to be visible) do not settle because, as Faraday speculated, they are lightly bonded to a "cloud" of ionized particles that repel the surrounding fluid media.

The chemistry of precipitates is more complex than the formula given above suggests, since gold chlorides exist in solution as $[AuCl_4]$- ions and various hydrolyzed ions as well.[2] These more complex species and reactions were not known to Faraday. As we discovered, however, the complexity of the reactions is reflected in a very complex phenomenology when gold salts are actually used. We had expected precipitating a gold salt to be a simple and straightforward process -- a "warm-up" exercise for us (as we thought it may have been for Faraday). In reality, our replications opened a new aspect of Faraday's work on gold, one not visible in the text of the diary itself. In the present paper, we describe our replications of Faraday's precipitates and compare them to colloids and solutions.

[1] The term "epistemic artifact" was used by Tweney (2002) to suggest a blending of the term "cognitive artifact" used by Zhang & Norman (1994) and "epistemic thing" used by Rheinberger (1997).

[2] See Puddephatt, 1978, for this and other details of the reactions of gold.

Faraday's Diary

Faraday's diary is well known because of its relative completeness, an aspect which permits reconstruction of his research practices (e.g., Steinle, 1996). In some cases, however, as in the case of his research on gold, much of the diary is hard to interpret by itself, since the visual context of Faraday's work is absent. As we show, even with that visual context present (in the form of the surviving specimens), there is more to be learned from the "manual" aspect.

Faraday wrote 1160 numbered entries on his gold research, roughly 300 manuscript pages dated from the 2nd of February, 1856 to the 20th of December (Martin, 1936). The distribution of entries (Figure 1) is roughly bimodal, the greatest density of entries occurring at the beginning of the series and toward the end. The first 47 entries (in the first peak of the distribution) are summaries of previous notes. They also include several dozen entries in which Faraday speculates on possible experiments, much as he had earlier kept an "idea book" to record possible studies (Tweney & Gooding, 1991). Faraday's (1857) published paper on the topic was submitted on November 15, 1856 and read before the Royal Society on February 15, 1857. Indeed, the character of the entries in the second peak suggests that he was "mopping up" prior to ending the research -- conducting some necessary control experiments, trying again to resolve some inconsistencies, replicating key preparations, and so on. His work with precipitates occurs near the beginning of the series, on February 5, and appears to record the first laboratory work on gold conducted in his own laboratory (earlier entries describe gold film preparations made at the home of a friend, Warren De La Rue; see Tweney, 2002). Thus, one question is why precipitates constituted the first task undertaken by Faraday.

It has been suggested (e.g., Williams, 1965) that Faraday's work on gold in 1856 manifests his "declining powers" (whether due to aging alone or to the effects of the many toxic exposures he was subjected to over the years). This judgement may stem, in part, from the seeming aimlessness of the precipitation experiments, especially since these occur at the beginning of the first burst of activity. Since the precipitation reaction of gold was long-familiar by 1856, Faraday could learn nothing new here and the text of the Diary alone does not indicate why he initiated his gold research with what seems like a rather prosaic procedure As we show, however, the experiments with precipitates were far from trivial -- by conducting the replications, we were able to detect a "confusion" that served a heuristic role in the important step of arguing that the colors of gold are due to particles interacting with light.

Method

Faraday is vague about exactly how he prepared the precipitates used in his research. In the Diary he indicated only that he "Prepared a standard weak solution of Gold" and a "standard solution of proto sulphate of Iron … consist[ing] of 1 vol. saturated solution at 54° F. plus 2 vols. Water, and a little sulphuric acid to keep all in solution during the changes" (Entry #s 14291 & 14292, 5 Feby. 1856). "Proto sulphate of iron" is "Ferrous Sulfate" in modern terms, and the fact that it was saturated allowed reproduction of the exact substance used by Faraday. But no clue is offered about the "standard weak solution of Gold." Thus, its concentration is unknown and, more importantly, because of the complex chemistry of gold salts and the solution processes by which they dissolve, several possibilities had to be explored for how to prepare the precipitates.

Today, "Gold Chloride" is typically sold in one of two forms; as "Gold (III)," that is, as gold in the valence state +3, in the form of "Tetrachloroauric Acid" ($HAuCl_4$), a yellow crystalline substance, or as "Gold (I) Chloride" (AuCl; valence state +1), in the form of yellowish-white crystals. Each was tried in turn as the basis for gold ion in solution, but neither proved satisfactory, in part because each is unstable. Further,

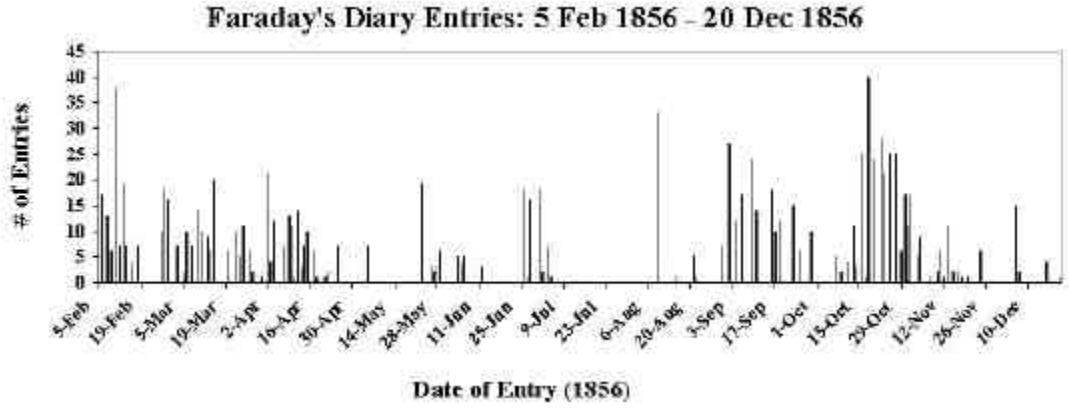


Figure 1

dissolving each in water is problematic; each leaves a precipitated deposit. In the case of tetrachloroauric acid, this is probably pure gold, an expected product when the substance hydrolyzes, but (barring extensive analytic procedures) we were unable to determine if the deposit was the expected gold or simply an undissolved portion of the original crystals. As a result, we could not be sure of the strength of the solutions we were preparing. "Gold (I) Chloride" (AuCl), in the presence of water, oxidizes to the III valence state. This substance seemed to dissolve readily, but again with traces of a deposit. Here again, knowing the strength of the resultant solutions was difficult. Accordingly, we decided to begin with pure gold, dissolving it in such a way that we could be sure of at least the quantity of gold in the solution.

Pure gold wire (0.025" diameter, 99.99%) was obtained from a vacuum technology supply house.[3] Aqua Regia, a 3:1 combination of hydrochloric acid and nitric acid, was used to dissolve a 3.5 cm length of the wire, weighing 217 mg, to create gold chloride solution.[4] Fifteen minutes after addition, the gold wire completely dissolved in the acid. The solution was then boiled in order to remove the hydrochloric and nitric acid. Water was added as needed to keep a constant volume of about 10 cc. The solution was boiled until the odor of the acids and the nitric oxide byproduct (all of which are pungent in even slight quantities) was no longer present.

To produce a gold colloid we used a modern method, the reduction of gold chloride solution by citrate ion (producing such colloids using Faraday's methods is part of an ongoing study and will be reported later). Gold (III) chloride (i.e., tetrachloroauric acid) (3 mg) combined with 10 ml of water produced the gold chloride solution. Gold and excess gold chloride remained at the bottom of the container. Ten mg of sodium citrate, a source of citrate ion, was dissolved in 10 ml of water. The gold chloride solution was agitated with a magnetic stirrer and heated to boiling temperature. Citrate solution (0.015 ml) was slowly added to the gold chloride and reacted immediately, producing a very pale slate-blue solution. Over the course of forty seconds the color of the solution evolved from slate-blue to amethyst to ruby red. When cooled, the product proved stable over many months. Although we cannot be sure of the quantities of gold that are actually in colloidal form, the properties of our colloid were exactly as described by Faraday.

[3] Gold wire, because of its malleability and lack of reactivity, is used as a component in high-vacuum O-ring seals. Jewelry gold, unfortunately, is always alloyed with other metals.

[4] Unlike other metals, gold will not dissolve in hydrochloric acid (HCl) alone because it requires both an oxidant and a ligand donor (Cl^-, in this case). When Aqua Regia is used, the result is $AuCl_4^-$ in solution and gaseous nitric oxide (NO).

The reduction of gold chloride solution by ferrous sulfate solution was used to form gold precipitates. A saturated solution was prepared by dissolving crystalline ferrous sulfate ($FeSO_4$) in heated water. When cooled, three drops of ferrous sulfate solution were added to 5 ml of gold chloride solution. No immediate reaction was apparent, but on the following day a yellow-orange residue of metallic gold had settled at the bottom of the experiment tube and could be re-dispersed by shaking.

Results

The three preparations showed the expected appearance in ambient (room) light; the solution was a clear, deep yellow fluid, the colloid was a clear ruby-red fluid, and the precipitate, when shaken, was a cloudy yellow-gold suspension in which individual particles could be seen in motion, and in which occasional glints of bright metallic gold could be seen. Except for the overall color, the solution and the colloid appeared to be very similar, while the shaken precipitate had a very different appearance.

The relative similarity of the three changed, however, when directional lighting was passed through the fluids. The principal results are summarized in Figure 2, which shows the effect of a parallel beam of light produced by a fiber-optic illuminator (entering from the left) on our prepared gold colloid, a solution of gold chloride, and the precipitated gold preparation, respectively. The overall colors of the preparations are not shown here, but can be viewed at http://personal.bgsu.edu/~tweney. The precipitate was shaken just before the photograph was taken. Note that the colloid shows a bright "Faraday-Tyndall Effect," that is, light is scattered to the side, illuminating the path of the beam through the colloid.

Figure 2. Colloid, Solution and Precipitate

The colloid (a ruby-red transparent fluid) tinges the scattered light a faint pink. The solution scatters no light, only some small reflections from the sides of the glass test tube being visible in the photograph. The precipitate scatters light rather more broadly than the colloid, although some of that is an incidental result of the widening of the initially parallel beam of light into a cone, as a result of its passage through the two prior

preparations. The overall color of the scattered light from the precipitate is a yellowish-gold, and individual particles are easily visible. Obviously the colloid and the precipitate resemble each other most closely under these optical conditions, in contrast to the appearances in ambient light. Although there is no record in the diary of Faraday placing these three in one context (as we have done in Figure 2), it is clear that he was attending these differences very carefully -- they constituted part of the basis for his conclusion that the colloids were in fact metallic particles of gold.

Conclusions

The change in apparent similarity of the three kinds of preparations when transmitted light is compared to ambient light (Figure 3) suggests a possible "confusion" (Cavicchi, 1997), and the need for a reorganization of the phenomenological domain of "divisible gold" (as Faraday referred to it in the 1857 paper). This confusion suggests an explanation for why Faraday began with precipitates on February 5, an explanation that corresponds with what Faraday does say in the Diary.

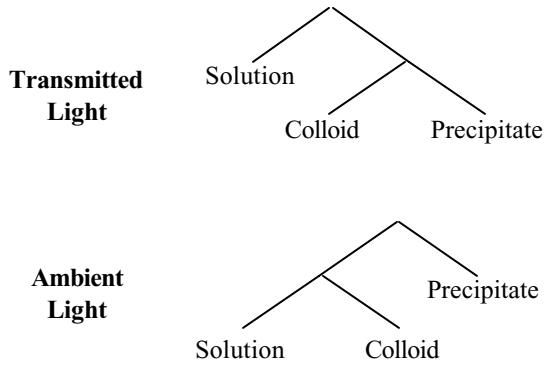


Figure 3. Taxonomy of Similarities

From the Diary, we learn that Faraday had visited De la Rue the week prior to February 5th, and the two had examined some gold leaf through the microscope; Faraday recorded this visit in his first diary entry on gold (#14243, 2 February 1856). On the 2nd, Faraday received some thin gold films prepared by De la Rue, who had used phosphorous to reduce the gold. On the 6th (one day after preparing his precipitates), Faraday established a careful optical method for the examination of precipitates, and recorded that, in the evening, he went to De la Rue's again, and observed how the thin gold films were made. At this point, Faraday noticed something odd; "A very fine red fluid is obtained [from] the mere washing" (Diary, #14321). This, of course, was a colloid, and Faraday saved the fluid, returning to it two weeks later on the 18th (#14437), after his experiments with precipitates and his first examinations of thin films. At that point, he was able to ask; "… the question is, is it [i.e., the gold] in the same state as whilst apparently dissolved in the fluid" (#14437, emphasis in original). It is interesting to note that during this 16 day period he referred to the red fluid using two terms interchangeably, "fluid" and "solution". Only later could he be sure that the red fluid was not a solution, although he must have had the idea very early.

The sequence of his ideas then must be something like the following. He first compares thin films (which he suspects are gold in a continuous state) to the precipitates, which he knows to be discrete particles. Since gold in a continuous state changes appearance in transmitted light and reflected light, he develops an "optical method" for examining precipitates under the same two conditions. Note that, to prepare the precipitates, he must have had before him the clear solution of gold chloride. Then, at De la Rue's, he explicitly notices the clear red solution and this must have suggested a comparison. He knew that the substances used to produce the clear red solution (phosphorous, carbon disulfide, and a gold chloride solution) produced metallic gold. But why then did it look like a solution? Resolving this "confusion" led him to examine the red fluids more closely – and it would be a natural extension to use both transmitted and ambient light, just as he had done with the precipitates. And the transmitted light (as our Figure 2 shows) would make the red fluid look very different from the clear solution. A real anomaly had been found, and a reorganization became necessary; the "red fluid" must be gold in a "divisible state," like the precipitates.

There was, of course, still much to do. More work was needed to explore the new optical effects, to examine other kinds of divisible gold (e.g., that produced by exploding gold wires), to examine other substances, and, most importantly, to explain the differences in the color of light produced by gold in different states (a goal only partially realized). Yet the anomalous appearance of the red fluids at the beginning of the series of experiments provided Faraday with a first important clue to the kind of inquiry he would need to make. Far from constituting a record of "declining powers," the replications allow us to see that Faraday's ability to notice and exploit an anomaly was undiminished in 1856.

Further, our replications revealed that the precipitation experiments are more important than can be discovered by examining the diary records alone, because their role in recognizing the divisible state of colloidal preparations is not evident otherwise. The text of the diary alone does not reveal what was obvious, visually, to Faraday – and was obvious to us only when

present as the result of our own "makings." Only in this fashion could we have noted a conceptual change reminiscent of that seen by Andersen (2002) in her analysis of 20th century particle physicists.

Faraday's gold research in 1856 provided Faraday with mental models based upon new conceptions about the interaction between thin gold films and light. And the differing visual properties of colloids, solutions, and precipitates were a crucial first step, because they showed that the particles of gold had specific optical properties. These in turn led Faraday to reevaluate his previous views of the distinction between continuous and "divisible" matter. Thus, there are similarities between the conceptual reorganizations we observed by replicating Faraday's precipitation experiments, and the larger reorganizations that constituted the outcome of the entire 1856 series of studies. Further replications and text analysis are in progress to extend the reach of this conclusion.

Acknowledgments

This research was partially supported by NSF Award 0100112.

References

Andersen, H. (in press, 2002). The development of family resemblance concepts. In L. Magnani & N.J. Nersessian (Eds.), Model-Based Reasoning: Science, Technology, Values. New York: Kluwer Academic/ Plenum.

Cavicchi, E. (1997). Experimenting with magnetism: Ways of learning of Joann and Faraday. American Journal of Physics, 65, 867-882.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), Mechanisms of insight (pp. 365-396). Cambridge, MA: MIT Press.

Faraday, M. (1857). Experimental relations of gold (and other metals) to light. Philosophical Transactions, 1857, 145-181 (Read Feb. 5, 1857).

Fisher, H.J. (2001). Faraday's Experimental researches in electricity: Guide to a first reading. Santa Fe, NM: Green Lion Press.

Gooding, D. (1990). Experiment and the making of meaning: Human agency in scientific observation and experiment. Dordrecht: Kluwer Academic Publishers.

Ippolito, M.F. & Tweney, R.D. (1995). The inception of insight. In R.J. Sternberg & J.E. Davidson (eds.) The nature of insight. (pp. 433-462) Cambridge, MA: The MIT Press.

James, Frank A.J.L. (1985). "The optical mode of investigation": Light and matter in Faraday's natural philosophy. In D. Gooding & F.A.J.L. James (Eds.), Faraday rediscovered: Essays on the life and work of Michael Faraday, 1791-1867 (pp.137-162). Basingstoke, UK: Macmillan.

Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, MA: MIT Press.

Kulkarni, D. & Simon, H.A. (1988). The processes of scientific discovery: The strategy of experimentation. Cognitive Science, 12, 139-176.

Kurz, E.M. & Tweney, R.D. (1998). The practice of mathematics and science: From calculus to the clothesline problem. In M. Oaksford & N. Chater (Eds.) Rational models of cognition, (pp. 415-438). Oxford: Oxford University Press.

Martin, T., ed., 1936, Faraday's Diary: Being the various philosophical notes of experimental investigation made by Michael Faraday during the years 1820-1862, vol. 7, G. Bell, London.

Nersessian, N.J. (1999). Model based reasoning in conceptual change. In L. Magnani, N.J. Nersessian, & P. Thagard (Eds.), Model-based reasoning in scientific discovery. New York: Kluwer/Plenum.

Puddephatt, R.J. (1978). The chemistry of gold. Amsterdam: Elsevier Scientific.

Rheinberger, H.J. (1997). Toward a history of epistemic things: Synthesizing proteins in the test tube. Stanford, CA: Stanford University Press.

Steinle, F. (1996). Work, Finish, Publish? The formation of the second series of Faraday's "Experimental researches in electricity," Physis, 33, 141-220.

Trickett, S.B., Fu, W.-T., Schunn, C.D., & Trafton, J.G. (2000). From dipsy-doodles to streaming motions: Changes in the representation of visual scientific data. Proceedings of the 22nd Annual Conference of the Cognitive Science Society (pp. 959-964). Mahwah, NJ: Erlbaum.

Tweney, R. D. (1992). Stopping Time: Faraday and the scientific creation of perceptual order. Physis, 29, 149-164.

Tweney, R.D. (in press, 2002). Epistemic artifacts: Michael Faraday's search for the optical effects of gold. In L. Magnani & N.J. Nersessian (Eds.), Model-Based Reasoning: Science, Technology, Values. New York: Kluwer Academic/ Plenum.

Tweney, R.D. & Gooding, D. (Eds.). (1991). Faraday's 1822 "Chemical Notes, Hints, Suggestions, and Objects of Pursuit". Edited with an introduction and notes by R. D. Tweney & D. Gooding. London: The Science Museum & Peter Peregrinus, Ltd.

Tweney, R. D. & Hoffner, C. E. (1987). Understanding the microstructure of science: An example. In Program of the Ninth Annual Conference of the Cognitive Science Society, pp. 677-681. Hillsdale, NJ: Lawrence Erlbaum.

Williams, L.P. (1965). Michael Faraday: A biography. New York: Basic Books.

Zhang, J. & Norman, D.A. (1994). Representations in distributed cognitive tasks. Cognitive Science, 18, 87-122.

Graphically Speaking:
Do Graphics Affect Perspectives in Event Conceptualization?[1]

Ichiro Umata (umata@atr.co.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan
Yasuhiro Katagiri (katagiri@atr.co.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan
Atsushi Shimojima (ashimoji@jaist.ac.jp)
Japan Advanced Institute of Science and Technology;
1–1 Asahi Tatsunokuchi Nomi Ishikawa 923-1292 Japan

Abstract

To see is to believe. A picture is often worth a thousand words in everyday communication settings. A graphical representation actually "talks" in such communication, integrated with some other representation systems like spoken language. Because of its power, however, a graphical representation can affect the way people grasp a target situation it describes. This paper presents an empirical investigation of language usage in graphical communication. Drawing on actual dialogue data, we show that the configuration of graphics affects linguistic expressions of motion when people collaboratively work on a task. This effect of graphics on language usage demonstrates that the configuration of graphics has an influence on perspectives in event conceptualizations.

Introduction

Daily communication is by nature multi-modal, and graphics often serve as strong visual aids for information exchange. People communicate with each other effectively by integrating information from linguistic and graphical sources (Neilson and Lee (1994); Umata, Shimojima, and Katagiri (2000)). People grasp described situations via graphics, taking advantage of their "handiness." Because of this role they play, however, the way people grasp target situations can be affected by graphics.

Recent investigations have demonstrated that the perspectives of spatial descriptions and linguistic expressions show some correspondence. Levinson (1996) observed that some languages make almost exclusive use of absolute coordinates while European languages tend to use egocentric or relative coordinates, reflecting people's strategies for spatial memory and inference. Taylor and Tversky showed that there are three perspectives of spatial descriptions (i.e. *a gaze tour/a route/a survey*) that roughly correspond to the frames of reference distinguished by Levinson in a linguistic method.

It has also been observed that the existence of graphics provides two graphics-based perspectives in addition to those used in only describing a world, and that there is a trade-off between cognitive costs and alignment-failure-robustness (Umata, Katagiri, and Shimojima (2002)). One of those perspectives is called the Observer-to-Graphics Perspective, in which people conceptualize the target events from the viewpoint of the observer relative to the graphics. Suppose one says: "From Baker Street, we'll travel on the Bakerloo Line and then go right at Oxford Circus to Holborn" while holding the map of the London Underground shown in Figure 1. The movement is described from the viewpoint of the observer relative to the map. On the other hand, if one says: "From Baker Street, we'll travel on the Bakerloo Line and then go left at Oxford Circus to Holborn," the speaker "goes into" the map world as an imaginary agent, taking the other perspective called the Protagonist Perspective.

When people see and talk about world situations through graphics, it is quite likely that the features of the graphics affect the way people conceptualize the target situations. In this research, we studied how the availability and configuration of graphics affect language usage in communication and problem-solving. We focused on the influence of graphical representations on the conceptualization of motion events.

Suppose that John and Mary are at the Goodge Street tube station, discussing where to have dinner together. Mary might suggest a place by saying (1) below, but she would not do it by saying (2):

(1) Let's go down to Waterloo Station on the Northern Line and eat at Livebait.

(2) Let's come down to Waterloo Station on the Northern Line and eat at Livebait.

The current position where the two people are located becomes the reference point of the movement in this case, and the movement can only be conceptualized as a movement away from the reference point, and hence the use of "go." Suppose, on the other hand, that John and Mary are discussing their evening plans over the map of the London Underground shown in Figure 1. Mary could use, in this case, either (1) or (2). The availability of the map and the configuration of icons on the map affect the conceptualization of the movement here: the nearness of the Waterloo Station icon from them makes it possible for her to conceptualize the movement, in addition to the previous distal movement conceptualization, as a movement *in the map-world* toward the reference point, their current position. Graphical representation can have an influence on language usage.

The use of "come" in (2) is possible because the map and the graphical objects contained in it are readily avail-

[1]This research was supported in part by the Telecommunications Advancement Organization of Japan.

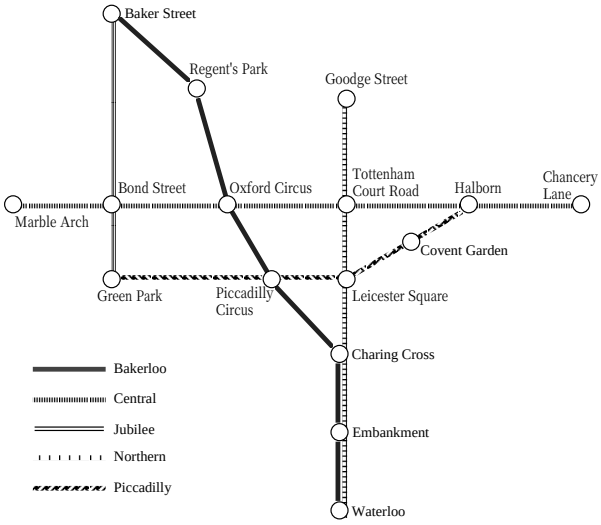


Figure 1: Route Map of London's Underground System

able to the speaker as a resource to formulate messages to be communicated and problems to be reasoned about. The locations and arrangements of objects can be expressed in terms of the relationships between graphical objects and the speaker, as well as those between objects themselves and the speaker. This availability, or the ease of accessibility, of graphical representations should work to amplify our communicative and reasoning capabilities by providing us with a novel set of possibilities for constructing perspectival event conceptualizations.

We investigated the effect of graphical representations on perspectival event conceptualizations through an empirical analysis of the use of motion verbs in actual two-party task-oriented dialogues that make use of diagrams. We first outline, in the next section, the classification of the types of perspectival conceptualizations available in communications that involve graphical representations. We then examine, in the following section, the Japanese dialogue data from our corpus involving a "Missionaries and Cannibals" type puzzle. We found that the configuration of graphics affects the way people grasp the target situations and that the perspectival conceptualizations of the graphical objects are more prominent than those of the real-world objects they represent.

Perspectives in Graphical Communication

In graphical communication, graphical representations work as "windows" through which we can see the target situations the graphics describe. However, they also serve as information processing "sites" because we can take advantage of their handiness. Graphical representations are so deeply ingrained in how we grasp and describe target situations that their existence raises the possibility of setting novel perspectives in conceptualizing events. Two such perspectives were observed in Umata, Katagiri and Shimojima (2002), in addition to two perspectives concerned solely with the target world. There are four possible categories of perspectives on motion events in graphical communication.

(a) Observer-to-World Perspective

> A movement is taken as a movement in the real world and conceptualized from the viewpoint of the observer within the real world. This perspective is concerned solely with the real world.

(b) Agent Perspective

> A movement is taken as a movement in the real world and conceptualized from the viewpoint of the agent of motion. This perspective is concerned solely with the real world.

(c) Observer-to-Graphic Perspective

> A movement is taken as a movement in the map space and conceptualized from the viewpoint of the observer relative to the map. This perspective concerns both the real world and the graphic space and creates a bridge between the two.

(d) Protagonist Perspective

> A movement is conceptualized from the viewpoint of an imaginary agent in a narrative world. In graphical communication situations, a graphic provides the narrative domain for this perspective. The agent can be identified with either the speaker or the listener. This perspective belongs solely to the narrative world.

The first two categories are perspectives concerned solely with the target world, which can also be observed in communication without the use of graphics. When you say "John is coming to my place from Goodge Street" without a map, you are taking the Observer-to-World Perspective. If you are actually driving to somebody's place, you might say "I'm now going to the right-hand side of Piccadilly Circus," taking the Agent Perspective. If you are explaining the way to somebody via a cellular phone, you might say, "Go south and turn left at Leicester Square," taking the Agent Perspective of the person to whom you are talking.

In graphical communication, the latter two perspectives are available in addition to (a) and (b). Examples from the HCRC Map Task Corpus analyzed in Umata, Katagiri and Shimojima (2002) are shown below.

HCRC Map Task Corpus

The examples shown in the following sections are from the HCRC Map Task Corpus. This map task is a cooperative one involving two participants. The two speakers sit opposite one another, and one speaker gives instructions for a route to the other. Each has a map that the other cannot see, and a route is marked on the Instruction Giver's map while no route is marked on that of the Instruction Follower. The speakers are told that their goal is to reproduce the Giver's route on the Follower's map. Their maps are not exactly identical and the speakers are

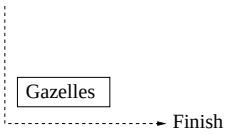


Figure 2: Movement described in (3)

told this explicitly at the beginning of their first session. It is, however, up to them to discover how the two maps differ. The maps describe fictitious areas.

Observer-to-Graphic Perspective

First, consider the following utterances:

(3) (The Giver is showing the Follower the movement shown in Figure 2.)

Giver: Okay? Now you need to drop straight down towards the gazelles.
Follower: Right, coming in at the top of them.
Giver: That's right, and then go round the bottom of the gazelles.
Follower: On the left-hand side?
Giver: And head off to the right-hand side ... So you go under the gazelles

The expression "drop straight down to" would not have been suitable without a map. It describes the motion from the Observer-to-Graphic perspective, making use of the spatial relation on the map. Other spatial expressions ("at the top of," "bottom of," "right-hand side," and "under") also describe the spatial relation of the target situation via the graphical relations. For example, "go round the bottom of the gazelles" does not mean actually going under the gazelles in the target situation. Also, notice that the deictic spatial expression "to the right of" is based on the Observer-to-Graphic Perspective. It would have been "to the left-hand side" if the giver was taking the Protagonist Perspective.

Protagonist Perspective

Now we will examine the examples of the Protagonist Perspective shown in (4):

(4) (The Giver is showing the Follower the movement shown in Figure 3.)

Giver: Then down.
Follower: What do you mean down? Towards the bottom of the paper?
Giver: Uh-huh
Follower: Uh-huh. Do I ... do I go by the collapsed shelter?
Giver: Uh-huh
Follower: Uh-huh
Giver: And then ... so that ... until you've got ...
Follower: The collapsed shelter's on my right?
Giver: Uh-huh

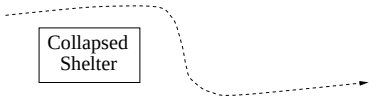


Figure 3: Movement described in (4)

Follower: Right
Giver: And then go round to your left
Follower: My left?
Giver: As you're the wee guy

The spatial expressions "on my right," "to your left" and "my left" are clearly based on the Protagonist Perspective. If the speakers were talking based on the Observer-to-Graphic Perspective, they would have said "on my left," "to your right" and "my right," respectively. The giver confirmed it with the utterance "as you're the wee guy," introducing an imaginary agent explicitly. Thus, we can find many such expressions spoken from the two graphic-based perspectives in communications that make use of graphics.

Perspectives and Alignment of Coordinates

We have observed that the existence of graphics provides two novel perspectives, namely the Observer-to-Graphics Perspective and the Protagonist Perspective. The former requires less cognitive resources because one can grasp motion events as seen on graphics. This perspective works if the coordinates are firmly aligned between speakers. However, misalignment of coordinates leads to serious miscommunication. On the other hand, the Protagonist Perspective is robust against misalignment, because conversation participants "go into the graphics," and the spatial relations between the actual speakers and their graphics are not crucial in this perspective. Such a "trade-off" between description cost and misalignment robustness was actually observed in the HCRC Map Task Corpus (Umata, Katagiri and Shimojima (2002)).

Dialogues Involving a "Missionaries and Cannibals" Type Puzzle

We have observed that the availability of graphics provides a novel set of perspectives based on graphics. Because people "see" the target situations via graphics from those perspectives, it is quite likely that the features of the graphics affect the way people conceptualize the target situations.

The dialogue data analyzed in this section were taken from two collaborative problem solving experiments that involved back-and-forth movement. Another important feature of this task was that it involved two real-world places that subjects were familiar with. We also examined how much effect the Observer-to-World Perspective had on the usage of motion verbs.

Motion Verbs in Corpus

Verbs like *come* and *go* reflect a speaker's reference point, as shown in (1) and (2). *Go* indicates motion to a location that is distinct from the reference point. *Come* indicates motion toward the reference point[2].

The Japanese language also has a pair of motion verbs similar to English *come* and *go*: *kuru* and *iku*. *iku* (go) and *tsurete-iku* (take) indicate motion to a location that is distinct from the reference point. *kuru* (come) and *tsurete-kuru* (bring) indicate motion toward the reference point[3]. *Iku*-type verbs are used more widely than *kuru*-type verbs in the sense that the former expresses movements that are neutral with respect to the reference point locations. There are several verbs that can be classified into these two classes. We examined the usage of these two classes of verbs in the following two experiments.

Data

The data analyzed here was gathered from experiments involving problem solving. In this task, two subjects collaboratively worked on a "Missionaries and Cannibals" type puzzles using a diagram given to them. The structure of the puzzle was basically the same as the original one, except that it involved two actual places that the subjects were familiar with. The time limit was seven minutes.

The subjects recruited from local universities were seated in separate, soundproof rooms and worked together as a pair using a shared virtual whiteboard and a full duplex audio connection. The diagram was shown on their whiteboards, and the subjects could draw and erase freely except that they could not erase the original diagram. All inputs to the screen were by stylus, and any writing or erasing by one participant would appear simultaneously on the partner's screen. A pointing action with the stylus was shown by a cursor on the screen, and the subjects could see what their partner was pointing to. The subjects were video-taped during the task.

The Motorcycle Gang Task

The puzzle was almost the same as the original one, except that we used two actual places and replaced the missionaries and cannibals with two teams of motorcycle gangs. The subjects were told to work out how all members of both gangs could be transported safely. This task involved only two kinds of motion: forward and backward motion between two places. The time limit was seven minutes, including the time they used to read the problem sheet.

[2]Actually, *come* and *go* have more complicated semantics, as shown in Fillmore (1997). The scheme presented here is a rather simplified version, but serves well enough for the present purpose.

[3]There is one clear difference between English and Japanese, though. When a speaker is trying to go to the hearer, s/he will say, "I'll come to you," while *iku* (go) is used rather than *kuru* (come) in Japanese. However, this difference is not relevant here.

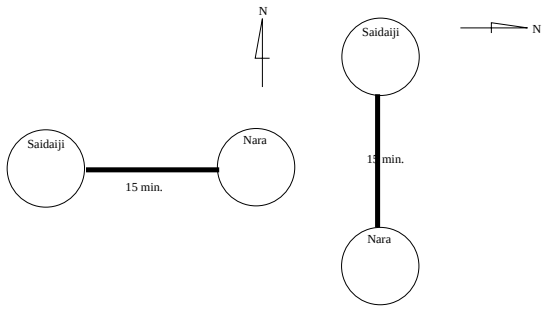


Figure 4: Horizontal and Vertical Diagrams

Experiment 1

The first experiment was conducted to examine the effect of the configurational features of graphics when a perspective based on graphics is taken. The motion in this task was much more simplified, though there was back-and-forth motion that was not in the HCRC Map Task. The problem involved motion between two actual places on a motorbike so that the subject could also directly access the real world. This task has a general direction of motion: all six boys have to move from *Saidaiji* to *Nara*. Each of these locations is about the same distance from the experiment site. The bike was supposed to be able to carry only two people at one time, and someone had to ride back on it. Two kinds of graphical representations were provided as shown in Figure 4. One had a horizontal configuration, in which the two icons of the places are at about the same distance from the subject[4]. The other one had a vertical configuration, which had a variation in the distance from the subject to each place. Each condition had eight pairs of subjects.

The speakers were not supposed to be the ones moving between the two places in this task setting, so the Agent Perspective was not possible. Thus, the possible perspectives were expected to be the Observer-to-World, the Observer-to-Graphic and the Protagonist Perspective. Because the motion is taken as that in the map from the real-world observer in the Observer-to-World Perspective, it is likely that the spatial relation between the speaker and the graphical objects plays an important role. The assumption was that, under that perspective, *kuru(come)*-type verbs would be used for the motion to Nara more frequently in the vertical condition than in the horizontal one, because they would be affected by the nearness of the Nara icon in the diagram to the speaker.

Results of Experiment 1 The distribution of motion verbs was as follows:

Iku-type verbs and *kuru*-type verbs exhibited significantly different distributions depending on the direction of movement in both conditions (horizontal: ${\chi^2}_{(1)} = 93.64, p < 0.01$; vertical (${\chi^2}_{(1)} = 53.98, p < 0.01$). In both the horizontal and vertical conditions, *iku*-type

[4]Note that it is common for maps in Japan to have a direction other than north at the top.

Table 1: Distribution of *iku*-type and *kuru*-type verbs.

		iku-type	*kuru*-type
Horizontal	Saidaiji → Nara	66	2
	Nara → Saidaiji	4	45
Vertical	Saidaiji → Nara	47	18
	Nara → Saidaiji	1	45

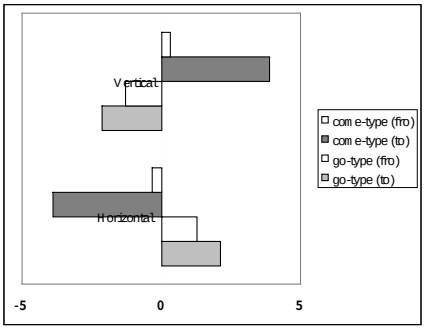


Figure 5: Difference in distribution of motion verbs

verbs were predominantly employed for the movement from Saidaiji to Nara, whereas *kuru*-type verbs were predominantly employed from Nara to Saidaiji. This distributional imbalance indicates that our four perspective types are not sufficient to explain the whole story of graphical communication in problem solving settings, since, for the horizontal condition, movement in each direction could be considered neither toward nor away from the reference point under all four perspectives. Hence one would predict that all movement in the horizontal condition should be expressed with *iku*-type verbs.

We therefore stipulated an additional perspective, the Problem Perspective, to account for this distributional imbalance. As the task for "Missionaries and Cannibals" type problems is to transport all parties collectively from one location to another, the nature of the problem itself induces a general direction of movement from the source location to the goal location. This is equivalent to positing the reference point at the source location, which is Saidaiji in our experimental setting. Under the Problem Perspective each movement is taken as a movement either toward or away from the reference point as posited by the structure of the problem itself, and this perspective is expected to appear in transportation type problems in both concrete and abstract domains.

Comparing the horizontal and the vertical conditions, we notice that *kuru*-type verbs are more frequently used for the movements from Saidaiji to Nara. Figure 5 shows the difference in the distribution of motion verbs between the two conditions. The motion verbs exhibit significantly different distributions between the horizontal and the vertical conditions ($\chi^2_{(3)} = 17.65, p < 0.01$). More concretely, the frequency of the *kuru*-type in Saidaiji to Nara motion is significantly larger in the vertical condition (adjusted residual: horizontal = -3.87, vertical = 3.87).

The Problem Perspective sets a general reference point at Saidaiji, the source of the whole transportation process. It appears, however, that the graphics configuration may be able to modify the reference point settings. The *kuru*-type showed higher frequency for Saidaiji to Nara in the vertical condition than in the horizontal condition. This shows that the spatial relation between the speaker and the graphical objects makes the transition to the Observer-to-Graphic Perspective, and this affects the reference point setting accordingly. The handiness of the graphical representation can cause a switch in perspectives and thus a shift in the reference point.

In contrast to the increase in *kuru*-type verbs in the vertical condition, we notice no increase in *iku*-type verbs for the movements in the opposite direction. The low frequency of the *iku*-type for Nara to Saidaiji suggests that the perspective switch from the Problem Perspective to the Observer-to-Graphic Perspective is preferred when the resulting perspective takes the movement as a toward movement rather than as an away-from movement. This asymmetry also suggests that the Protagonist Perspective was not playing a significant role, as switching to it would have increased the occurrence of *iku*-type verbs here.

Thus, it was shown that the effect of the Problem Perspective defined by the task was the most prominent factor in reference point setting, but that the configuration of a graphical representation often affects this reference point setting.

Experiment 2

The previous experiment showed that the Problem Perspective was the most influential factor, while the configuration of graphics also affects the usage of motion verbs. Now we will look into the effect of the real-world configuration in conversations involving graphics. The setup of Experiment 2 differs from that of our previous experiment as follows:

- Both of the two conditions had a vertical configuration.
- One of the locations was the current position of the subject (ATR), and the other was a place some distance away.
- One condition was consistent with the physical world, while the other condition was inconsistent with the physical world.

The difference was that one diagram had a configuration consistent with the real-world relationship, while the other did not; that is, the nearer icon in the graphics represented a farther place in the real world. The general starting point was placed at the top of both diagrams. These diagrams are shown in Figure 6.

If the real-world configuration has some effect on setting the reference point, the motion verbs would show

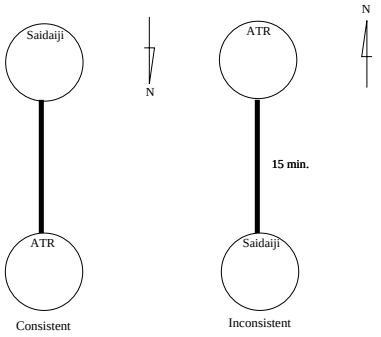


Figure 6: Consistent and inconsistent diagrams

different distributions between the consistent condition and the inconsistent condition. The frequency of *kuru*-type verbs in start-to-goal motion is expected to be lower in the inconsistent condition than in the consistent one. If the real-world configuration did not have much effect, then the distribution would be almost the same between these two conditions. Each condition had eight pairs of subjects.

Results of Experiment 2 The distribution of motion verbs was as shown in Table 2.

Table 2: Distribution of *iku*-type and *kuru*-type verbs in Exoeriment 2.

		iku-type	*kuru*-type
Consistent	Start $\rightarrow$ Goal	70	17
	Goal $\rightarrow$ Start	1	48
Inconsistent	Start $\rightarrow$ Goal	47	25
	Goal $\rightarrow$ Start	4	42

No significant difference was observed between the consistent and inconsistent conditions ($\chi^2{}_{(3)} = 7.00 < 7.81, p = 0.05$). Furthermore, a comparison of each condition with the vertical condition in Experiment 1 also did not show any significant difference (consistent vs. vertical: $\chi^2{}_{(3)} = 2.14 < 7.81, p = 0.05$; inconsistent vs. vertical: $\chi^2{}_{(3)} = 2.83 < 7.81, p = 0.05$). The distribution in each condition was similar to the distribution in the vertical condition of Experiment 1. The *kuru*-type was observed in the Start-to-Goal motion in both conditions. The frequency of the *iku*-type for the Nara to Saidaiji motion was again quite low.

The results show that spatial consistency did not contribute to shifting the reference point. The Observer-to-World Perspective did not have a strong influence in conversation with a diagram. The Protagonist Perspective was also weak in this setting.

It turned out that the spatial property of graphics had a stronger effect on event conceptualization than that of its target world. This suggests that the Observer-to-Graphic Perspective is stronger than the Observer-to-World Perspective in graphical communication settings.

Conclusion

We have analyzed the effect of graphics on language usage in communication. Based on an empirical analysis of the uses of movement verbs in actual conversational data, we have shown that the configuration of a graphical representation affects the reference point setting when people conceptualize motion events from graphics-based perspectives. We found that: (1) the task settings provided yet another kind of perspective, the Problem Perspective, setting a general reference point to the general origin; (2) the Problem Perspective is stronger than graphics-based perspectives; (3) the Observer-to-Map Perspective is the next strongest; and (4) the real-world perspective does not contribute, in comparison with the graphics-based perspective, to the reference point shift.

These results suggest that we are mainly grasping an event of the target world via its representation, rather than from the event itself, in graphical communication situations. The point of using graphical representations is the convenience and the ease of access they give us, which helps us to grasp an event through the mediation of graphics. This mediation makes it possible to talk about distal objects by manipulating their proximal counterparts, thereby facilitating both communication and reasoning processes. This provides us with a novel set of perspectives based on graphics, and conceptualization of target events may be affected by the features of the graphics when people rely on those perspectives.

References

Fillmore, C. J. (1997). *Lectures on Deixis*. Stanford, CA.: CSLI Publications.

Levinson, S. (1996). Frames of Reference and Molyneux's Question: Cross-linguistic Evidence. In P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett. (Eds.) *Space and Language*, 109–169. Cambridge: MIT Press.

Neilson, I., and J. Lee (1994). Conversations with Graphics: Implications for the Design of Natural Language/Graphics Interfaces. *International Journal of Human-Computer Studies 40*, 509–541.

Taylor, H. A., and B. Tversky (1996). Perspective in Spatial descriptions. *Journal of Memory and Language 35*, 371–391.

Umata, I., A. Shimojima, and Y. Katagiri (2000). Talking through Graphics: An Empirical Study of the Sequential Integration of Modalities. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 529–534.

Umata, I., Y. Katagiri, and A. Shimojima (2002). Movement Conceptualizations in Graphical Communication. In Hegaty, M., Meyer, B. and N. Hari Narayanan (Eds.) *Diagrammatic Representation and Inference LNAI 2317*, 3–17. Berlin: Springer.

When Participants Are Not Misled They Are Not so Bad After All: A Pragmatic Analysis of a Rule Discovery Task

Jean-Baptiste Van der Henst (jvanderhenst@hotmail.com)
Laboratory of Experimental Psychology, K.U. Leuven, Tiensestraat 102
Leuven, 3000 Belgium
Institut Jean Nicod, 1 bis avenue de Lowendal
Paris, 75007 France

Sandrine Rossi (rossi@scvie.unicaen.fr)
Laboratoire de Psychologie Cognitive et Pathologique, Université de Caen, Esplanade la Paix
Caen, 14032 France

Walter Schroyens (Walter.Schroyens@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology, K.U. Leuven, Tiensestraat 102
Leuven, 3000 Belgium

Abstract

In this paper we present a pragmatic analysis of a widely used task in the field of hypothesis testing: the 2-4-6 problem (Wason, 1960). In this task participants have to discover the rule "three increasing numbers" by testing triples of numbers and are given the "2-4-6" as an example of triples compatible with the rule. We argue that most people fail because the givens of the task are conversationally misleading: first because the 2-4-6 is communicated and is thus presumed to be relevant (Sperber & Wilson, 1995) and second because the rule to be discovered is too simple in the context of the task. In a first experiment we showed that providing the triple without communicating it improved performance in the task. In a second experiment we contextually increased the relevance of the rule and observed that people were thus more inclined to discover it.

Introduction

Imagine that you have to discover a rule that generates triples of numbers. Some triples are consistent with the rule and some are not. Now, somebody who knows the rule – a trustworthy person like an experimental psychologist – is telling you that '2-4-6' is a triple that is consistent with the rule. Will you consider this as helpful information or not? Surely, you will and you will also probably think that the experimental psychologist would expect you to regard this triple as helpful in order to succeed in the task. Hence, you will consider that the salient properties conveyed by '2-4-6' (like "evenness" or "increase by 2") must be taken into account in order to discover the rule. However, considering that these properties are important is in fact deceptive since the rule to be discovered does not relate to them: the rule is simply *"three increasing numbers"*. Focusing on 2-4-6's most salient properties is thus not the good way to solve such a task!

This task is the well-known '2-4-6' problem designed by Peter Wason more than forty years ago (Wason, 1960) in order to investigate hypothesis testing ability (see Gorman, 1995 and Poletiek, 2001 for reviews). It has become the most commonly used task by researchers in the field of hypothesis testing. In its standard version, it consists in proposing sequences of triples to discover a rule the experimenter has in mind. For each triple, the experimenter indicates whether or not it is consistent with the rule. Participants have to test triples until they are sure of having discovered the rule. As for the other famous Wason's task, namely the selection task (Wason, 1968), one stimulating aspect of the '2-4-6' problem is that few people succeed in it despite of its apparent simplicity. In the initial study (Wason, 1960), only 21% of participants succeeded in discovering the rule in their first announcement. Typically, the rules proposed by participants inherit the salient properties of '2-4-6' and are more specific than the rule to be discovered. For instance, they propose rules such as *"three even numbers", "numbers increasing by 2", "even numbers increasing by 2"*.

The failure in the 2-4-6 task has often been viewed as a sign of irrationality. Wason argued that participants exhibited a "confirmation" bias, and Evans (1983; 1989) argued that people exhibited a "positivity" bias. The protocols indeed reveal that people tend to propose instances of triples compatible with their hypothesis whereas the most efficient strategy consists of proposing instances inconsistent with the held hypothesis. It is commonly accepted that people overly rely on a positive testing strategy and focus on too narrow hypotheses (Poletiek, 2001). What is the reason for this? Researchers assume that positive testing is

simply a natural way of thinking typical of human beings (Evans, 1989; Klayman, 1995) and that the consideration of restrictive hypotheses is due to the salient properties of 2-4-6. In contrast, we think that the incidence of positive testing and the size of "restrictiveness" bias (Poletiek, 2001) have been overestimated. We believe that one of the most important analysis required for understanding the psychological mechanisms underlying the task has been systematically overlooked, namely the *pragmatic analysis*. In the present paper, we argue that a key cause of the failure in the 2-4-6 problem derives from *communication.* In particular, we claim that the task is difficult precisely because the triple 2-4-6 is obtained by communication.

The pragmatics of the 2-4-6 problem

Nobody can contest that conversation plays an important role in the task. First, like many reasoning tasks, the 2-4-6 problem sets up a situation of communication. The experimenter communicates the givens of a problem to a participant and the participant has to communicate the experimenter a conclusion in order to provide him with some information about his or her inferential skills. The participant tries to determine the experimenter's communicative intention and has some expectations about what he or she is interested in. She or he may thus tailor her/his answer according to these expectations. Second, and more importantly, communication is noticeably misleading in the 2-4-6 problem. In Gricean terms, one can view the experimenter as being uncooperative (Grice, 1975) since the triple he/she intentionally choose is overly specific and does not illustrate the level of generality of the rule. Communicating the triple '2-4-6' to illustrate a typical example of the rule is thus a violation of the second maxim of quantity which stipulates that the speaker should not make his/her contribution more informative than required (Grice, 1975). Of course choosing a triple whose most salient properties are consistent with much more specific rules than the one to be discovered was done in purpose. Wason and other subsequent researchers wanted to see if people were able to come up with general hypotheses by attempting to falsify specific hypotheses drawn from the triple 2-4-6. However, what has been neglected is the fact that the consideration of specific hypotheses is made on the basis of a triple that is communicated. Giving the participant a specific triple has not been seen as a violation of a rule of communication but rather as a way to suggest specific hypotheses. Consequently, researchers have not assessed the impact of misleading communication on weak performance in the task. To which extent does the fact that 2-4-6 is communicated contribute to the restrictiveness bias? In this paper, we aim at investigating this issue.

Our pragmatic analysis of the 2-4-6 relies on relevance theory (Sperber and Wilson, 1995). The concept of relevance is characterized by cognitive effects and cognitive effort, and the degree of relevance relies on these two factors: on one hand, the greater the cognitive effects resulting from processing an information, the more relevant the information; on the other hand, the greater cognitive effort required to achieve these effects and process that information the lesser its relevance. Sperber and Wilson (1995) argue that human communication is governed by a *communicative principle of relevance.* According to this principle, each utterance conveys a presumption of its own relevance. This makes an important difference between information received from a communicator and information not obtained by communication. A communicated information raises expectations of relevance. The communicator manifestly intends the information to be relevant enough to deserve the consideration by the addressee. Presuming that the information is relevant implies first that the effects will be sufficient to offset the effort required to process the communicated information. Second, it implies that the effort required is presumed to be minimal to reach the level of expected effects given the communicator's ability. The presumption of relevance sets up a comprehension strategy, which consists in following a least-effort path: considering cognitive effects in order of accessibility and stopping the processing effort when the level of expected relevance is met.

Let's now turn to the task itself. When the experimenter gives the participant '2-4-6' as an example consistent with the rule to be discovered, this triple is accompanied by a presumption of relevance. In other words, the addressee should presume that this triple is relevant to discover the rule. What type of cognitive effects may the addressee expect to draw from processing the triple? She/He will expect that the triple will look as having been generated by a rule and thus will search for properties common to the three numbers or for properties about the way these numbers are ordered. There are many properties that can be attributed to the 2-4-6 triple, but some of these properties are much more salient (i.e. they immediately come to mind with minimal processing) than others. Given the presumption of relevance, the properties to be considered in order to discover the rule are those that are easily accessible from processing 2-4-6 (for instance "evenness" and "increase by 2").

This does not necessarily mean that the rule should correspond exactly to one or several of the most salient properties but this means that these properties indicate the directions to investigate in order to discover the rule. Actually, the rule may still be quite hard to discover and integrate many other characteristics than the one conveyed by the triple, but what is saliently

given in the triple is relevant to discover it. Hence, because the triple is presumed to be relevant, the most accessible properties it conveys cannot be considered as inappropriate clues to discover the rule. When the experimenter communicates the triple '2-4-6' while he wants the participant to discover the rule "three increasing numbers", he violates the participants' expectations of relevance.

Another cause that may contribute to mislead participants is the rule to be discovered itself. When a person is taking part in a reasoning experiment she/he should normally expect to provide the experimenter with some information about her/his inferential skills. Trivial tasks and trivial answers are not well suited for this: they require little cognitive competency and should not interest the experimenter. In order to provide an answer relevant to the experimenter, the participant will probably expect the task to be of a certain level of difficulty. Faced with a task consisting of discovering a rule about numbers, participants should thus think that this rule should be a bit challenging to discover. However, "increasing numbers" is actually one of the most obvious rules applied to order numbers and in daily life we very often encounter set of numbers sorted according to it. Consequently, participants may be reluctant to think of discovering such an undemanding and widespread rule and may try to achieve relevance by seeking for more difficult rules. As he/she is misleading in communicating the triple, the experimenter is also misleading about his own expectations since he actually expects the participants to find out one of the less relevant (i.e. the simplest) rule to discover in the context of the task.

In our experiments we aimed at showing that misleading participants in their expectations of relevance influence task performance. We used a less deceptive way to convey the givens of the problem while still keeping the same provided example (i.e. 2-4-6) and the same target rule (i.e. increasing numbers). In the first experiment, the presumption of relevance, which accompanies the triple 2-4-6 in the standard version of the task, was removed: 2-4-6 was *given* but *not communicated*. In the second experiment, we increased the relevance of the target rule "three increasing numbers".

Experiment 1

In this experiment, there was no presumption of relevance accompanying the triple 2-4-6 in one of the two conditions. Before receiving the instructions of the rule discovery task, participants had to manipulate a "jackpot" generating triples of numbers at random. After several trials, the experimenter gave the participant the instructions about the rule discovery task. He/She then asked the participant to trigger the jackpot for a last time and told him whether or not the obtained triple was compatible with the rule. However, the jackpot was biased in such a way that the sequence 2-4-6 came out on this trial. The participant obviously did not know that we rigged the jackpot on this trial and she/he could expect the triple to be consistent or inconsistent with the rule. Hence, from the participant's perspective, the salient properties of the 2-4-6 just result from chance. Even if this triple suggests specific hypotheses, the participant cannot consider them as ones the experimenter necessarily wanted him/her to think about. This because the triple has not been chosen intentionally, in contrast with the standard version of the task. Participants should rely much less on the salient properties of the 2-4-6 and should perform better than in the standard version. We predict that subjects would focus less on the specific properties conveyed by the triple. Consequently, they should test a greater variety of triples: we should thus observe a greater rate of triples increasing in an irregular way (i.e. triples whose numbers do not increase with the same interval, see also Vallée-Tourangeau, Austin & Rankin, 1995) or counterexamples of the rule to be discovered (i.e. triples which are not increasing).

Participants

Fifty-eight undergraduate psychology students from the University of Caen (France) participated in this experiment. They were tested individually.

Procedure and materials

In the control condition (N=29) participants received the task with the standard instructions. They were required to discover a rule the experimenter had in mind by proposing sequences of three numbers and were given '2-4-6' as an example of triples compatible with the rule. To make sure that the participant well understood the instructions, which were printed on an instruction sheet, the experimenter re-explained them and asked the subject if he/she had any questions. Participants kept a written record of the triples they proposed, their hypothesis about the target rule, as well as the experimenter feedback.

In the "jackpot" condition (N=29), participants were faced with a computer screen resembling a jackpot machine. Participants were informed that it randomly generates sequences of three numbers. The experimenter asked the participants to trigger the jackpot by pressing the key "ENTER". After five trials, the experimenter stopped the "jackpot" session. At this point, the participant did not know yet the purpose of the task and did not know what the use of the jackpot was for. After the jackpot session, the participant was given the rule discovery task as in the standard version. However, instead of receiving an example communicated by the experimenter, he/she had to

trigger the jackpot for a last trial. The experimenter would thus tell him/her whether or not the triple supposed to be randomly generated was consistent with the rule. For each participant, the sequence 2-4-6 came out at this trial and the experimenter told the subject that it was consistent with the rule to be discovered. The participant subsequently had to generate triples by herself/himself, like in the standard version. The jackpot was thus not used to generate triples after "2-4-6" was obtained.

Results

Performance As predicted participants performed better in the jackpot condition than in the control condition: Only 24% of participants gave the correct rule on their first announcement in the control condition (21% in Wason' study) while 55% did so in the jackpot condition (χ^2 (1) = 5.84, p <.02). Moreover, the mean number of rules announced to reach the correct solution was higher in the control condition than in the jackpot condition (2.38 vs. 1.59; *Mann-Whitney* $U_{29,24}$= 214, Z=2.39, p<.01) ; for the five participants of the control group who failed in the task, the mean number of proposed rules was 2.6).

Number and types of triples Participants tested more triples before proposing a rule in the jackpot condition than in the control condition. The mean number of proposed triples per rule by participants who succeeded in the task was higher in the jackpot condition than in the control condition (8.15 vs. 6.11; $U_{29,24}$=202, Z=-2.6, p<.009). The mean proportion of counter-examples (i.e. triples that received negative feedback) for successful subjects was lower in the control condition than in the jackpot condition (0.17 vs. 0.25; $U_{29,24}$=207.5, Z=-2.51, p<.01; this rate is equal to 0.06 for the 5 participants who failed in the control condition). Similarly, the mean proportion of irregular increasing triples was lower in the control condition than in the jackpot condition for these subjects (0.18 vs. 0.29; $U_{29,24}$=218, Z=-2.32, p<.02; this rate is equal to 0.02 for the failing participants). These results indicate that subjects in the jackpot conditions were more prompt to explore a greater variety of triples than subjects in the control condition who focused more on triples exhibiting the salient properties of the 2-4-6.

Discussion

Removing the presumption of relevance of the triple '2-4-6' was helpful: twice as many subjects discovered the correct rule in a single annoucement when the triple 2-4-6 was provided without any presumption of relevance and 83% of participants in the jackpot condition succeeded in the task in no more than two announcements. Moreover, even though giving 2-4-6 in the jackpot condition may still suggest specific hypotheses related to the salient features of this triple, the fact of removing its presumption of relevance entails that these features do not necessarily have to be taken into account in order to succeed in the task. Consequently, participants in the jackpot condition were more inclined to consider alternative properties and thus tested a greater variety of triples than in the control condition.

Experiment 2

In contrast with the Wason selection task literature, in which the content question has led to a vast amount of research, no study has ever investigated content effects with the 2-4-6 problem. That is in the earlier experiments, numbers in the triples proposed by participants or given by the experimenter never referred to real quantities. However, when we use numbers in daily life, we most of the time refer to concrete quantities like books, people, dollars, reasoning errors and so on. We conjectured that framing the task with a real content situation might influence performance. In particular, we think that using an appropriate content and context can enhance the relevance of the rule "three increasing numbers".

In a previous section, we claimed that the rule of increase was too simple to deserve the interest of the participant in the context of an experimental task. However, in real life situations people definitely not avoid looking for an information for the reason that it is too easy to access. They try to look for information that matters for them. We believe that in some contexts, searching for a rule of increase is likely to be cognitively efficient. Indeed, in many real life situations, following a rule of increase is actually highly relevant. A paradigmatic example is economic activity. An economic agent always aims at following such a rule: he or she wants his/her turnover, sailings, productivity, profits, or market scope to increase over the time. Hence, we framed the 2-4-6 task in the context of economic activity. In such a context, the task was to discover not a rule an experimenter − who studies human reasoning and hypothesis testing − had in mind but a rule about car sailings that a garage owner imposes to his new employee. From the participant's perspective, what is relevant for an experimental psychologist who studies cognitive skills is likely to be different from what is relevant for a garage owner. Indeed, searching for the most common way to order numbers (i.e. rule of increase) may be seen as too simple in one case, whereas it becomes highly relevant in the other. Hence, we predict that people will discover the rule "three increasing numbers" more often when the search for such a rule occurs in the context of economic activity.

Participants

One hundred and twenty undergraduate psychology students from the University of Leuven (Belgium) participated in this experiment.

Procedure and materials

The procedure used in this experiment substantially differed from the one we used in Experiment 1. Indeed, participants did not interact with the experimenter. The instructions were given via a computer and participants had to enter the triples they wanted to propose in the computer. The feedback about the triple proceeded in the following way: when the proposed triple was consistent with the rule it appeared in green and when it was inconsistent it appeared in red. The consistent/inconsistent triples remained on the screen where they were presented. After each trial participants had the press the [1] key when they wanted to test another triple, and the [2] key when he/she was sure of having discovered the rule, after which he/she had to type in the rule. After one rule announcement, the experiment ended. This contrasts with Experiment 1 in which participants could propose rules until they discovered the target rule. Participants were tested in groups of ten to twenty people each on their individual computer. In the control condition (N=62), participants had to discover a rule implemented by the experimenter in the computer. In the "economic" condition (N=58) participants received the following real life context:

"Mister Jansens is a prosperous garage owner. He has recently posted an advertisement in order to recruit a car salesman. Bert answered the ad and obtained an interview with Mr Jansens, he told him that he was very motivated for the job but he also informed Mr Jansens that he never sold any car before. Mr Jansens had a good feeling about Bert, but he thinks that an interview is not enough to decide if Bert will be a good car salesman. Hence, Mr Jansens offers Bert to work in his garage for three months as a sale-training period. If Bert does a good job, then he will decide to hire him. In particular, in order to be recruited, Bert's sales during these three months have to follow a rule required by Mr. Jansens. Bert willingly accepts the proposition.

Three months later… It's time for balance! The first month, Bert has been able to sell 2 cars, the second month 4 cars and the third 6 cars. The verdict of Mr. Jansens is very clear: "Perfect! Your sales have well respected the rule. So, I can hire you!"

Results and discussion

In line with our prediction, more participants discovered the rule in the economic condition than in the control condition (3.2% vs. 29.3%; $\chi^2(1)= 15.3$, p=.0001). Moreover participants proposed a greater number of triples in the economic condition than in the control one (2.45 vs. 3.73; *Mann-Whitney* $U_{62,58}$= *1121.0, Z = -3.645,* , p <.05). This indicates that people abstract more beyond the hypotheses suggested by the 2-4-6 triple and thus tested a greater variety of triples. In particular, the mean proportion of irregular increasing triples was higher in the economic condition than in the control one (.090 vs. .125, *Mann-Whitney* $U_{62,58}$= *1553.5, Z = -1.689, p < .05* one-tailed). The mean rate of counter-examples was also higher in the economic condition than in the control one but this difference was not significant (.10 vs. .12). Hence, the results reveal that increasing the relevance for searching the rule "three increasing numbers" improved performance and thus show that people adapt their cognitive skills in order to maximize relevance (Sperber & Wilson, 1995): on one hand, it is indeed not really relevant for a participant to search for a trivial rule like "increasing numbers" in order to exhibit to a cognitive psychologist her/his own cognitive skills, especially if the experimenter provides an example which does not suggest such a rule; on the other hand, it becomes much more relevant to search for a rule of plain increasing in the context of economic activity.

General discussion

In this study we provided the first extensive conversational analysis of the 2-4-6 problem. We argued that communication was highly misleading in this task and that this explains why people focus on overly narrow hypotheses. We claimed that being misled is not a clue of irrationality. We showed that when participants were not misled, they were not so bad after all. In the standard task, people are misled because they rationally consider that the givens of the problem are relevant to solve the task.

Our experiments aimed at providing less deceptive tasks and show that this increased performance. In Experiment 1, we designed a task in such a way that there was no presumption of relevance accompanying the triple 2-4-6. This implied that the salient properties of the 2-4-6 did not have necessarily to be considered by the participant in order to succeed in the task. The results showed that participants performed better when the salient characteristics of the 2-4-6-triple resulted from a random procedure (a jackpot) than when a presumption of relevance accompanied such a triple (as in communication). In Experiment 2 we manipulated the content of the task: triples did not refer to abstract numbers as it has always been the case in previous studies but to numbers of cars sold within a three months period. We framed the task within the context of economic activity in order to increase the relevance of searching for the rule "three increasing numbers". Our study shows that participants tailor the search of their hypotheses according to what they expect to be relevant in the task. When expectations of relevance

coincide with the correct rule, they are more prone to discover it. This is line with the pragmatic analysis of the Wason selection task made by Sperber, Cara and Girotto (1995). According to them, performance in this task is determined by expectations of relevance. Subjects fail in this task because intuitions of relevance do not coincide with the logical answer. Increasing the rate success consist in constructing a context in which intuitions of relevance will match with the logical answer.

A pragmatic analysis of the conversational structure of a cognitive task may be highly helpful in assessing the quality of participants' skills. There is now an increasing body of research in the domain of high-level cognitive processes revealing that lack of rationality is mistakenly attributed on the basis of misleading tasks (for reviews see, Politzer, 1986; Hilton, 1995; Politzer & Macchi, 2000). These studies as well as ours should alert psychologists that weak performance in a task might be overestimated in the absence of a pragmatic analysis.

Acknowledgements

We are very grateful to Professor D. Trotoux and his students of the University Institute of Technology of Caen for the design of the "jackpot" we used in Experiment 1.

References

Evans, J. St. B. T. (1983). Selective processes in reasoning In J. St B. T. (Ed.), *Thinking and Reasoning: Psychological Approaches.* London: Routledge & Kegan Paul.

Evans, J.St.B.T. (1989). *Bias in human reasoning : Causes and consequences*. Hove, UK : Lawrence Erlbaum Associates.

Grice, H.P. (1975). Logic and conversation. In P. Cole, & J.L. Morgan (Eds), *Studies in syntax, Vol 3: Speech acts.* New York: Academic Press.

Gorman, M. E. (1995). Hypothesis testing In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on Thinking and Reasoning. Essays in honour of Peter Wason*. Hove, UK: Lawrence Erlbaum Associates.

Hilton D.J. (1995) The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*, 248-271.

Klayman J., Ha Y.W. (1987) Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, *94*, 211-228.

Klayman, J. (1995). Varieties of Confirmation Bias In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Decision Making from the Perspective of Cognitive Psychology* . New York: Academic Press.

Poletiek, F. (2001). *Hypothesis-testing behaviour*. Psychology Press/Taylor and Francis, Philadelphia.

Politzer G. (1986) Laws of language use and formal logic. *Journal of Psycholinguistic Research*, *15*, 47-92.

Politzer, G. & Macchi, L. (2000). Reasoning and pragmatics. *Mind and Society*, *1*, 73-93.

Sperber D., Cara F., Girotto V. (1995) Relevance theory explains the selection task. *Cognition*, *57*, 31-95.

Sperber, D. & Wilson, D. (1995). *Relevance: Communication and Cognition,* Oxford, Blackwell. 2nde Edition.

Vallée-Tourangeau, F., Austin, N. G., & Rankin, S. (1995). Inducing a Rule in Wason's 2-4-6 Task: A Test of the Information-Quantity and Goal-Complementarity Hypotheses. *Quarterly Journal of Experimental Psychology, 48,* 895-914.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Deriving a Conclusion From Relational Premises

Jean-Baptiste Van der Henst (jvanderhenst@hotmail.com)
Laboratory of Experimental Psychology, K.U. Leuven, Tiensestraat 102
Leuven, 3000 Belgium

Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology, K.U. Leuven, Tiensestraat 102
Leuven, 3000 Belgium

Abstract

Literature on relational reasoning mainly focuses on the performance question. It is typically argued that problem difficulty relies on the number of "mental models" compatible with the problem. However, no study has ever investigated the wording of conclusions that participants formulate. In the present work, we analyse the relational terms that people use in drawing conclusions from spatial relation problems (A is to the left of B, B is to the left C, D is in front of A, E is in front C, What is the relation between D and E?). We show that the linguistic form of premises, the presentation format, the orientation of the question and the internal inspection of the mental model are important factors in determining the wording of conclusions. Our study shows that the type of conclusion produced provides a key to identifying the mental processes involved in solving these problems.

Introduction

Consider the following relational reasoning problem:

A is to the left of B
B is to the left of C
D is in front of A
E is in front of C
What is the relation between D and E?

How might this problem be solved? One possibility consists of building in one's mind an analogical representation, which exhibits the relation between D and E:

A	B	C
D		E

Another is to use inference rules and apply these rules to the propositional form of the premises to derive the required relation. During the last forty years, numerous studies have attempted to discriminate between these two approaches. In the sixties and seventies, they were respectively labelled the 'analogical' approach (DeSoto, London & Handel, 1965; Huttenlocher, 1968) and the 'linguistic' approach (Clark, 1969a; b) and related to representational processes. In the eighties and nineties, they were labelled the "mental model" (Johnson-Laird, 1983; Byrne & Johnson-Laid, 1989) and the "mental logic" approaches (Hagert, 1984; Rips, 1994; Braine & O'Brien; 1998) and related to inferential processes.

To compare these two approaches, researchers have relied on performance results (see Evans, Newstead & Byrne, 1993 for review). They either consider the correctness of the conclusion or the response time. However, no study has ever investigated the type of conclusions that people formulate. It is rather surprising that psychologists of reasoning have ignored this question since in another field of deductive reasoning, namely reasoning involving quantifiers like "all", "none", and "some", this issue was one of the first to be investigated under the guise of the 'atmosphere' effect (Woodworth & Sells, 1935). This effect, which has since been extensively explored (see Evans, Newstead and Byrne, 1993 for review) refers to the fact that universal premises (All A are B) prompt universal conclusions, particular premises (Some A are B) prompt particular conclusions, affirmative premises prompt affirmative conclusions and negative premises prompt negative conclusions.

The question we will address is that of the wording of conclusions in relational reasoning. In the above example, the answers "D is to the left of E" and "E is to the right of D" are both correct. But which of these two conclusions do people actually draw? In this article, we want to show that taking into account the wording of the conclusions can reveal several important mechanisms that occur in representing and in reasoning from spatial relational descriptions. We now present several factors that may influence the wording of conclusions.

The linguistic form of the premises

According to mental model theory, there are two stages in representing the premises. The first stage, which is compatible with a linguistic approach, consists of forming a propositional representation that is close to the surface form of the sentence. The second stage consists of using this propositional representation as a basis for constructing a mental model that is analogous to the situation described in the premises. Once the model is built, the linguistic details of the premises and their propositional representation tend to be forgotten (Mani & Johnson-Laird, 1982). The formulation of the

conclusion should then rely only on the mental model, which does not keep track of the comparative expression in the premises. Hence, the comparative used in the premises should not be used more often than its contrary in formulating the conclusion. 'Left-left' problems, whose first two premises contain the comparative 'left' (problems 1-2-9-10 in Table 1a) should not prompt more 'left' conclusions than 'right-right' problems (problems 7-8-15-16 in Table 1b).

On the other hand, according to the linguistic approach (Clark, 1969a; b; Hagert, 1984) the comparative used in the conclusion should be congruent with the comparative used in the premises. In the above example, the first two premises will be represented by two independent propositions: LEFT (A,B); LEFT (B,C). As a result, the predicates used in the inference rule will match those used in the premises: 'LEFT (X,Y) & LEFT (Y,Z) → LEFT (X,Z)'. Hence, the inferred relation between A and C will be stored with the predicate 'left': LEFT (A,C). Since the relation between D and E is identical to that between A and C, the D-E relation will be stored with the same predicate: LEFT (D,E). More generally, 'left-left' problems should prompt more 'left' conclusions than 'right-right' problems.

It also follows from the linguistic approach that the relevance of the premises may affect the wording of the conclusion. Consider Problem 11 from Table 1. The first premise is irrelevant (as for all problems in Table 1b) since it does not have to be taken into account to answer the question: The relation between D and E relies on the relation between A and C, which is explicitly given by the second premise. Hence, the comparative used in the conclusion is likely to be congruent with the comparative used in that premise. Given this assumption, Problems 11 and 12 (irrelevant premise with 'left', relevant premise with 'right') should prompt more 'right' conclusions than Problems 13 and 14 (irrelevant premise with 'right', relevant premise with 'left').

Table 1: The 16 spatial problems

Table 1a: One-model problems

Pb1	**Pb2**	**Pb3**	**Pb4**
A left B	A left B	A left B	A left B
B left C	B left C	C right B	C right B
D front A	E front C	D front A	E front C
E front C	D front A	E front C	D front A
Pb5	**Pb6**	**Pb7**	**Pb8**
B right A	B right A	B right A	B right A
B left C	B left C	C right B	C right B
D front A	E front C	D front A	E front C
E front C	D front A	E front C	D front A

Table 1b: Two-model problems (first premise always irrelevant)

Pb9	**Pb10**	**Pb11**	**Pb12**
A left B	A left B	A left B	A left B
A left C	A left C	C right A	C right A
D front A	E front C	D front A	E front C
E front C	D front A	E front C	D front A
Pb13	**Pb14**	**Pb15**	**Pb16**
B right A	B right A	B right A	B right A
A left C	A left C	C right A	C right A
D front A	E front C	D front A	E front C
E front C	D front A	E front C	D front A

Problem difficulty and presentation format

Most researchers currently agree that the difficulty of a relational reasoning problem is a function of the number of models it supports. For instance, the following problem,

A is to the left of B
A is to the left of C
D is in front of A
E is in front of C,
What is the relation between D and E?

is compatible with two models,

A	B	C	A	C	B
D		E		D	E

and is more difficult than a one-model problem since two models are harder to construct and store than a single one (Byrne & Johnson-Laird; 1989). It has also been shown that when the number of models increases, people are less likely to construct such models and are more prone to stay at the propositional level of representation (Mani & Johnson-Laird,1982). Mani and Johnson-Laird argued that the indeterminacy introduced by multiple-model problems disrupts the model construction process. They showed that people were more likely to recall linguistic details for indeterminate than for determinate descriptions. Hence, the premises may have a stronger influence over the wording of conclusions for two-model problems than for one-model problems: The comparative used in the conclusion should be more often congruent with the comparative used in the premises with two-model problems than with one-model problems.

In addition, the presentation format might have an effect on the type of representation involved (see Potts & Scholz, 1975; Ormrod; 1979; Schaeken & Johnson-Laird, 2000; Roberts, 2000) and, consequently, on the type of conclusion formulated. One can distinguish between two ways of presenting the premises. With <u>simultaneous</u> presentation, all premises are presented together

with the question and remain available when one solves the problem. Sequential presentation places more load on working memory: the premises are presented one after the other and disappear each time a new premise or the question occurs. It has been argued (Ormrod, 1979) that an analogical representation is more likely to occur with a sequential presentation. The reason is that in a linguistic representation, each relation – given in the premises or inferred from them – is stored separately whereas in a model representation, all premises are integrated within a single representational format. Thus, when working memory load increases it becomes harder to keep track of the premises and inferences separately and a mental model becomes a more efficient and concise mode of representation. Hence, we should observe fewer conclusions with a relational comparative congruent with that of the premises in the sequential presentation than in the simultaneous presentation.

Scanning the mental model

The wording of the conclusions might reveal how individuals scan the model they constructed. If people scan their model in a 'left-to-right' direction, they will be likely to make a 'left' conclusion since the first element they encounter, which is likely to be the first element mentioned in the conclusion, is on the left part of the model; alternatively; if people scan their model in a 'right-to-left' direction, they will be likely to make a 'right' conclusion. One might suppose that the direction of scanning is driven by left-to-right reading habits (Cicirelli, 1977), leading to 'left-to-right' inspections of mental models and to 'left' conclusions.

Another factor that might govern the direction of model-inspection is the question. The question given in the above examples ("What is the relation between D and E?") initially directs attention to the left side of the models since D is mentioned first in the question and is located on the models' left side. Consequently, such a question is likely to induce a 'left-to-right' inspection of the model and a 'left' conclusion. Inversely, the question "What is the relation between E and D?" is likely to induce a 'right' conclusion.

Finally, the order in which the items are inserted within the model might direct inspection of the model. If the premise containing the D item is provided before the premise containing the E item, D will be inserted before E in the model and the construction of the D-E line will proceed from left to right (granted that D is to the left of E as in all the problems of Table 1). Payne (1993) has shown that people keep track of the construction process. One can extend this approach and assume that keeping track of the construction process may induce people to scan their model in the direction of its construction.

Experiment

Before describing the method of the experiment let us first recall the predictions:

- The linguistic form of the premises should influence the wording of the conclusion according to a linguistic approach but not according to an analogical approach.
- Problems conveying an indeterminacy (i.e. two-model problems) could favor more the occurrence of linguistic processes than determinate problems (i.e. one-model problem).
- A sequential presentation could favor more the occurrence of analogical processes than a simultaneous presentation.
- 'Left-to-right' reading habits could prompt 'left-to-right' inspections of mental models.
- 'Left-to-right' questions could prompt 'left-to-right' inspections of mental models.
- 'Left-to-right' constructions of mental models could prompt 'Left-to-right' inspections of mental models.

Method

Participants The participants were 174 first-year psychology students from the University of Leuven.

Design Each participant received 16 relational reasoning problems (Table 1). Half of these problems were one-model problems (Table 1a) and the other half two-model problems (Table 1b). In half of the problems, the first two premises had the same relational term (four 'left-left' problems and four 'right-right' problems); in the remaining half, the first two premises had different relational terms (four 'left-right' problems and four 'right-left' problems). Moreover for half of the problems (Type-1 problems), the premise introducing the item located on the left (i.e. the item D) was given before the premise introducing the item located on the right (i.e. the item E; Problems 1-3-5-7-9-11-13-15 in Table 1). For the other half (Type-2 problems), this presentation order was reversed (Problems 2-4-6-8-10-12-14-16).

There were two between-participant manipulations: 2 types of presentation format × 2 types of question. First, participants received the premises and question in a simultaneous presentation format or in a sequential presentation format. Second, the first item mentioned in the question was either the item mentioned in the third premise (i.e. "what is the relation between D and E?" for Type-1 problems and "what is the relation between E and D?" for Type-2 problems) or the item mentioned in the fourth premise (i.e. "what is the relation between E and D?" for Type-1 problems or "what is the relation between D and E?" for Type-2 problems).

Procedure and Materials. Participants were tested in groups of 12 to 20 individuals. The instructions

and the problems were given in Dutch and were displayed on a screen via a data projector. Participants received 2 training problems and 16 randomly ordered test problems with contents relating to fruits and vegetables (see Table 1). In the simultaneous conditions, each problem was displayed for 50 seconds. In the sequential conditions, each premise and the question appeared for 10 seconds. Participants wrote their answers on a sheet of paper.

Results. Performance was in line with the relational reasoning literature since one-model problems were easier to solve than two-model problems (83% vs. 73%, Wilcoxon's $T = 1787$, $n = 131$, $p < .00001$). We now turn to the analysis of the conclusions that people expressed. We discarded incorrect answers and 0.1% of correct but imprecise answers (i.e. 'D is next to E'). We will now discuss the relevant main effects and interactions.

First, participants had a clear preference for 'left' conclusions. Overall, there were 68.6% of 'left' conclusions and 31.4% of 'right' conclusions, suggesting a tendency for mental models to be scanned in a 'left-to-right' direction.

Second, the type of question influenced the wording of the conclusion. With 'left-to-right' questions (D-E?) there were 83.1% of 'left' conclusions (and consequently 16.9% of 'right' conclusions) whereas with 'right-to-left' questions (E-D?) there were only 54.2% of 'left' conclusions (Wilcoxon's $T = 1082$, $n = 135$, $p < .00001$).

Third, the extent to which the question influenced the incidence of 'left' and 'right' conclusions depended on presentation format. Given the 'left-to-right' question, which prompted a 'left-to-right' inspection of the model, simultaneous presentation gave rise to 78.5% of 'left' conclusion while sequential presentation led to 88.5% of 'left' conclusions (Mann-Whitney U's $= 3029.5$, $n1 = 92$, $n2 = 82$, $p < .01$). A similar, but non-significant trend was obtained for the 'right-to-left' question, which prompted a 'right-to-left' inspection of the model. Here, simultaneous presentation gave rise to 57% of 'left' conclusions, but sequential presentation resulted in 50.7% of 'left' conclusions. The directional scanning induced by the question is apparently enhanced by sequential presentation, in accordance with the notion that analogical processes are more likely under such conditions.

Fourth, the wording of the premises influenced the wording of the conclusions in the simultaneous condition since 'left-left' problems elicited more 'left' conclusions than 'right-right' problems. Whereas, 75.5 % of 'left' conclusions were observed for 'left-left' problems, there were 57.4% of 'left' conclusions for 'right-right' problems (Wilcoxon's $T = 385.5$, $n = 61$, $p < .00005$). In contrast, in the sequential condition there was no influence of the linguistic form of the premises since the rate of 'left' conclusions was essentially the same in each type of problem (71.5% for 'left-left' problems, and 71.1 % for 'right-right' problems). This indicates that when the premises were not available, participants were not inclined to use the comparative introduced by the premises. The results suggest that participants tend to adopt a linguistic representation given simultaneous presentation and an analogical representation given sequential presentation.

Similarly, in the simultaneous presentation condition, the findings obtained for two-model problems, which all had an irrelevant first premise, indicated that participants were prone to formulating conclusions congruent with the relevant premise. When the relevant premise contained the comparative 'left' and the irrelevant one contained the comparative 'right', 77.9% of 'left' conclusions occurred; when the relevant premise contained the comparative 'right' and the irrelevant one contained the comparative 'left', 57.4% of 'left' conclusions occurred (Wilcoxon's $T = 236$, $n = 49$, $p < .0001$). However, given sequential presentation, participants were not really inclined to formulate a conclusion congruent with the relevant premise. When the relevant premise contained the comparative 'left', 69.1% of conclusions were 'left' and when it contained the comparative 'right', the percentage of 'left' conclusions was 67.6.

Fifth**,** one might have expected that when the insertion of the last two items in the mental model proceeds from left to right (Type-1 problems), more 'left-to-right' inspections and 'left' conclusions would occur than when it proceeds from right to left (Type-2 problems). This was not the case and the results even show a tendency in the opposite direction. There were 66.6% of 'left' conclusions for Type-1 problems and 71% of 'left' conclusions for Type-2 problems.

Sixth, it could be argued that when the number of models increases participants should be more prone to relying on the linguistic form of the premises. However, the differences in 'left' and 'right' conclusions were almost identical in one and two-model problems. Indeed, 'left-left' one-model problems gave rise to 73.9% of 'left' conclusions and 'left-left' two-model problems gave rise 73.6% of left conclusions. Similarly, 'right-right' one-model problems gave rise to 62.6% of 'left', and 'right-right' two-model problems gave rise to 64.9% of 'left' conclusions. Hence, people who gave a correct answer to two-model problems did not rely more on the linguistic form of the premises than in the case of one-model problems.

General discussion

Our study is the first to analyze the wording of the conclusions people draw in relational reasoning. We have shown that the wording of conclusions exposes several psychological mechanisms. Two of the effects we have demonstrated are compatible with the analogical approach to reasoning and provide new insight in the way people inspect their

mental model: First, the preference for 'left' conclusions is nicely explained by the fact that people construct mental models and inspect them from left to right. Second, the nature of the question is an important factor in determining the direction of model inspection**:** a 'left-to-right' question prompts 'left-to-right' inspection of the model and a 'left' conclusion. On the other hand, the congruence effect is compatible with the linguistic approach to reasoning. It shows that linguistic details of the premises, like the type of comparative, are stored in memory and are used in the inferential phase.

However, the occurrence of analogical and linguistic processes and the degree to which they are involved largely depend upon the presentation format. Sequential presentation increases the incidence of analogical processing: the "preference-for-left"-effect and the question-effect were stronger in the sequential condition than in the simultaneous condition. Simultaneous presentation induces linguistic processes: the congruence effect was present in the simultaneous condition but not in the sequential condition.

Interestingly, two effects were not observed. First, the number of models did not influence the wording of conclusions: the congruence effect was not greater in indeterminate problems. At first sight this seems to contradict the results of Mani & Johnson-Laird (1982), who found that people more often recalled linguistic details when the description was indeterminate. But they also reported that recall of the gist was lower when the description was indeterminate. Hence, having a weak representation of the gist of the description was related to high retention of the linguistic details of the sentences supporting the description. However, in the results taken into account here the representation was not weak since only correct conclusions were considered. This might explain the absence of a greater congruence effect with 2-model problems, and shows that when the representation of the description is correct people do not rely more on the linguistic level given indeterminate vs. determinate problems.

Second, the direction of model inspection was not congruent with the direction of model construction. People did not scan their model in the direction they constructed it. When construction of the D-E part of the model proceeded from left to right, it did not prompt more frequent 'left-to-right' inspection than when it proceeded from right to left.

In conclusion, some of the data presented here support the analogical framework and others support the linguistic framework. This contrasts with many reasoning experiments in which the data are considered to be entirely compatible with one approach and entirely incompatible with the other (see also Roberts, 1993)

Our findings show that both linguistic and analogical processes do contribute to the wording of conclusions in spatial reasoning. However, the impact of the different kind of processes is unequal: whereas the simultaneous presentation format provides evidence supporting both approaches (i.e. the congruence effect and effects related to model scanning were observed), the sequential presentation condition provides empirical evidence supporting the analogical approach and not the linguistic approach (i.e. the congruence effect was not observed for this condition). This pattern of data seems to indicate that analogical processes are pre-eminent in reasoning from spatial premises.

However, a comprehensive theoretical account of these results has to take into account both types of processes. Such a mixed model has been adopted previously by several researchers like Shaver, Pierson & Lang, (1974) Sternberg (1980) and Johnson-Laird (1983; Mani & Johnson-Laird, 1982). According to Sternberg and Johnson-Laird, the premises are first decoded into a linguistic format and are subsequently represented by a spatial mental model. However, this view concerns only the representational phase but not the inferential phase. Accordingly, it seems then that the inferential phase, during which the reasoner produces a conclusion, relies only on the inspection of the mental model. However the data we obtained indicate that the formulation of a conclusion is influenced by both analogical and linguistic factors in the simultaneous presentation, and support the idea that both factors influence not only the representational phase but also the inferential one. A possible explanation is that when people have achieved the construction of the mental model, they go back to the premises, when they are available, and use the premises as a guide to inspect the mental model. According to this view, linguistic factors play a role, but it does not necessarily imply that inferences rules are used while our data clearly indicate that mental models are constructed.

References

Byrne, R.M.J. & Johnson-Laird, P.N. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*, 564-575.

Clark, H.H. (1969a). Linguistic processes in deductive reasoning. *Psychological Review, 76,* 387-404.

Clark, H.H. (1969b). Influence of language on solving three-term series problems. *Journal of Experimental Psychology, 82,* 205-215.

Cicirelli, V.G. (1977). Children's problem solving in relation to perceptual habit. *Perceptual and Motor Skills, 44*, 883-888.

De Soto, C.B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology, 2,* 293-307.

Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning. The Psychology of Deduction.* Hove, UK: Lawrence Erlbaum Associates.

Hagert, G. (1984). Modelling mental models: Experiments in cognitive modelling spatial reasoning. In T. O'Shea (Ed.), *Advances in*

artificial intelligence, pp. 389-398. Amsterdam: North-Holland.
Huttenlocher, J. (1968). Constructing spatial images: a strategy in reasoning. *Psychological Review, 75,* 550-560.
Johnson-Laird, P.N. (1983). *Mental Models.* Cambridge: Cambridge University Press.
Johnson-Laird, P.N. & Byrne, R.J.M. (1991). *Deduction*. Hove, U.K: Lawrence Erlbaum Associates.
Mani, K., & Johnson-Laird, P.N. (1982). The mental representation of spatial descriptions. *Memory and cognition*, *10*, 181-187.
Ormrod, J.E. (1979). Cognitive processes in the solution of three-term series problems. *American Journal of Psychology, 92*, 235-255
Payne, S.J. (1993). Memory for mental models of spatial descriptions: An episodic-construction-trace hypothesis. *Memory and Cognition, 21*, 591-603
Potts, G.R., & Scholz, K.W. (1975). The internal representation of three terms series problem. *Journal of verbal Learning and Verbal behavior, 14,* 439-452.
Roberts, M.J. (1993). Human reasoning: deduction rules or mental models, or both? *The Quarterly Journal of Experimental Psychology, 46A*, 569-589.
Roberts, M.J. (2000). Strategies in relational reasoning. *Thinking and Reasoning, 6*, 1-26.
Rips, L.J. (1994). *The psychology of proof. Deductive reasoning in human thinking.* Cambridge, Massachusetts, MIT Press, Bradford
Schaeken, W. & Johnson-Laird, P.N. (2000). Strategies in temporal reasoning. *Thinking and Reasoning, 6*, 193-219.
Shaver, P., Pierson, L., & Lang, S. (1974). Converging evidence for the functional role of imagery in problem solving. *Cognition, 3*, 359-375.
Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General, 109*, 119-159.
Woodworth, R.S. & Sells, S.B. (1935). AN atmosphere effect in syllogistic reasoning. *Journal of Experimental Psychology, 18*, 451-460.

Working Memory Capacity and the Nature of Generated Counterexamples.

Niki Verschueren (Niki.Verschueren@psy.kuleuven.ac.be)
Wim De Neys (Wim.Deneys@psy.kuleuven.ac.be)
Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)
Géry d'Ydewalle (Géry.Dydewalle@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology
University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Abstract

This article presents a taxonomic system for generated disablers (based on Elio, 1998) and generated alternatives. Based on the taxonomy, we distinguish three different types of knowledge that are advocated during generation tasks (1) situations that are semantically strongly related to the content of the premises (2) more remote situations and (3) the invalid or low quality counterexamples. Second, we look at the effect of working memory capacity on the nature of generated counterexamples. We found that participants with a high working memory capacity can generate more counterexamples and are flexible in their search process. Participants with low working memory generate less counterexamples and restrict themselves to the first type of counterexamples.

Introduction

Deductive reasoning with causal propositions is one of the core activities of human cognition. The prototypical causal rule is formulated as an 'if-then' sentence. The if-part of the conditional expresses the cause and the then-part contains the effect. The four reasoning problems that are traditionally used to investigate causal reasoning are (1) modus ponens - MP: does the effect follow when the cause is present (2) denial of the antecedent - DA: does the effect follow in absence of the cause (3) affirmation of the consequent - AC: did the cause occur when the effect is observed (4) modus tollens - MT: did the cause occur although the effect did not occur. The answers participants produce to these problems are classically discussed in terms of conditional answers. Schematically, the reasoning problems and answers look as follows. The conditional sentence is: 'If cause, then effect'

	Categorical premise	Conditional answer
MP	The cause occurs.	The effect follows.
DA	The cause does not occur.	The effect does not follow.
AC	The effect occurs.	The cause preceded.
MT	The effect does not occur	The cause did not precede.

Cummins (Cummins, Lubart, Alksnis, & Rist, 1991; Cummins, 1995) found that the tendency to deduce AC and DA is related to the number of alternative causes the reasoner can activate from background knowledge. The number of disabling conditions, on its turn influences the making of MP and MT. Alternative causes is a cause other than the one given, that is capable of evoking the effect. Disabling conditions is an event that can prevent an effect from occurring in the presence of the given cause.

For each of these four reasoning problems Markovits (2000) gives a detailed description of the underlying cognitive mechanism. His theory is based on the mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). The mental model theory assumes that reasoners build internal models representing the premise content, and through manipulation and extension of these models they generate a conclusion. We will briefly discuss Markovits' (2000) account of how the four reasoning problems are solved (applied to causal reasoning).

The mental model theory assumes that (1) reasoners start by representing the content of the conditional sentence in an economical way, for instance '*cause→effect'*. This model represents a possible situation and is often called the initial model. Other possible models of situations are left implicit. When they are asked what follows from the categorical premise reasoners verify whether they can produce a conclusion based on the initial model. In case of MP they can initially conclude that the effect occurs, and for the AC they can conclude that the cause preceded. For the other two reasoning forms, there is no explicit information regarding the absence of effect or cause, so no conclusion can initially be generated. (2) In case of DA and MT, reasoners create explicit models of other possible situations. According to Markovits (2000) the preferred second model is '*no cause→no effect'*. Based on this extra model it is possible to generate an initial conclusion for DA (the effect does not follow) and for MT (the cause did not precede). At this point, a first conclusion is formulated for all four reasoning problems; this conclusion corresponds to the conditional answer. (3) Most reasoners will then validate their initial conclusion by searching for possible counterexamples. For MP and MT, the falsifying model is '*cause →no effect'* (disabler). For AC and DA, the falsifying situation is '*no cause→effect'* (alternative). If a counterexample is found, reasoners become aware that there is more than one conclusion possible, and reject the initial conclusion. When no counterexamples are found, reasoners give conditional answers to all four reasoning forms. Hence, the probability of finding counterexamples informs us about the probability of giving conditional answers.

The probability that reasoners find a counterexample depends on the number of counterexamples that are present in semantic memory. When there are many counterexamples, the probability of retrieving at least one is higher than when there are only few counterexamples. In order to check how many counterexamples reasoners can retrace from memory,

researchers ask participants to generate possible alternatives and/or disablers for a conditional sentence. The number of counterexamples generated in this way reflects the number of counterexamples present in background knowledge, thus reflecting the probability that reasoners find at least one counterexample during reasoning.

Previous research has focussed on specific characteristics of generated counterexamples: the absolute number of counterexamples (see e.g., Cummins et. al., 1991; Cummins, 1995), the salience of counterexamples (Markovits, 2000), and the strength of association between (alternate) causes and a consequent (Quinn & Markovits, 1998). For disablers, Chan and Chua (1994) and De Neys, Schaeken & d'Ydewalle (2001) described the importance of the perceived strength of the connection between cause and effect. Dieussaert, Schaeken, & d'Ydewalle (2002) investigated the differential effect of disablers referring to the item itself or to speaker control.

Another important characteristic is the type of the generated counterexamples. It is possible that some counterexamples are considered to be of greater importance regarding their falsifying strength, than others are. A first step in this research domain is to develop a taxonomy, which enables us to distinguish different types of disablers and alternatives. Elio (1998) has proposed a taxonomy of 'disablers'. Although she constructed this taxonomy from the perspective of belief revision, we consider this taxonomy also useful for research on 'deductive' conditional reasoning. Elio (2001) herself points out that both research area's are complementary, as 'endorsement and entrenchment of a conditional are opposite sides of the same coin'. No taxonomy has been lined out for alternatives. Developing a taxonomy for alternatives will be the first aim of the present study.

Furthermore, we presume that the ability of generating counterexamples is influenced by working memory capacity. Retrieving a counterexample is considered to be a semantic search process (Markovits, Fleury, Quinn & Venet, 1998) and since the efficiency of a semantic search process is linked to working memory capacity (Rosen & Engle, 1997) we deduce that the retrieving of counterexamples is linked to working memory capacity (see also De Neys, et al., 2002). The present study will provide some preliminary data on this topic. Secondly, we will investigate whether there are differences in the nature of generated counterexamples corresponding to differences in working memory capacity.

TAXONOMY

Disablers

Elio (1998) proposed her taxonomy for disablers in the context of belief-change. She first induced a belief-state about the rule by presenting an MP problem and its conditional answer. Then, the participant finds out that this stated conclusion is contradicted by observed facts and is asked to give some sort of rationalisation. Elio (1998; 2002) distinguishes seven categories of disablers. We will illustrate these categories for the sentence: '*If a plant is watered well, the plant stays green'*. A disabler for this sentence explains why the plant doesn't stay green although it is watered well.

The first category contains the *'real' disablers* (1), e.g. 'there is no sunlight'. These answers state that normally the cause produces the effect but in the situation under description there is an extra conditon present which prevents the effect from occuring. Instances of the second category, *demote to default* (2) merely indicate that the given rule is probabilistic in nature, e.g., 'in most cases the plant stays green, but there are exceptions'. The next category contains the *missing enablers* (3), e.g., 'the plant received too little water'. These responses indicate that a condition necessary for the cause to take effect is absent. The fourth category holds *generalisations* (4) of the effect, e.g., 'the plant stays healthy'. The rationale behind this kind of disabler is that the cause produces an effect, but not specifically the effect mentioned in the rule. Another category contains responses that indicate an *invalid relation* (5) between cause and effect; e.g., 'water is not enough for the plant to stay green'. The sixth category contains *exceptional instances* (6), e.g., 'the plant is an oak with brownish leaves'. The rule remains valid, but the participants lists an instance to which the rule exceptionally doesn't apply. The last category contains answers which make reference to *intervening variables or the passage of time* (7), e.g., 'the plant was watered well until last month'. These responses indicate that the cause was indeed followed by the effect, but something happened that cancelled the effect.

Alternatives

In line with the categories proposed by Elio for disablers, we can construct a taxonomy for alternatives. We will use the same sentence to illustrate the different types of alternative causes*, 'If a plant is watered well, the plant will stay green'.* An alternative explains why the plant stays green, even when he is not watered well.

A first category contains the '*real' alternatives* (1), e.g., 'the plant receives a lot of fertilizer'. These are causes, which can also produce the effect, even when the given cause is absent. The second category is called *demote to default* (2), e.g., 'normally the plant needs water to stay green, but not always'. This category contains answers that point out that normally the cause produces the effect, but there are some exceptions, which are not explicitly mentioned. The third category contains *non-missing enablers* (3), e.g., 'the plant needs practically no water'. This category mirrors the missing-enabler category of the disabler taxonomy. The fact that the plant does not need a lot of water is no cause of the plant staying green. It just enables the effect to occur even if the required cause is absent. A fourth category contains the *generalizations* (4), e.g., 'if you take good care of a plant, the plant stays green'. Watering a plant well is an instantiation of the superordinate category 'taking good care of a plant'. The fifth category is the *invalid rule* (5); e.g., 'a plant does not need water to stay green'. This 'alternative' cancels the stated relation between antecedent and consequent. The sixth category contains the *exceptional instances* (6), e.g., 'the plant is a Mexican cactus'. Instances of this category point out that the conditional sentence is valid, but for this

particular example of a plant, the rule does not apply. Finally, the seventh category contains alternatives referring to intervening variables or passage of time (7), e.g., 'after a while the plant learned to live on little water'.

In addition to these 7 parallel categories, we distinguish three extra categories. The first extra category contains answers referring to luck or magic (8), e.g., 'Harry Potter came by and the plant turned green'. The second category contains answers for which the conditional sentence is read in its non-literal meaning (9), e.g., 'the plant sees that other plants receive water and turns green with envy'. A last category is reserved for invalid answers (10), e.g., 'the plant stays green by its photosynthesis', 'the plant does yoga', ... In most experiments where participants are asked to generate disablers or alternatives, these answers are excluded from the analysis. From the perspective of building a taxonomic system, we preferred to put them in a special category. As with Elio's (1998) taxonomy for disablers, we assume that some of these categories have fuzzy boundaries. The category 'luck or magic' is related to 'demote-to-default' and 'exceptional instances'.

The extra three categories are also valid for disablers. They cannot be reduced to one of the seven categories Elio proposed, so for sake of completeness, we will add them to her taxonomy. Since the taxonomy for disablers then fully parallels the taxonomy for alternatives, we can compare the distribution of the answers.

Overall, for disablers as well as for alternatives, we can say that the categories labeled 'disablers' and 'alternatives' contain the '*real' counterexamples*. Instantiations of this category appear to be semantically closely related to the content of the premises. The categories 'demote to default', '(non)missing enabler', 'generalization', 'invalid rule', 'time' and 'exceptional instance' are more *remote*. They either refer to exceptional situations or some of the basic assumptions of the conditional sentence are denied. The categories of 'luck or magic', 'non-literal interpretation' and 'invalid answers' contain counterexamples that can be given to any kind of sentence, regardless of the exact semantic content. We consider these counterexamples to be of *low quality*.

Experiment

Applying the Taxonomy

First of all, we will apply the two taxonomic systems on generated counterexamples. This way we can get some indication of which type background knowledge participants use when asked to produce counterexamples.

Method We used twenty causal 'if-then'-sentences. The sentences covered a broad range of semantic domains. Based on previous research we choose an equal proportion of sentences for the four categories: (1) many disablers and many alternatives, (2) many disablers, few alternatives, (3) few disablers, many alternatives, and (4) few disablers and few alternatives. Our generation task was similar to the one used by Cummins (1995). First, we presented the participants with a causal rule. Subsequently we stated that the cause occurred but it did not produce the effect (disablers) or that the effect occurred in absence of the given cause (alternatives). Participants were then asked to write down as many explanations as possible (maximum 5). It was explicitly mentioned that the given explanations had to be different from the stated cause, different from each other, and that they could only give valid answers, answers such as 'the person came from Mars' are not tolerated.

Sixty-two subjects participated in the experiment as part of course requirements. Thirty-two subjects were given the disabler-generation task, while thirty other subjects received the alternative-generation task. Each participant generated either disablers or alternatives for each of the 20 sentences. The order of the sentences was randomized over participants. The participants were given 15 to 20 minutes to complete the task. For each situation that participants generated, two independent raters determined to which category type the answer belonged. Interrater reliability was .93 for the alternative and .84 for the disabler generation task.

Results and Discussion Both for the alternative and disabler generation task we first divided the sentences into two groups. For one group of sentences (n=10) there are few alternatives or disablers (dis/alt) generated, while in the other group (n=10) there are many dis/alt. The few-group contains sentences for which the total number of generated counterexamples for the sentence is less than the overall mean of all sentences. For the sentences of the many group the number of generated counterexamples for each sentence is higher than the overall mean.

For the alternatives as well as for the disablers, we determined the number of times each category type occurred, this separately for the few and the many sentences. Table 1 gives an overview of the results.

Table 1: Proportion of answers for different categories.

	Disablers		Alternatives	
Category	Few	Many	Few	Many
1. real alt/dis	76.8	81.9	74	94.5
2. demote to default	0.6	3.1	3.6	1.5
3. (non)missing enabler	12.4	7.1	1.9	-
4. generalization	0.6	2.2	2.1	3
5. invalid rule	-	-	-	-
6. exceptional instance	4.1	3.6	10.6	0.1
7. time/ intervening	3.6	1	0.4	-
8. luck/magic	-	0.8	3	0.2
9. non-literal	-	-	0.4	-
10. invalid	1.9	0.2	3.8	0.6
Total N	531	869	470	976

Within the many-group there are relatively more 'real' dis/alt generated then in the few-group. This difference is significant for disablers (p_1=.768, n_1=531 versus p_2=.819, n_1=869, p<.0209) as well as for alternatives (p_1=.74; n_1=470 versus p_2=.945, n_1=976, p<.0001). Additionally, we found that for the few disabler group, more missing enablers are generated than for the many group (dis: p_1= .124, n_1=531 vs. p_2=.071, n_1=869, p=.0008). For alternatives we observe that

in the few group participants more often list exceptional instances than in the many group (p_1=.106, n_1=470 vs. p_2= .01, n_1=976, p<.0001). All other differences between proportions are non-significant.

We assume that when participants are asked to generate dis/alts they start to look for straightforward examples, namely the 'real' dis/alts (category 1). This is because the 'real' counterexamples are semantically strongly related to the content of the conditional sentence. In addition to these 'real counterexamples' participants dispose of another pool of possible counterexamples, namely, the more remote situations. Instantiations of this type are semantically not directly linked to the premise content. They refer to exceptions to the normal situation (category 2,6) or to conversational implicatures that are suspended (category 3,4,7). The assumptions that are normally valid, such as 'promises are kept', 'birds can fly', 'coffee contains caffeine' are examined in order to account for the apparent contradicting premises. In general, this more remote category contains counterexamples sprouting from suspended conversational implicatures (Levinson, 2000).

For some assumptions you find that when the assumption not holds, the relation between antecedent and consequent changes, and can account for the apparent contradiction.

When only few dis/alt can be found, it is harder to find a full range of 'real' counterexamples. As a result, participants search also for the more remote type of disablers and alternatives.

Conclusion The two taxonomic systems can be used to categorize the answers participants give when asked to generate disablers or alternatives. Although some of the presented categories are conceptually related, the raters consistently classified the answers.

By applying the taxonomy we found that more 'real' alternatives and disablers were generated in the many groups than in the few groups. This difference is compensated by a shift to the more remote types. We assume that participants start searching for counterexamples from the pool of 'real' counterexamples, because these counterexamples are semantically close to the content of the premises. In addition, the search can be directed to the more remote categories.

In the second part of this experiment we will investigate whether working memory capacity affects the type of the generated dis/alts.

Working Memory Capacity

Double task experiments showed that working memory capacity puts a constraint on the ability to generate counterexamples (De Neys, Schaeken, & d'Ydewalle, 2002). First, we will investigate the effect of working memory capacity on the *number* of generated counterexamples. We expect that participants with high working memory capacity generate more counterexamples than those with low working memory capacity. We expect this difference to be larger for sentences with few counterexamples. For the many sentences we assume that the difference may be blurred due to a ceiling effect. Second, we will look at the effect of working memory on the *type* of generated counterexamples. Do differences in working memory capacity affect somehow the sort of counterexamples participants come up with?

All first year psychology students had fulfilled a Dutch version of the OSPAN test (La Pointe, & Engle, 1990; Dutch version: De Neys et al., 2002) for measuring working memory capacity. As such, we can link the number and nature of the generated answers to differences in working memory capacity.

Results and Discussion The subjects are divided in three groups depending on their working memory capacity. The high participant group consists of the top third (dis: Min: 37; Max: 54 - alt: Min: 39; Max: 54). The low group contains participants with scores of the bottom third (dis: Min: 18; Max: 24 - alt: Min: 39; Max: 54). Table 2 displays the distribution of the relative proportion of answers.

Participants with high working memory capacity generate more disablers than participants with a low working memory capacity (dis: p_1=.45; n_1=964 versus p_2=.55; n_1=964, p<.0001). This difference is not significant for alternatives.

Table 2: Proportion of generated counterexamples for each category (numbers refer to categories of Table 1). The shaded regions refer to the three types category 1 equals Type 1, category 2 to 7 corresponds to Type 2, categories 8 to 10 are labeled Type 3.

	Disablers						**Alternatives**					
	Low			High			Low			High		
Category	Few	Many	Total	Few	Many	Total	Few	Many	Total	Few	Many	Total
1	82.5	86.4	85	71.7	80.1	76.7	79.4	95.2	90.4	76.7	93	87.7
2	5.8	2.9	3.9	0.9	4.2	2.8	6.1	1	2.6	2.8	1.5	1.8
3	3.9	5.7	5.1	14.6	8.7	11.1	1.8	-	0.6	4.4	-	0.5
4	0.6	1.4	1	0.5	2.9	1.9	-	3.2	2.2	1.1	3.5	2.7
5	-	-	-	-	-	-	-	-	-	0.6	-	0.2
6	3.3	2.5	2.8	5	2.6	3.6	5.5	0.3	1.9	8.9	-	2.9
7	3.3	-	1.2	3.7	0.6	1.9	0.6	-	0.2	-	-	-
8	-	1	0.7	-	0.6	0.4	4.2	-	1.3	2.8	0.5	1.3
9	-	-	-	-	-	-	-	-	-	0.6	0.3	0.4
10	0.6	-	0.2	3.7	0.3	1.7	2.4	0.3	0.9	5	1.5	2.5
Total N	154	279	433	219	312	531	165	375	540	180	371	551

We will now discuss the effects of working memory capacity separately for the sentences with few and many disablers.

A first important result is that the observation that participants with high working memory capacity generate *more disablers* than those with low working memory capacity, is only found on the sentences with few disablers (p_1=.413; n_1=373 vs. p_2= .587; n_1=373, p<.0001). This finding can be explained as follows. For sentences with only few disablers, it is inevitably harder to generate counterexamples than for sentences with many disablers. In general, for sentences with only few disablers most participants quickly run out of inspiration (only in 5% of the trails there were more than 3 disablers given). Because participants with low working memory capacity experience more difficulty in generating disablers, we can expect that their searching process takes more time than that of participants with high working memory capacity. For sentences with many disablers, we assume that there is a ceiling effect. As participants can choose from a large pool of possible counterexamples, the differences in working memory capacity on the generated number of counterexamples does not show.

A second striking finding is that participants with a low working memory seem to restrict themselves to a single type of counterexamples. Participants with a high working memory capacity generate overall more 'real' disablers, this group represents a larger proportion of the generated responses of the participants with low working memory (p_1= .85; n_1=433 vs. p_2= .77; n_1=531, p<.0013). This decrease in 'real' disablers is mirrored by a significant increase in the proportion of disablers of the 'remote' type. Participants with a high working memory score generate more missing enablers (category 3) than participants with low working memory capacity (p_1=.051, n_1=433 versus p_2=.111, n_1=531, p<.0009). As stated above we assume that participants start their search for counterexamples by checking situations, which are semantically related to the content of the premises. In addition, reasoners can check situations that are semantically more remotely related to the premise content. Thus, the results suggest that participants with a high working memory capacity can more easily shift from the straightforward type of counterexamples to the more remote type. Participants with a low working memory capacity are rather conservative in their search for counterexamples. It can be argued that the flexibility to change from one semantic domain to another yields a substantial profit in finding counterexamples. Based on our results we can add that working memory capacity is a crucial mediator of this flexibility.

The significant effects on disablers are paralleled by non-significant trends for alternatives. The absence of any significant working memory effects on alternatives can be explained with reference to the structural difference between the two types of counterexamples. When you are asked to generate a disabler, you have to find a situation in which the effect does not occur in presence of the given cause. The presence of the given cause constitutes an important element of the situation you have to generate. As a result you have to maintain two different propositions, the cause as well as the effect, in memory. For alternatives it is not necessary to maintain the given cause in memory. You just need to look for some alternate causes, and maintain the effect in memory. Markovits argues that for young children, it is harder to search for disablers than to search for alternatives (Jeanveau-Brennan, & Markovits, 1999; Markovits, 2000). We assume that his finding can be generalized to adults. No effect of working memory capacity is observed on alternatives because the generation of alternatives does not challenge working memory capacity in the way that the generation of disablers does.

General Discussion

This article addressed four main issues. First of all, we applied Elio's (1998) taxonomic system for disablers in belief-revision to data gathered in a conditional reasoning perspective. The described system was equally valid for categorizing disablers, as almost all the answers could readily be categorized. Elio (1998) pointed out that belief revision and deductive reasoning are complementary fields of research. This experiment proves that the use of her taxonomic system can be generalized to the domain of conditional reasoning.

Second, we constructed a taxonomic system for generated alternatives. By applying this categorization system we can distinguish different types and categories of alternatives. Although some categories are interrelated, only few answers were subject to discussion. Hence, we conclude that the taxonomy serves its purpose well. In line of Elio's (1998) proposal we suggest that this taxonomic system can also be used for categorizing the alternatives generated in the context of belief-revision. By applying this system researchers are able to shed light on the type of knowledge that is used during the process of belief revision, or in terms of Elio (1998), on the belief-revision operators that people use for resolving everyday contradictions.

Recent research emphasizes the importance of pragmatic and semantic aspects in theories on conditional reasoning (see e.g., Chan & Chua, 1994; Newstead, Ellis, Evans, & Dennis, 1997; Quinn & Markovits, 1998). By outlining two taxonomic systems we provide researchers with an additional methodological weapon for disclosing how the search for counterexamples takes place. These taxonomic systems can also be used for categorizing counterexamples for other types of conditionals than causal ones.

Third, we used the categorization of alternatives and disablers to examine the sort of background knowledge that is advocated during the search for alternatives or disablers. Three broad types of counterexamples were distinguished. The first type contains the 'real' disablers or alternatives. These are descriptions of situations that are semantically close to the stated premises. The second type of answers is of a more 'remote' type. They include answers in which the normal conversational implicatures are suspended. Participants go beyond the usual scheme's (Chan & Chua, 1994) that the premises refer to, in order to find some condition that could not apply (enabler, truthfulness of the speaker, …). A third category contains answers that are

given just to lengthen the list of alternatives. They include answers referring to some magical interference, plain luck, a non-literal reading or just invalid responses. For a possible rule to decide to start looking in another pool, we can refer to the stopping rule proposed by Johnson-Laird (1994); when it gets too hard to generate another imagined situation, participants stop their search (for alternative stopping rules, see Elio, 2002).

Fourth, we looked at the effect of working memory capacity on the taxonomic distribution of the generated disablers and alternatives. We found that reasoners with high working memory retrieve more disablers and are more flexible in their search. They tend to retrieve different types of disablers while reasoners with low working memory capacity are more conservative. Reasoners with low working memory capacity generated more 'real' disablers then reasoners with high working memory capacity. This result suggests that reasoners with low working memory capacity start by searching the pool of semantically related disablers and are conservative in their search. In contrast, reasoners with high working memory capacity are more flexible in redirecting their search.

We like to add that asking participants to generate as many disablers or alternatives as possible could reflect a somewhat different cognitive process than the process active during the validation phase of reasoning (see also Markovits, Fleury, Quinn & Venet, 1998).

In sum, participants with a high working memory capacity can retrieve more counterexamples and are flexible in their search process. Participants with low working memory generate less counterexamples and restrict themselves to counterexamples of the first type. We find that considering qualitative aspects of generated counterexamples provide valuable information on the underlying cognitive search process. In this study we describe three different types of counterexamples. Furthermore, we argue that working memory capacity is not only a crucial mediator for maintaining and searching information but determines also a reasoner's flexibility to search different semantic domains.

Acknowledgments

This work is carried out thanks to the support of the Fund for Scientific Research Flanders.

References

Chan, D., & Chua, F. (1994). Suppression of valid inferences: syntactic views, mental models and relative salience. *Cognition, 53*, 217-238.

Cummins, D.D. (1995) Naïve theories and causal deduction. *Memory and Cognition, 23,* 646-658.

Cummins, D.D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19,* 274-282.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2000). *Causal conditional reasoning and strength of association: the disabling condition case.* Psychological Report N°: 271. Leuven: University of Leuven, Laboratorium of Experimental Psychology.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2001). Does pure water boil, when it's heated to 100°C?': The associative strength of disabling conditions in conditional reasoning. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society, 249-245*. Mahwah, NJ: Lawrence Erlbaum Associates.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). *Working memory capacity and causal conditional reasoning*. Manuscript in preparation.

Elio, R. (1997) What to believe when inferences are contradicted. The impact of knowledge type and inference rule. *Proceedings of the Nineteenth Annual* Conference of the Cognitive Science Society, 211-216. Hillsdale, NJ: Lawrence Erlbaum Associates.

Elio, R. (1998). How to disbelieve p=>q: Resolving contradictions. *Proceedings of the Twentieth Meeting of the Cognitive Science Society, 315-320.* Mahwah, NJ: Lawrence Erlbaum Associates.

Elio, R. (2002) *Belief revision and plausible inference*.. Manuscript submitted for publication.

Elio, R., & Pelletier, F.J. (1997). Belief revision as propositional update. *Cognitive Science, 4,* 419-460

Jeanveau-Breannan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology, 35*, 904-911.

Johnson-Laird P., Byrne, R., Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99,* 418-439.

La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1118-1133.

Markovits, H., (1984). Awareness of the 'possible' as mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology, 75*, 367-376.

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Thinking and Reasoning, 6,* 335-347.

Markovits, H., Fleury, M., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69*, 742-755.

Newstead, S. E., Ellis, M. C., Evans, J. St. B. T., & Dennis,I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3*, 49-76.

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: strength of association as a predictive factor for content effects. *Cognition, 68*, 93-101.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. Journal of Experimental Psychology: General, 126, 211-227.

Thompson, V. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*, 742-758.

Thompson, V., (2000). The task-specific nature of domain-general reasoning. *Cognition, 76,* 209-268.

A Study of Object-Location Memory

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)
Todd R. Johnson (Todd.R.Johnson@uth.tmc.edu)
Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu)
Yue Wang (Yue.Wang@uth.tmc.edu)
School of Health Information Sciences, University of Texas Health Science Center at Houston
7000 Fannin, Suite 600, Houston, TX 77030 USA

Abstract

This paper aims to study the representational nature of human object-location memory. Two experiments are reported, including both performance data and eye movement data. The results show that multiple allocentric frames of reference are used to encode spatial relationships among objects and late computation in object-location memory retrieval in object-cued conditions is often inevitable. The implications on developing a general model of human spatial cognition are discussed.

Introduction

One important aspect of human spatial memory has to do with remembering the location of objects relative to each other. For example, you might recall that the book you read last night is on your office desk between your computer and the desk lamp. This type of memory for spatial relationships is an essential component of a more general type of memory for spatial layout and is obviously critical for many spatial tasks including locating and navigation (see Tversky, 2000, for a review).

It is not clear, however, how spatial relationships among objects are encoded in memory. While it seems apparent that allocentric frames of reference (i.e., locations defined relative to external objects) rather than egocentric frames of reference (i.e., locations relative to the observer self) are often used to describe these spatial relationships, the representational and computational nature of this description is controversial (see Hunt & Waller, 1999; Klatzky, 1998). For example, are object-based spatial relationships encoded and stored directly (early computation)? Or do they have to be inferred much later at the retrieval stage (late computation)? What factors determine which representational scheme is used?

In this paper we report two experiments we conducted to directly address these issues. The results show that multiple allocentric frames of reference are used to encode spatial relationships among objects and late computation in object-location memory retrieval in object-cued conditions is often inevitable.

This paper consists of three major parts. In the first section, the experimental paradigm is briefly introduced. In the second section, the experiments are reported, including both performance data and eye movement data. In the final section, we briefly discuss our ongoing work of developing a computational model of the object-location memory and its implications on modeling human spatial cognition in general.

The Experimental Paradigm

We adopted an experimental paradigm developed by Milner and colleagues in the 1990s, which we call the Milner paradigm (e.g., Johnsrude, Owen, Crane, Milner, & Evans, 1999; Milner, Johnsrude, & Crane, 1997; Owen, Milner, Petrides, & Evans, 1996). Though their focus was on neuroimaging studies of the brain foundations of object-location memory, the Milner paradigm offers an elegant experimental design that allows a systematic evaluation of multiple schema for representing spatial relationships. In addition, the availability of neuroimaging data provides invaluable constraints on both understanding behavioral results and developing computational models (e.g., Wang, Johnson, & Zhang, 2001).

There are two phases in the Milner paradigm. In the encoding phase (Figure 1A), eight drawings (objects) are individually presented on a computer screen to subjects, with each drawing accompanied by two landmarks (solid squares). Subjects are asked to remember the locations of drawings, relative to the landmarks. In the retrieval phase, subjects are presented some cues plus two identical drawings. One of the two identical drawings (target) is presented in its original location, and the other one (noise) is presented in a different location (or more accurately, it occupies the original location of another object). Subjects are required to perform a forced-choice recognition of the target, relative to the cues. Milner and colleagues originally used four retrieval conditions:

1. In the fixed-landmark condition (Figure 1C), the two landmarks were presented as cues, along with the target-noise pair. The absolute location of landmarks and objects on the screen was unchanged from their original encoding positions.
2. In the shifted-landmark condition (Figure 1D), the two landmarks were presented as cues, along with the target-noise pair. Though the spatial relationships among the landmarks/drawings remained unchanged, the absolute locations of the landmarks/drawings on the screen were shifted.
3. In the fixed-object condition (Figure 1E), two encoded drawings instead the two landmarks were

presented as cues, along with the target-noise pair. The absolute locations of drawings on the screen were unchanged.

4. In the shifted-object condition (Figure 1F), two encoded drawings instead the two landmarks were presented as cues, along with the target-noise pair. Though the spatial relationships among the drawings remained unchanged, the absolute locations of the drawings on the screen were shifted.

One significant feature of the Milner paradigm is that it simultaneously involves multiple spatial representations, including screen-based, landmark-based, and object-based. While landmark-based representations are perceptually accessible in the encoding phase (because an object was always presented along with the two landmarks in the encoding phase), object-based representations are not (because no two objects are presented at the same time in the encoding phase). Therefore, this fact alone might suggest that object-based retrieval would be harder than landmark-based retrieval. Systematic alignment of the different testing conditions allowed Milner and colleagues to use a subtraction method to determine the brain areas that dominate in the different test conditions.

Behavioral data was only briefly reported in Johnsrude et al (1999). It was found that the shifted-object condition was harder (e.g., longer RTs and lower accuracy) than any other conditions, which did not differ from each other. Neuroimaging data suggested that object-location memory in general involved the parahipocampal system, and the shifted conditions, as compared to the respective fixed-conditions, activated the posterior inferotemporal cortex. Both areas have been believed to subserve important functions of spatial cognition (e.g., Burgess, Jeffery, & O'Keefe, 1999).

Experiment 1

In Experiment 1 we added one more testing condition to the original Milner paradigm. In this additional condition, called the fixed-nocue condition, no cues were presented along with the target-noise pair in the retrieval phase. Subjects had to make the forced-choice based solely on the absolute location of objects on the screen. This condition was added to explicitly test the effect of screen-based spatial representations in location retrieval.

Another purpose of experiment 1 was to collect eye movement data. Both perceptually encoding and cognitively computing spatial relationships invite eye movements. The trace of natural eye movements during the task provides an indication of the deployment of attention (e.g., Corbetta et al., 1998) and may shed light on the underlying spatial representations and operations (e.g., Colby & Goldberg, 1999).

Subjects, Apparatus, and Materials

21 subjects, 8 females and 13 males, with normal or corrected-to-normal vision, were paid to participate in the experiment. Five sets of stimuli (each consisting of eight drawings) were created using digitized black and white representational drawings of common objects, selected from the database of (Snodgrass & Vanderwart, 1980). The drawings, 100x100 pixels in size, were presented against a white background on a 19'' VGA monitor with a resolution of 1024x768. The monitor was in front of the subjects within 2 feet. Subjects were asked to respond by clicking with a mouse which was within comfortable reach. 11 subjects wore a head-mounted ISCAN eye-tracker while they were doing the experiment.

Design and Procedure

Each subject performed all 5 experimental conditions, each with a different stimulus set. The design is illustrated in Figure 1.

In each encoding trial, subjects were presented one drawing and two landmarks and were instructed to remember the location of the drawing relative to the landmarks. Subjects clicked the drawing to go on to the next trial. There were 32 encoding trials in each condition, with each drawing presented 4 times. The presentation order was randomized. During the study subjects did not know which testing condition would follow.

In each retrieval trial, subjects were presented the cues and the target-noise pair according to the testing condition. Subjects were instructed to choose the target, by clicking, as quickly as possible and as accurately as possible. As soon as the subjects clicked, the next trial was presented. Each drawing was tested 4 times.

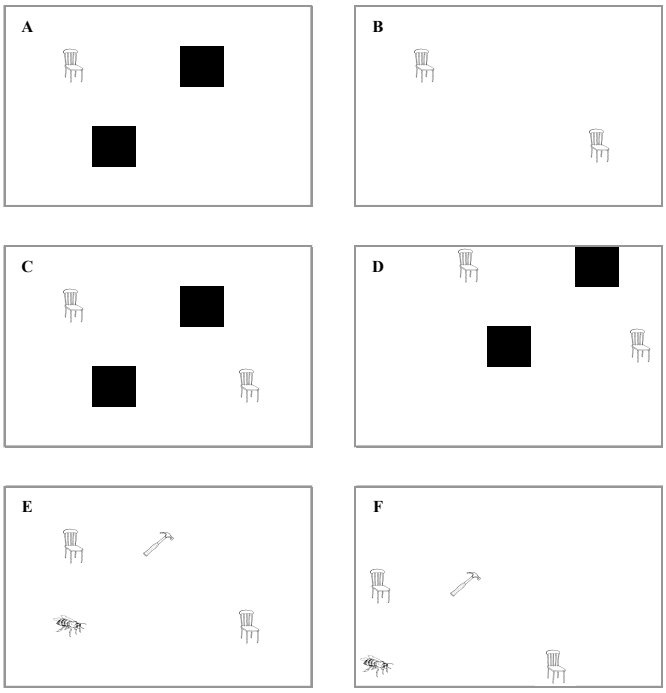


Figure 1. The design of Experiment 1. A, an encoding trial; B, fixed-nocue retrieval; C, fixed-landmark retrieval; D, shifted-landmark retrieval; E, fixed-object retrieval; F, shifted-object retrieval.

Results

Accuracy data. The average accuracy was at least 93%, and there was no difference among the 5 testing conditions.

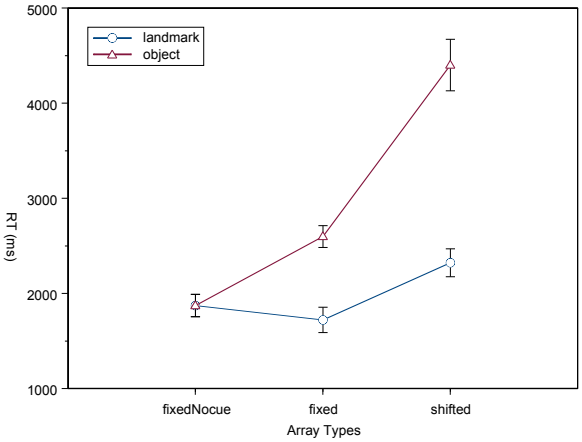


Figure 2. RT data in Experiment 1. The error bars represent 95% confidence intervals.

RT data. The reaction time data is shown in Figure 2. An ANOVA shows a significant interaction between the cue type (landmark vs object) and the array time (fixed vs shifted). In addition, a post-hoc comparison shows that the shifted-object condition takes significantly longer than any other conditions, consistent with Johnsrude et al (1999) results.

Eye movement data. Eye movements are needed to search the scene and measure spatial relationships. The number of eye fixations in each trial is counted and reported here. The result is shown in Figure 3. It is interesting to note that the eye fixation pattern is remarkably similar to the RT pattern, indicating that the number of eye fixations is a good predictor of RT.

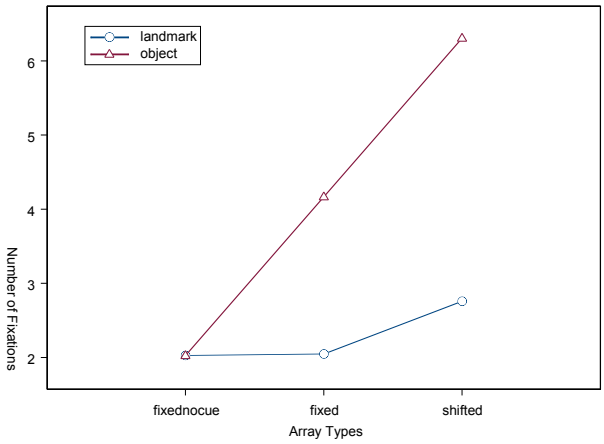


Figure 3. The number of eye fixations in Experiment 1.

Summary & Discussion

Experiment 1 resulted in two major findings that have not been reported by Milner and colleagues. First, the reaction time in the fixed-nocue condition was not different from that in the fixed-landmark condition. Since a screen-based spatial representation had to be used in order to perform the fixed-nocue condition, this result indicates that a screen-based spatial representation might be *implicitly* encoded and stored (because subjects were specifically instructed to pay attention to the drawing's location relative to the landmarks), and be adopted to perform the fixed-landmark condition. This hypothesis was further supported by the eye movement data. The number of eye fixations in both conditions was about 2, the minimal fixations needed to identify the target if a conservative check-both-target-and-noise-before-click strategy was used. The eye movement traces also indicated that subjects often ignored landmarks in the fixed-landmark condition.

Second, the significant interaction between the cue type (landmark vs object) and the array type (fixed vs shifted) was surprising. The reaction time in the shifted-object condition was significantly longer than that in any other conditions (the RT in the shifted-object condition was about 1800ms, 2100ms, and 2700ms longer than that in the fixed-object, shifted-landmark, and fixed-landmark conditions, respectively, see also Table 1), indicating some additional operations occurred in that condition. An analysis of the computational differences among conditions sheds light on what these operations could be. a) Landmark cues (solid squares) were much more perceptually distinct than object cues. In both object-cued conditions, an additional search operation was necessary in order to distinguish the target-noise pair from the two object cues. b) Compared to the fixed conditions, both shifted conditions required explicit access of spatial relationships, either landmark-based or object-based. While landmark-based spatial relationships might be directly encoded in the encoding phase and later directly retrieved in the retrieval phase, it seems that object-based spatial relationships had to be derived through late computation because subjects never saw any two objects at the same time.

Eye movement data, however, indicated that this hypothesis might be oversimplified. In the encoding phase, we quite often observed that subjects moved his/her eyes back and forth between the currently presented object and the location of the previously displayed object (in the previous trial, which has already disappeared), indicating some form of object-based spatial relationships might be encoded directly and quite early. In general, however, it seems likely that shifted-object conditions involved quite extensive late computation in determining object-based spatial relationships.

We speculate that a race model can be used to explain the data. Specifically, when multiple types of representations for spatial relationships are available to solve the task at hand, they compete. Though often the representation that affords easiest operations dominates, sometimes they

interfere with each other. A decomposition of the representations/operations for each condition is summarized in Table 1. It seems that the race model explains the RT data reasonably well.

Table 1: A representational decomposition

	RT (ms)	Accessible representations/operations: Early computation	Late computation	Addn. ops
Fixed-nocue	1874	Screen-based		
Fixed-landmark	1723	Screen-based Landmark-based		
Fixed-object	2599	Screen-based	Object-based	Search
Shifted-landmark	2324	Landmark-based		
Shifted-object	4402		Object-based	Search

Experiment 1 raised two issues. The first one is the role of search in the object-cued conditions. Since the target-noise pair and the object cues are visually indistinguishable, a non-spatial visual search component is necessary to perform the task. The search component was a free parameter in the above race model that could be estimated but it obviously confounded the results. It would be useful to eliminate this confound. The second issue also has to do with the object-cued conditions. In Experiment 1, the two objects that were chosen to be cues in each trial were randomly selected from all possible objects (i.e., those that were not the target-noise pair). This made the task hard in the sense that all possible object-based spatial relationships (there were 7^8 of them!) might be relevant in the retrieval phase. This was in sharp contrast with the landmark-cued conditions, which had only 16 relevant spatial relationships (8 for each landmark). Therefore, it might be the pure number of relevant spatial relationships but not the late computation of object-based representations that made the object-cued conditions more difficult. We designed experiment 2 to explore these two issues.

Experiment 2

Experiment 2 adopted the same Milner paradigm, but differed from Experiment 1 in three aspects. First, the landmarks were changed from solid black squares to a white-filled black square. Second, in the object-cued conditions, the object cues were framed in a black squared to make them visually salient. The purpose of the change was to eliminate the search component in retrieval. Third, we added two more object-cued conditions, which we called consistent mapping conditions. In these conditions, instead of selecting two cue objects at random for every trial, the two objects were selected at the beginning of the testing session and consistently served as the object cues for every trial in that session. This change greatly reduced the relevant spatial relationships and could be viewed as a middle condition between the pure object-cued condition and the pure landmark-based condition.

These changes resulted in six testing conditions, as shown in Figure 4. 14 subjects were paid to participate in the experiment.

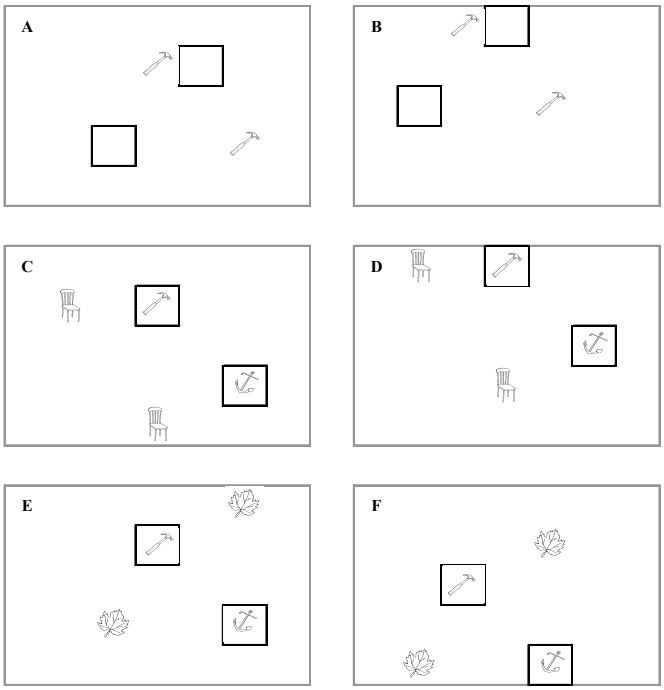


Figure 4. The design of Experiment 2. A, fixed-landmark retrieval; B, shifted-landmark retrieval; C, fixed-object retrieval; D, shifted-object retrieval; E, fixed-object consistent mapping retrieval; F, shifted-object consistent mapping retrieval.

Results & Discussion

Accuracy data. The average accuracy was at least 88%, and there was no difference among the 6 testing conditions.

RT data. The reaction time data is shown in Figure 5. An ANOVA reveals similar effects to Experiment 1, including the significant interaction between cue types (object vs landmark) and array types (fixed vs shifted).

The effects of the two manipulations we adopted in Experiment 2 were evident. First, combining the results from both experiments, it is clear that the elimination of the search operations (by framing the object cues) did decrease reaction time in certain object-cued conditions. However, this time saving had a surprising interaction with the array types. Specifically, while the time saving in the fixed-object condition was not significant (2599ms in Experiment 1 vs 2494ms in Experiment 2) the saving was significant in the shifted-object condition (4402ms in Experiment 1 vs 3548ms in Experiment 2). It is not so obvious how to explain this interaction. Second, the manipulation of consistent mapping in object-cued conditions also reduced

the reaction time. However, again, a reliable interaction with array types was found. While the time reduction was about 450ms in the fixed-object conditions (2494ms vs 2044ms), the reduction was about 900ms in the shifted-object conditions (3548ms vs 2640ms). Similarly, it is not obvious how this interaction occurred without a detailed computational model.

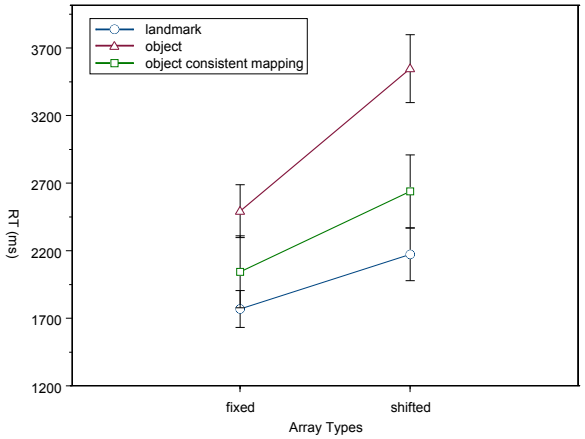


Figure 5. RT data (in ms) in Experiment 2. The error bars represent 95% confidence intervals.

Eye movement data. Similar to Experiment 1, eye movement data corresponded quite well with the reaction time data (see Figure 6). Fewer numbers of eye fixations were observed in the fixed array conditions than in the shifted array conditions. In addition, the object-cued conditions induced more eye fixations than the landmark-cued conditions. In particular, both the elimination of the search component and consistent mapping in object-cued conditions significantly reduced the number of eye fixations (by about 1 and 1.5, respectively), indicating both manipulations successfully reduced the efforts of object recognition and late computation of spatial relationships.

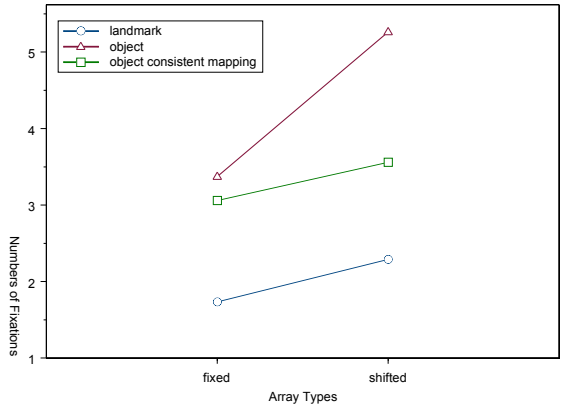


Figure 6. The number of eye fixations in Experiment 2.

General Discussion

Memory for object-location is an essential aspect of human spatial memory. However, the underlying representational mechanisms and computational operations are controversial. The empirical study reported here adopted and extended the Milner paradigm and produced interesting data toward a better understanding of the problem. In this section, we would like to discuss three issues related to the implications of the study and the future work.

First, the current study supports the argument that memory for spatial relationships can take multiple forms of representations, each encoded in a different frame of reference. Some of these representations may result from an early computation, often due to a direct perceptual experience in the early encoding phase. These representations can be encoded implicitly, such as the screen-based representations, or explicitly, such as the landmark-based representations and some object-based representations (e.g., spatial relationships between objects presented in consecutive trials). However, most of the object-based spatial relationships have to be inferred when necessary through a late computation, resulting in longer reaction time in the object-cued conditions. When multiple forms of representations are simultaneously available, a race model seems plausible. The processes supported by each representation compete with each other, and typically the one that affords fast response dominates. Eye movement results support this hypothesis

Second, the results from the current study are also consistent with the neuropsychological evidence that suggests there exist multiple spatial representational systems in the brain (e.g., Burgess et al., 1999; Wang et al., 2001). The PET imaging results from Milner and colleagues (1997) revealed that when object-location memory is retrieved, brain activity increases in the parahippocampal system, an area that is generally believed to subserve allocentric spatial representations.

Finally, while the current study generated interesting results, it is clear that to fully understand these results a detailed computational model is necessary. Questions about how multiple forms of spatial relationships are represented and how they interact can be better explored only when a computational model is developed. Efforts are being taken to develop such a model in the Act-R cognitive architecture (Anderson & Lebiere, 1998). The long-term goal is to develop a framework that can be used to model human spatial cognition in general, including object-location memory and spatial layout memory.

Acknowledgments

This work is supported by a grant from the Office of Naval Research (Grant No. N00014-01-1-0074).

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Press.

Burgess, N., Jeffery, K. J., & O'Keefe, J. (Eds.). (1999). *The hippocamal and parietal foundations of spatial cognition*. New York: Oxford University Press.

Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience, 22*, 319-349.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., Linenweber, M. R., Petersen, S. E., Raichle, M. E., Essen, D. C. V., & Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*, 761-773.

Hunt, E., & Waller, D. (1999). *Orientation and wayfinding: A review*.Unpublished manuscript, Arlington, VA.

Johnsrude, I., Owen, A. M., Crane, J., Milner, B., & Evans, A. C. (1999). A cognitive activation study of memory for spatial relationships. *Neuropsychologia, 37*, 829-841.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In K. F. Wender (Ed.), *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*. New York: Springer-Verlag.

Milner, B., Johnsrude, I., & Crane, J. (1997). Right medial temporal-lobe contribution to object-location memory. *Phil. Trans. R. Soc. Lond. B, 352*, 1469-1474.

Owen, A. M., Milner, B., Petrides, M., & Evans, A. C. (1996). A specific role for the right parahippocampal gyrus in the retrieval of object-location: A positron emission tomography study. *Journal of Cognitive Neuroscience, 8*, 588-602.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6*, 174-215.

Tversky, B. (2000). Remembering spaces. In F. I. M. Craik (Ed.), *The Oxford handbook of memory*. New York: Oxford University Press.

Wang, H., Johnson, T. R., & Zhang, J. (2001). *The mind's views of space*. Paper presented at the Fourth International Conference of Cognitive Science.

Combining Belief and Utility in a Structured Connectionist Agent Architecture

Carter Wendelken and Lokendra Shastri
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704
{carterw,shastri}@icsi.berkeley.edu

Abstract

The SHRUTI model demonstrates how a system of simple, neuron-like elements can encode a large body of *relational* causal knowledge and provide the basis for rapid inference. Here we show how a representation of utility can be integrated with the existing representation of belief, such that the resulting architecture can be used to reason about values and goals and thereby contribute to decision-making and planning.

Introduction

To understand how the brain creates the mind, one could work mainly from the top down, characterizing mental processes, or from the bottom up, trying to understand the capabilities of neurons and simple circuits. In developing the SHRUTI model we have pursued both these approaches simultaneously in order to understand how networks of neurons can perform complex cognitive tasks. In past work, we have demonstrated how such networks can make predictive and explanatory inferences with respect to a large body of causal knowledge. In this paper, we show how the SHRUTI architecture can be extended to represent and reason not only about beliefs but also about utilities, values and goals. The resulting model uses a single causal structure to seek explanations, make predictions, and identify expected utilities of world states and actions.

The SHRUTI architecture

First we present the basic elements of the SHRUTI architecture. The model is described in considerably more detail in [Shastri, 1999, Shastri and Ajjanagadde, 1993, Shastri and Wendelken, 2000]. SHRUTI is a neurally plausible (connectionist) model that demonstrates how a network of neuron-like elements could encode a large body of structured knowledge and perform a variety of inferences within a few hundred milliseconds. SHRUTI suggests that the encoding of relational information (frames, predicates, etc.) is mediated by neural circuits composed of *focal clusters* and that the dynamic representation and communication of relational instances involves the transient propagation of *rhythmic* activity across these clusters. A role-entity binding is represented in this rhythmic activity by the *synchronous* firing of appropriate cells.

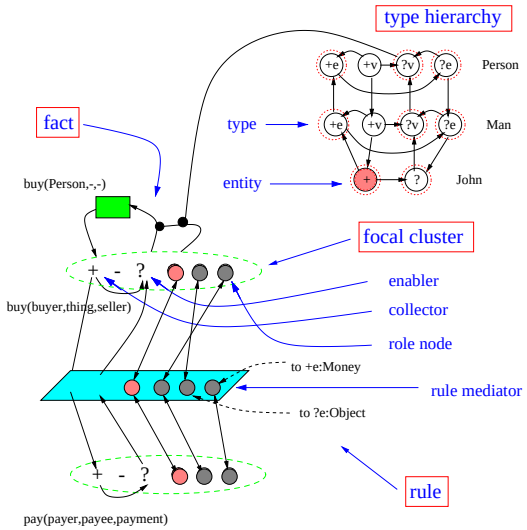


Figure 1: Diagram showing core elements of the SHRUTI model, including relational focal clusters, a fact, a rule, and a simple type hierarchy.

A focal cluster is a collection of nodes with varying functionality all subserving a common representation. A relational focal cluster consists of a positive (+) and a negative (-) collector node, an enabler (?) node, and role nodes. The activity of the positive (negative) collector node reflects the amount of evidence collected in support of belief (disbelief) in the given relation. Activity of the enabler (?) node reflects the strength with which information about the relation is being sought. A link from collector to enabler ensures that the system automatically seeks explanation for what it believes. Role bindings are represented by synchronous firing of relational role nodes with nodes in a connectionist type hierarchy. A relational cluster with active role bindings represents a relational instance. Rules are encoded with links that enable the propagation of rhythmic activity from one relational focal cluster to the next. Specifically, a rule is formed by linking the antecedent collector to the consequent collector, the consequent enabler to the antecedent enabler, and matching role nodes in both directions, through an intervening focal cluster

termed the *rule mediator*. Type restriction and instantiation of unbound variables are handled via connections between the rule mediator structure and the type hierarchy. Long-term facts are encoded in SHRUTI as temporal pattern matching circuits. *Episodic facts* (E-facts) are tuned to particular relational instances and represent specific knowledge or memories, while *taxon facts* (T-facts) are typically responsive to a range of relational activations and represent more general statistical knowledge about the world.

Probabilistic reasoning

Previous work has shown that the inferential behavior of SHRUTI does not, in most cases, stray far from a probabilistic ideal [Wendelken and Shastri, 2000]. With appropriate assignment of link weights, a simple rule structure can be shown to compute probabilities correctly in both the forward and backward direction. A set of evidence combination functions allows for flexible combination of evidence from multiple sources, while maintaining a relatively simple connectionist structure in which each antecedent communicates with the consequent via a single weighted link [Shastri and Wendelken, 1999]. Explaining away occurs via inhibitory interconnections between antecedents, so false patterns of circular reasoning are not introduced.

Inference in SHRUTI is essentially an anytime algorithm. Unlike in a belief net, responses to a query are generated almost immediately, based on the prior information stored for the queried relation. As inference is allowed to progress, early estimates are repeatedly refined as more and more evidence is brought in from further up or down the causal chain. In a neural system, the depth to which this search for evidence occurs would be limited, such that only evidence within a certain distance (along any casual chain) would be considered. Presumably, this depth could be modulated by attention or other factors. Importantly, this is a model which scales up naturally to large domains without performance loss (with reference to a parallel network of nodes and links).

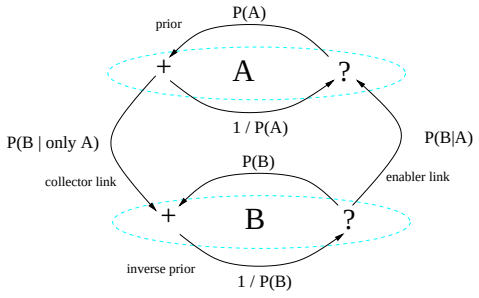


Figure 2: An illustration of the link weights for a simple rule (roles not shown). If B is believed true (+:B active with value 1.0) then activity at +:A will equal P(A—B)

Representing utility in SHRUTI

SHRUTI's representation of utility [von Neumann and Morgenstern, 1947] is analogous to its representation of belief. This consists primarily of a set of utility nodes associated with each relational focal cluster, reward facts denoting reward and punishment, value facts denoting learned utility values, probabilistically weighted utility-carrying connections between relations, and various modulatory mechanisms that affect utility flow differently in different situations. Thus belief and utility in SHRUTI are tightly integrated, sharing much of the same structure, and are not separate modules in any conventional sense.

Utility nodes

Recall that the representation of beliefs in SHRUTI is built around relational focal clusters, which contain several different types of nodes including positive and negative collectors, an enabler, and role nodes. Alongside these nodes representing belief, there are additional nodes representing associated utility. Thus there is a utility node tied to each of the two collectors, with activation range [-1,1]. These nodes are denoted by $+ and $-; positive activity of $+ ($-) indicates that positive utility value is associated with the truth (falsity) of the relation, while negative activation value of $+ ($-) indicates that negative utility is associated with the truth (falsity) of the relation. Links from each utility node to the enabler node ensure that whenever something is marked as having utility, it is automatically investigated by the system.

Activation of a relational utility node can indicate that reward is currently being experienced, or that it is expected. In either case, it reflects not only reward that is directly associated with its relation (as, for example, satisfying a sweet tooth is associated with eating cake), but also sources of reward that are more distantly related (such as potential weight gain). In this respect, the utility node is comparable to the value funcion of traditional reinforcement learning; however, utility node activity is transient and cannot by itself represent any permanent learned value associated with a relation instance (how this information is maintained will be described shortly). Instead, activity at a relational utility node reflects the combination of more permanent representations of value with the transient factors that make up current context.

Reward facts

Some relations have *reward facts* (R-facts) tied to them, designating certain relational instances as goals. Reward facts represent the source of reward and punishment in the system. Activation of a positive reward fact indicates the attainment (real or imagined) of some reward, while activation of a negative reward fact indicates the suffering (real or imagined) of some punishment. Like episodic facts in the belief system, reward facts are temporal pattern matching circuits that respond only when the specified set of role-fillers are active. In this case, activation of a relational collector along with synchronous activation of role nodes and appropriate type

node role fillers leads to activation of an associated fact node, which in turn leads to activation of that relation's appropriate utility node. Many different reward facts can be linked to a single relation; for example, a relation like $eat(x)$ might have associated with it positive reward facts such as $eat(Cake)$ as well as negative reward (punishment) facts such as $eat(Dirt)$.

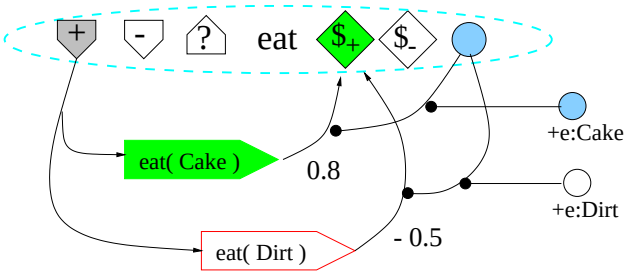


Figure 3: Two reward facts for the relation eat(x)

Research with rats and brain-stimulus reward suggests that both idiosyncratic and common currency representations of utility exist in the brain [Shizgal, 1998]. The representation of utility as activation values of relational utility nodes is a common currency representation which allows the activity of one node to be directly compared to the activity of another in order to guide decision-making. This is vital in order to allow successful decision making that takes into account disparate sources of reward and punishment. More domain-specific representations of utility must also exist, since the relative weighting of utilities from different sources can vary. The utility of eating, for example, is greatly influenced by degree of hunger, while the utility of play is not. Reward facts represent the connection between the common currency and the more domain-specific representations of utility. In order to model the latter, we allow that the weights on reward facts might vary depending on some internal state of the agent.

Value facts

While relational utility nodes represent value estimates in the current context, and reward facts represent basic goals, the task of storing learned value estimates rests with the *value facts*, or V-facts. Value facts are similar in form to reward facts, but instead of directly representing reward, they represent predicted future reward. For both value facts and reward facts, utility values are stored as link weights (specifically, as the weight on the link leading from the fact node to the associated relational utility node). The value fact associated with a relation plays a similar role to the value function in traditional reinforcement learning, and the update function for a value fact, depending as it does on local reward and maximization (or some other combination) of utility values of possible consequents, closely resembles the Bellman equation [Bellman, 1957]. Note, however, that value updates in SHRUTI depend only on activity of a few connected predicates, and not on the entire system state. Because of the similarity in the Bellman equation and SHRUTI's value-updating algorithm, the latter has been termed Causal Heuristic Dynamic Programming (CHDP) [Thompson and Cohen, 1999]. Like taxon facts in the belief system, value facts hold a statistical summary of past activity. They too are associative, meaning that matching of relational activity to the fact is stronger with more role matches, but is not necessarily blocked by a single role mismatch; this helps with generalization of value to multiple related instances.

A typical relation has many value facts associated with it, some very specific and some quite general. In this way, particularly important or salient items are explicitly encoded, whereas novel or less important items can fall back on more general representations. For the hypothetical agent for which eating cake is a paramount goal, $find(Cake)$ should be a highly-rewarding value fact. Eating other things may still be beneficial, so the more general $find(Food)$ may also appear as a weaker value fact; finding anything is more often good than bad, so even the most general value fact $find(Thing)$ might appear in the agent's internal representation. When the agent with these value facts happens upon a dollar bill, it will immediately perceive this as a positive situation according to the value of the $find(Thing)$ value fact. If finding money turns out to be significantly more rewarding than finding that average-value random thing, then this should be learned and explicitly represented as a new value fact.

Communication of utilities

Links connect utility nodes of different relations in the same way that they connect belief nodes. These links run parallel to the belief system connections, but in the consequent to antecedent (backward) direction. Figure 4 provides a simple illustration of these connections: for the rule $A \wedge B \Rightarrow C$, there are utility connections from the utility nodes of C, through the rule mediator, back to those of A and B. Weights on these connections are similar to the weights on the collector-collector links. Their purpose is to introduce probability into the calculations of value, such that the value estimate at some antecedent relation is based on both the value of its consequent (activity at its utility node) and the probability that it will be reached (weight on the connecting link). For the rule $A \Rightarrow C$, where the utility node of C ($:$C$) has a value of α, the utility node of A ($:$A$) should obtain the value $\alpha \times P(C|A)$.

This structure has the effect that assertion of a particular goal, via activation of a utility node, leads in the simplest case to assertion of its potential causes as subgoals, via spreading activation backwards along the causal chain. Belief in some relation, represented as activation of a collector node, leads to internal reward or punishment (activation of a reward fact) or recognition that such reward or punishment is likely (activation of a value fact) if there is an intact causal chain leading from that relation to some goal relation.

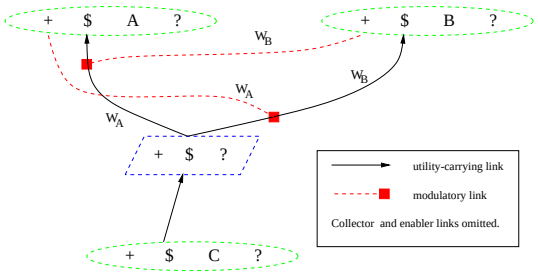


Figure 4: A diagram showing structure of utility connections for a two-antecedent rule.

Utility modulation

The model of utility propagation described so far is perfectly adequate for simple cause-effect relationships or chains of these. However, with multiple-antecedent or multiple-consequent rules, or with multiple rules involving a common relation, additional mechanisms must be introduced. Consider first a rule with two antecedents, such as $find(x) \wedge edible(x) \Rightarrow eat(x)$. The utility of finding something, which is derived from the utility of eating something, depends directly on whether or not that thing is edible. Thus, there should be an interaction between the two antecedents such that if $edible(x)$ is false, then the propagation of utility from $eat(x)$ to $find(x)$ is at least partially blocked. The reverse holds true as well - utility of a thing being edible depends not only on the utility of eating it, but also on whether or not it has been found.

The interaction described here is appropriate for the *and*-combination, but different interactions should occur when different relations hold between the antecedents and the consequent. For example, when antecedents are combined with an *or* function, then belief in the truth of one should tend to discount the propagation of utility to the others. In this case, when one cause is established, then redundant causes are no longer particularly useful. For the *avg* (weighted average) function, each antecedent contributes independently to the total, and so belief in the truth of one antecedent should have no impact on the perceived utility of another.

In general, the utility value at an antecedent relation should reflect the value of any associated consequents times the extent to which truth of the antecedent affects truth of the consequent. For a rule with antecedents A and B_1 through B_n and consequent C, this might be stated as "What difference does A make, in the context B_1 ... B_n, for the attainment of C", or in terms of probabilities, $P(C|A,B_1 = b_1,\ldots,B_n = b_n) - P(C|\neg A,B_1 = b_1,\ldots,B_n = b_n)$.

If the above expression is expanded for each different combination function, an interesting result is obtained, namely, that it is possible to compute it exactly for each different combination function using only the existing weight on the utility link along with a single additional weight from each associated antecedent. This relative simplicity of the resulting connectionist structure is important, since it lends plausibility to the notion that such a mechanism could be learned in the brain. Results for three combination functions, *and*, *or*, and *avg*, are shown below. The connectionist structure that computes these functions is shown in figure 4.

ECF	\$:A/\$:C
and	$W_A \cdot \prod_{i=1}^{n}(1-(1-b_i)W_{B_i})$
or	$W_A \cdot \prod_{i=1}^{n}(1-b_i W_{B_i})$
avg	W_A

Action focal clusters are given special treatment within this framework. Since the agent has control over whether or not an action is performed, activity of an action collector does not modulate the utility values flowing to any sibling antecedents. Also, while activity of an action's utility node indicates that the action is beneficial or harmful, activity of its enabler simply indicates that the action is potentially relevant.

Distribution and recombination

Just like beliefs, utilities from different sources must be combined. In general, the same approach is used here as with calculation of belief - a range of simple evidence combination functions are available and can be inserted into the connectionist structure as appropriate. Because many rewards are generally better than one, combination functions selected should generally have the property that a combined utility value is greater than any of the individual utilities; summation and *or* are two likely candidates. However, using such a combination function leads to a difficult problem when we allow multiple paths to exist between two relations. Consider the scenario, illustrated in figure 5 where exploration can lead to finding fruit or finding game, and that either of these consequents can lead to the goal relation of being able to eat. Utility associated with eating is propagated in full to both $findFruit$ and $findGame$ (assuming an *or*-combination and that neither is currently true), and from each it is further propagated back to *explore*. Now if *explore* has the sort of combination function described above, it can obtain a local utility value greater than that originating at the goal *eat*. This is clearly an unacceptable situation, and it comes from the fact that locally there is no information to distinguish between utility arriving from different sources (which should be added together) and utility values that originate from the same source (which should not).

One solution to this problem might be to disallow multiple paths between relations. Indeed, this is the solution adopted for belief nets to solve essentially the same problem. However, connections between relations are assumed to be learned from experience based only on local information; it is difficult to imagine any plausible mechanism by which learning of multiple paths could be inhibited when these provide the best fit for experience. Another solution would be to reduce the amount of utility distributed along each path according to the number

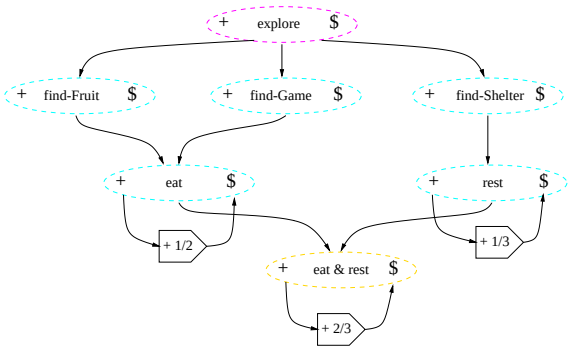


Figure 5: A proposed solution to the problem of utility combination.

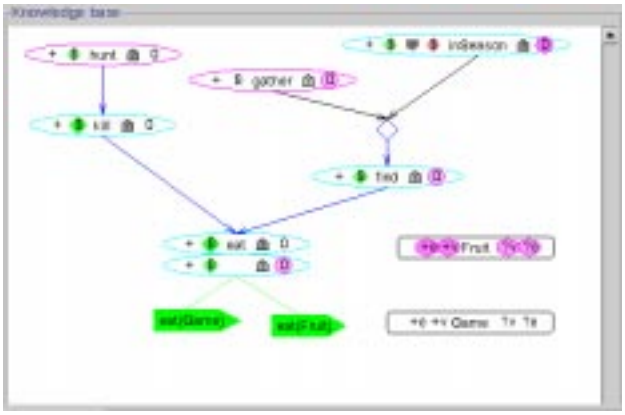


Figure 6: A captured moment from the simulation of the caveman scenario. Activity of -:inSeason blocks the propagation of utility to $:gather, resulting in a higher valuation of the hunt action.

of such paths; in this case that would mean that only half of utility at the *eat* relation is propagated to $findFruit$ and $findGame$. But this clearly leads to an underestimation of utility along each. Finally, we might abandon the use of summation and similar combination functions for utility and instead use something like *max*. This solves the problem at hand but also makes it impossible to productively combine multiple sources of utility. The "commom currency" representation of utility becomes somewhat modified; utilities can be directly compared within this framework but can no longer be directly combined. Instead, reward combinations must be explicitly represented in order to be used. This is illustrated in figure 5, where the basic goals *eat* and *rest* are supplemented by a combination goal *eat&rest*.

Simulation example

The operation of the network is illustrated here. The screen capture from the Shruti-Agent Simulator in Figure 6 shows a simple network representing the caveman's dilemma of whether to hunt or gather. Successful hunting yields the greatest reward (represented by the reward fact $eat(Game)$). Gathering, on the other hand, is more reliable, but only productive during the right season. We examine the propagation of beliefs and utilities around this simple network in detail. First, suppose that the caveman agent is hungry, and hence reward facts related to eating are fully active. Eating game or eating fruit are the current active goals of the system. Activity from the reward facts flows to multiple banks of the eat relation and from there back to $kill(Game)$ and $find(Fruit)$. The agent has realized that either killing game or finding fruit would be useful eventualities. Alongside the propagation of utility, a querying belief state is also being transmitted from relation to relation; this is represented in the activity of the enabler nodes. Since neither eventuality is thought to be true of the current world state, there is no competitive modulation of utility values; thus, $kill(Game)$ has the full 0.8 value from the $eat(Game)$ reward fact while $find(Fruit)$ has the full 0.6 from $eat(Fruit)$.

Utility value propagates further back to the *hunt* relation, this time modified by the uncertainty of hunting, such that $hunt(Game)$ has associated with it a utility of 0.4. In order for gathering fruit to be perceived as useful, the agent must have some knowledge that the fruit is in season. Suppose first that the query $inSeason(Fruit)$ is answered in the negative, either as a result of immediate knowledge of the agent or of further reasoning along paths not illustrated here. Then, according to the equation for distribution of utility values around an *and*-combination given above, and by means of a simple inhibitory mechanism, the flow of utility to the *gather* relation is blocked. Simlarly, If *inSeason* is uncertain, utility propagation to *gather* will be partially blocked. In either case, the *hunt* action, with a higher utility, will be favored. This is the situation illustrated in figure 6 and indicated by a numeral one in figure 7. If on the other hand the agent is reasonably certain that fruit is in season, then sufficient utility will propagate from the $find$ relation and *gather* will obtain a higher utility value than *hunt*, marking it as the preferred action.

Figure 8 illustrates a larger domain wherein the possibility of moving to a location where food can be found is included, as is the possiblility of being injured while hunting. When the assumption is made that *skilled* is true, (i.e. caveman is a skilled hunter), utility and belief propagate in this network such that $moveTo(RiverBank)$ (i.e. go to where the game is) is marked as a useful action.

Conclusion

We have demonstrated that SHRUTI, a neurally plausible model of knowledge representation and reasoning, can be enhanced to deal effectively with utilities, values, and goals. The resulting connectionist machinery is sufficient to guide an agent through a wide range of decision-making tasks, such as those illustrated in the previous examples. However, there is a class of decision problems for which the model presented here is inadequate.

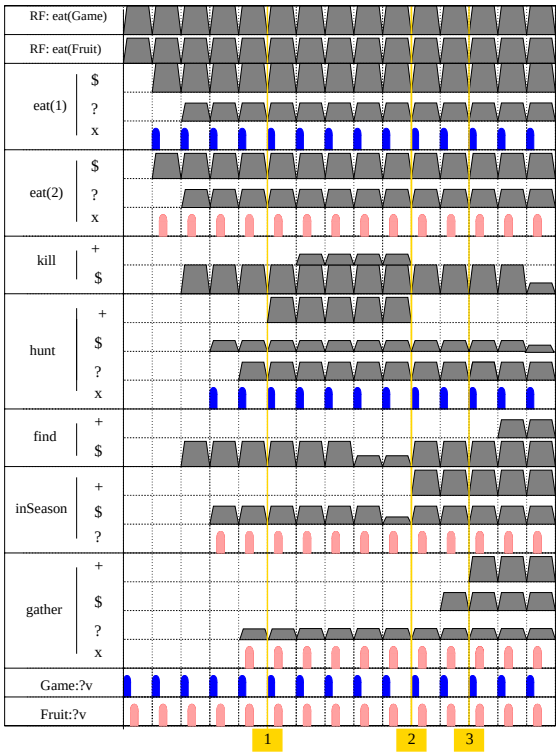


Figure 7: A stylized trace of node activations during execution of the caveman scenario.

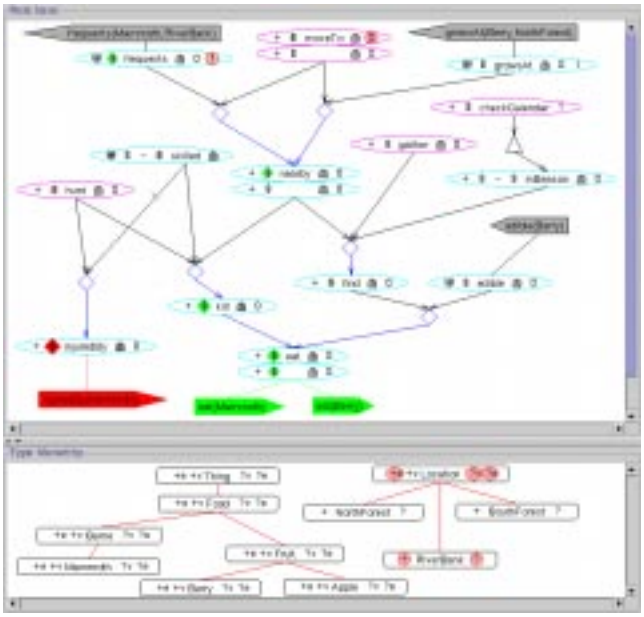


Figure 8: An expanded version of the caveman scenario.

In order to deal effectively with complex decision tasks, a measure of higher-level control must be introduced. Extensions to the model described here that enable it to perform complex decision-making and planning are described elsewhere [Wendelken and Shastri, 2002, Garagnani et al., 2002].

References

[Bellman, 1957] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

[Garagnani et al., 2002] Garagnani, M., Shastri, L., and Wendelken, C. (2002). A connnectionist model of planning via back-chaining search. In *Proc. 24th Conf. of the Cognitive Science Society*.

[Shastri, 1999] Shastri, L. (1999). Advances in SHRUTI - a neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, 11.

[Shastri and Ajjanagadde, 1993] Shastri, L. and Ajjanagadde, V. (1993). From simple associations to systematic reasoning. *Behavioral and Brain Sciences*, 16(3):417–494.

[Shastri and Wendelken, 1999] Shastri, L. and Wendelken, C. (1999). Soft computing in SHRUTI. In *Proc. 3rd Int. Symposium on Soft Computing*, pages 741–747, Genova, Italy.

[Shastri and Wendelken, 2000] Shastri, L. and Wendelken, C. (2000). Seeking coherent explanations - a fusion of structured connectionism, temporal synchrony, and evidential reasoning. In *Proc. 22nd Conf. of the Cognitive Science Society*, Philadelphia.

[Shizgal, 1998] Shizgal, P. (1998). *Foundations of hedonic psychology: Scientific perspectives on enjoyment and suffering*, On the neural computation of utility: implications from studies of brain stimulation reward.

[Thompson and Cohen, 1999] Thompson, B. and Cohen, M. (1999). Naturalistic decision making and models of computational intelligence. In A. Jagota et al. editors, *Connectionist Symbol Processing: Dead 0r Alive?*, volume 2 of *Neural Computing Surveys*, pages 1-40. http://www.icsi.berkeley.edu/ jagota/NCS.

[von Neumann and Morgenstern, 1947] von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press.

[Wendelken and Shastri, 2000] Wendelken, C. and Shastri, L. (2000). Probabilistic inference and learning in a connectionist causal network. In *Proc. 2nd Int. Symposium on Neural Computation*.

[Wendelken and Shastri, 2002] Wendelken, C. and Shastri, L. (2002). Decision-making and control in a structured connectionist agent architecture. In *submitted*.

Computer Augmented Psychophysical Scaling

Robert L. West (robert_west@carleton.ca)
Department of Psychology, Department of Cognitive Science, Carleton University, Ottawa, Canada

Ronald L. Boring (rlboring@ccs.carleton.ca)
Department of Cognitive Science, Carleton University, Ottawa, Canada

Stephen Moore (srmoore@chat.carleton.ca)
Department of Cognitive Science, Carleton University, Ottawa, Canada

Abstract

In this paper we present a methodology for improving the reliability of observers in magnitude estimation tasks by using the computer to augment the cognitive components of the task.

Psychophysical scaling is the study of how to accurately measure perception. More specifically, the goal is to find methodologies that allow people to accurately communicate the magnitudes of specific dimensions of conscious experience, such as brightness, loudness, temperature, and heaviness. Psychophysical scaling can also be used for measuring the magnitude of subjective experiences such as level of happiness (e.g., West & Ward, 1988). The goal of psychophysical scaling is to find the mathematical functions that map the magnitudes of external stimulus dimensions to the conscious perception of magnitude. This enterprise is extremely useful for both scientific and applied research.

Numerous different scaling techniques exist. However, our focus is on magnitude estimation, which is one of the most commonly used psychophysical methods. Magnitude estimation (ME) was invented by Stevens (1956) and involves exposing subjects to a set of stimuli and asking them to match the magnitude of a particular dimension of each stimulus to the magnitude of a number. This is repeated for multiple trials to provide multiple responses for each stimulus value. To avoid the influence of outliers, the median or the geometric mean of the responses for each stimulus value is calculated. Numerous studies have shown that plotting these values against the stimulus values produces functions that are closely approximated by power functions. This is known as, the Power Law, or, Stevens' Law.

The form of the power law is,

$$R=KS^B,$$

where R is the observer's response, S is the stimulus magnitude, B is the exponent value, and K is a constant. Logging both sides of the equation produces,

$$Log(R)=B\cdot Log(S)+Log(K),$$

which is a straight line with B estimated by the slope and K by the intercept. The exponent, B, can be interpreted as a metric for stimulus compression. This reflects the fact that people use a power function or something closely approximating a power function to compress stimuli, just as audio and video files can be compressed to save on bandwidth. In fact, audio and video compression go unnoticed to the extent that the compression function maps onto the human compression function for the same stimuli. Generally speaking, in ME the goal is to put as few restrictions on the observer's choice of numbers as possible. Often free ME (e.g., see Zwislocki & Goodman, 1980) is used, in which observers are instructed to match the perceived magnitude of the stimulus to whatever number seems most natural. This is quite different from the common psychological practice of imposing scales on people. The reasons for this are both theoretical and practical. From a mathematical standpoint, if any two stimuli are set equal to any two responses then you have determined what the exponent value must be. Thus, if an observer uses the lowest value on a scale to match the lowest perceived magnitude and the highest value to match the highest perceived magnitude, the power function exponent has been fixed. To get around this one could assign a value to a middle value on the scale and not impose a top end or bottom end, but this has been shown to produce confusion and poor results (Stevens, 1975). However, the fact that peoples' backgrounds cause them to use different ranges of numbers in their responses is not a problem as these differences are captured by the K constant (since response range is usually not of interest, K values are usually not reported).

ME can be considered a special case of cross modal matching (CMM). In cross modal matching, the observer adjusts the magnitude of one stimulus dimension to match the magnitude of another stimulus dimension (e.g., adjusting the brightness of a light to match the loudness of a tone). Like ME, CMM results also produce power functions. Furthermore, ME and CMM results are consistent in that they can be used to predict each other (e.g., the ME exponents for brightness and loudness can be used to predict the exponent relating brightness and loudness in a CMM experiment). Also, both the power functions and the specific exponent values found through ME are consistent with ratio scaling experiments, in which magnitude scales are derived by asking observers to set or report ratios between stimuli. These approaches to scaling are known as direct scaling techniques (Stevens, 1971).

Problems

ME forms the foundation for a potentially accurate and consistent way of measuring perceived magnitude. However,

ME, as well as the other methods with which it is consistent, have been found to be limited in terms of accuracy. Although a considerable amount of evidence indicates that subjects do obey the power law (see Stevens, 1975; and Bolanowski & Gescheider, 1985 for reviews), the specific exponent values that Stevens found could not be reliably replicated with the level of accuracy one would expect for measuring sensory processes in normal, healthy individuals. Exponent values vary considerably across individuals in the same experiment (e.g., Algom & Marks, 1984; Luce & Mo, 1965; Marks & J. C. Stevens, 1965; Rule & Markley, 1971; Wanschura & Dawson, 1974; Logue, 1976) and can also vary across time within individuals (Logue, 1976; Marks, 1991; Teghtsoonian & Teghtsoonian, 1983). Stevens also found strong individual differences, which he attributed to various response biases. Stevens' solution was to treat response bias as a random factor and to average across individuals to get the true exponent value (Stevens, 1971). However, Marks (1974) reviewed the literature and found that in addition to individual differences, the average value of the exponent varies significantly across ME experiments done in different labs. These results suggest that the distribution of individual response biases differs from lab to lab, indicating that they cannot be treated as random. Indeed, it is well known that some labs get systematically higher or lower exponent values than others, suggesting that response bias can be influenced by minor procedural differences.

In addition to limitations on accuracy, ME results are not consistent with partition scaling (also called interval scaling) results for prothetic continua, although they are consistent for metathetic continua (according to Stevens, metathetic continua are more qualitative in nature, e.g., pitch or hue; while prothetic continua are more quantitative in nature, e.g., loudness or brightness; see Stevens, 1971 for a more detailed discussion). Partition scaling includes a variety of techniques that require observers to partition the stimulus continuum. Category scaling (e.g., 1 to 5 scales; 1 to 7 scales; scales partitioned by word labels such as good, bad, very bad) is a form of partition scaling, and is by far the most commonly used scaling technique. The problem is that partitioning techniques tend to produce power functions with lower exponents than direct scaling techniques (Stevens, 1971). Stevens' argument for accepting the results of direct scaling techniques rather than partition scaling techniques was that partition scaling is less direct because it requires the extra step of partitioning the stimulus range, and that the discrepancy can be attributed to biases introduced by the partitioning task (see Stevens, 1971). However, like direct scaling, partition scaling also produces excessive variability (Marks, 1974).

Because of these problems, psychophysical scaling still has issues concerning reliability and validity. In terms of the power law, the validity problem can be stated as the problem of which, if any, method will produce the "true" exponent. The reliability problem is that we do not have a methodology that we can use to make reliable statements about individual differences or inter-lab differences in exponent values. In our opinion, the reliability problem needs to be solved before tackling the validity problem. Our work attempts to address this. The reliability problem can be broken down into a theoretical and a practical problem. The theoretical problem is that if bias differs from individual to individual and within individuals across time, we cannot get reliable measurements without being able to somehow predict or control the bias. The practical problem is that even if we solve the theoretical problem, to be useful we need a system that does not require huge numbers of responses from individuals who have limited amounts of time and limited attention spans. We have focused our efforts on the reliability issue and attempted to solve both of these problems by cognitively augmenting our human observers through the use of computerized support.

Bias

The process of magnitude matching can be represented in the following way (Marks, 1991),

$$M(S) = R$$

where S is the stimulus magnitude, R is the response magnitude, and *M* is the function relating them. The *M* function can then be decomposed into an initial, perceptually based function, P, that is the same (or highly similar) across healthy, normal individuals; followed by a function, C, representing cognitively imposed constraints that account for the excessive variability:

$$M(S) = C(P(S))$$

Since most psychophysicists study perception, the emphasis has been on getting rid of C so as to reveal P. Considerable effort has been expended in this enterprise. Approaches taken include trying to identify the sources of C to avoid or control for them (see Poulton, 1989 for a review); trying to minimize C by encouraging observers to respond naturally, without thinking about it too much (e.g., Stevens, 1975; Zwislocki & Goodman, 1980); trying to measure C and then partial it out (e.g., Berglund, 1991); trying to stabilize C across scaling tasks to get rid of intra-observer variability (e.g., J. C. Stevens & Marks, 1980); and avoiding C by developing methods that allow the scale to be derived from judgments of "greater than" or "less than" for paired stimuli sets (e.g., Schneider, 1980, 1988). However, success in these endeavors has been limited and a consensus as to the best method is lacking.

Our approach to dealing with C was quite different. As cognitive scientists, we viewed the variability of C as the inevitable consequence of the sort of problem presented to the observers, i.e., create and maintain a consistent mapping from P to R. The problem of creating a mapping may or may not be difficult but it is definitely open ended, with very few constraints on the solution. Also, the problem of maintaining the mapping once it has been created could tax the limits of working memory. In fact, Petrov and Anderson (2000) and Petrov (2001) were able to model a number of different bias effects associated with various factors using the ACT-R (Anderson & Lebiere, 1988) architecture to model the memory processes involved. Based on this view, our approach has been to attempt to eliminate these effects by

providing computerized support for establishing and maintaining the scale.

Constrained Scaling

Constrained scaling is a form of magnitude estimation (i.e., observers report numbers to match stimulus values). The goal of constrained scaling is to calibrate observers to the same C function before scaling the stimulus dimension of interest, similar to the way that physical measuring instruments are calibrated before use (Ward, 1991). Constrained scaling (West, Ward, & Khosla, 2000) is based on four claims about C: (1) that C is cognitively penetrable, (2) that C is heavily influenced by ad hoc decisions made early in the scaling process, (3) that the C process makes heavy demands on working memory which leads to instability across the task, and (4) that C is independent of the perceptual modality being judged (i.e., if the perceptual modality is changed it does not directly cause a change in C, although an interruption in the process could disrupt and indirectly alter C). Provided these assumptions are true, it should be possible to train observers to use a predetermined C function, and to support the maintenance of it in memory by refreshing it through a computerized feedback system.

Constrained scaling involves two phases, a learning phase and a test phase. In the learning phase, feedback is used to train observers to respond to a standardized set of stimulus magnitudes according to a predetermined response scale. This is done across several trials by presenting learning set stimuli and having the observer rate the perceived magnitude by entering an R value. On the interface we have been using this can be done by entering a value in a text box or by using a specially designed scroll bar that allows the observer to move the slider by units of 10, 1, 0.1, and 0.01. The scroll bar runs from 0 to 100 (although the observers are instructed that they may enter R values above 100). After this the observer clicks a button marked, "OK," and their R value is replaced with the correct R value. The point of this is to build C functions that are the same across observers and to give them the practice they need to become familiar with it. Provided that P is highly similar across observers, training the observers so that they all correspond to the same function relating S and R, implies they have the same C function, although it is possible that the details of how they cognitively implement and maintain the C function may differ.

The choice of the scale to be learned should be based on learnability and the mathematical desirability of the scale. Similar to West et al (2000), we used a power function with an exponent similar to what would be found using ME (i.e., we accept, to some extent, Stevens' argument that free ME produces scales that people find more natural to use) and K was set so that the scale range was approximately from 1 to 100 (as we believe this is a range that people are familiar with).

Research has shown that, with feedback on each trial, people can learn these scales quite accurately (King & Lockhead, 1983; Koh & Meyer, 1991; Koh, 1993; West & Ward, (1994); Marks, Galanter, & Baird, 1995). However, we have found that once the feedback is taken away, people start to drift off of the learned scale. Therefore, during the test phase the learned scale is presented on every second trial followed by feedback, so that the form of the scale is constantly refreshed in memory. On the alternate trials, test stimuli, different from the learned stimuli, are presented without feedback. The observers are instructed to use the learned scale to respond to the test stimuli as well as the learned stimuli. They are also told that the response range of the test stimuli may be greater or less than the response range of the test stimuli.

This general approach was used in West et al (2000) and the results were compared to other psychophysical methods. In that study, the learned scale stimuli were 1000 Hz tones between 32 dB and 99 dB, spaced at 1 dB intervals. The learned scale responses were numbers from 1 to 100 related to the stimulus magnitudes by a power function with an exponent of 0.600 (taken from the International Organization for Standardization, 1959). The test stimuli were 65 Hz tones and light brightness. The results, a full discussion of the psychophysical meaning of the results, and a comparison to other methods is presented in West et al (2000). Here we will just point out that constrained scaling produced very low levels of inter-observer variability compared to ME and CMM. Furthermore, the only method that we could find that produced similar low levels of inter-observer variability was conjoint measurement as applied to combined pairs of tones (Schneider, 1988). However, this methodology exploits the fact that, under the right conditions, loudness is additive for two tone combinations, which limits its application to auditory stimuli. It also requires a large number of trials.

Scaling Video Frame Rates

The results from West et al (2000) clearly demonstrated that training observers and using external means to constantly refresh their memory produces highly reliable scaling results. This indicates that arbitrary decisions about how to structure a scale and insufficient resources for maintaining the scale in memory are the primary source of inter-observer variability in direct scaling. However, it was still unclear how observers use the feedback to maintain a representation of the scale. We speculated that observers memorized a limited number of perceived magnitude/response pairs and interpolate to get responses in-between (see Ward & West, 1988, for an example of people using this strategy in a similar type of task). If this is the case then constrained scaling should work if the observers are only supplied with feedback on a limited number of S/R pairs instead of many pairs covering the whole range (as in West et al 2000).

We applied this methodology in a study designed to look at the effect of content type on the perception of frame rate in video clips. Specifically, we were interested in whether or not speed of movement in the clip alters the perception of frame rate. To do this we began with a pilot study using magnitude matching. Magnitude matching is a version of ME in which two different stimuli are alternately presented in the same scaling task (J. C. Stevens & Marks, 1980). In this case we used a fast paced video clip and a slow paced video clip. The results, averaged across observers, indicated that the exponent for frame rate was approximately 0.90. No

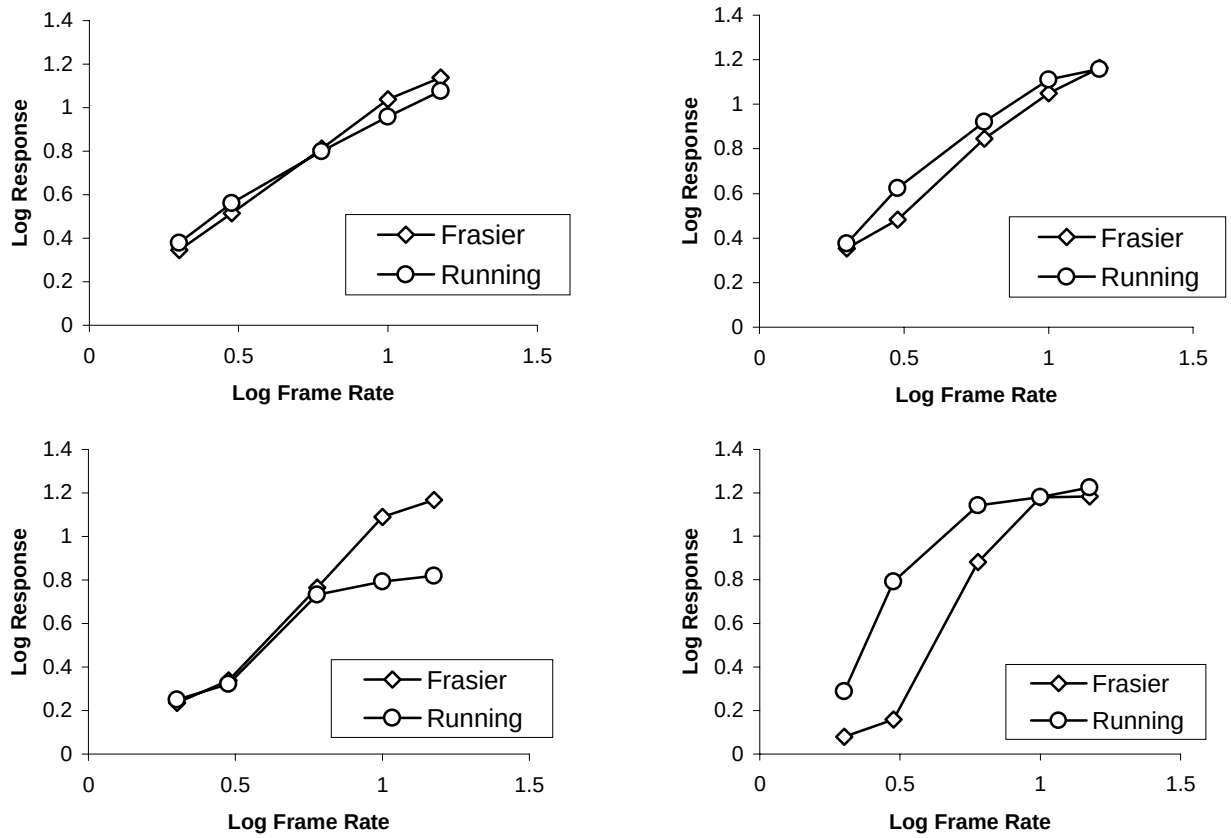


Figure 1. Psychophysical functions for four representative observers. The top row shows two observers who obeyed the power law and the bottom row shows two observers who deviated from it

significant effects for content were found (Boring, West, & Dillon, 2000).

For the constrained scaling experiment we used only five stimulus levels for training (2, 3, 6, 10, and 15 frames per second). Observers were taught, using feedback, to respond to these frame rate levels according to a power function with an exponent of 0.90. The observers were given 50 trials to learn the scale and the stimuli were presented randomly. The content of the video clip was moderate in speed (medium speed hip hop dancing).

During the test phase, the observers were instructed that the same hip-hop clips would be presented with feedback on every second trial, and that on the alternative trials a different video clip would be presented. The Observers were told to respond to the other clip using the same scale they learned for the hip-hop clip, but that the frame rate levels would not necessarily be the same and that there would be more than five versions of the new clip. This was actually not true; the test stimuli were generated using the same frame rate levels as the learning stimuli. However, the observers did not know this as the stimuli were spaced less than one JND (just noticeable difference) apart. We mislead our observers so that they would be open to responding with the whole range of responses. The observers all completed two test phase sessions, one using a fast content clip (children running) and one using a slow content clip (a clip from the Fraser show of Fraser talking). The order of the sessions was counterbalanced and another 50 trials of training were presented in-between. All stimuli were presented in random order.

Results

As in West et al (2000), we found that constrained scaling did not produce outliers, so we used mean response values for scaling the responses instead of medians. From a visual inspection of the graphed functions from the test phase trials it was clear that four observers produced functions with relatively large nonlinear trends (see Figure 1). This is actually not uncommon in ME (Luce, & Mo, 1965). The normal procedure would be to throw them out or to average across them, along with the functions of the other observers. However, since we are interested in individual differences, we note that these four were less able than the other

observers to exploit the external scaling aids offered by constrained scaling. This indicates that individual differences in strategy, cognitive ability, and/or effort still play a role. Since these deviations were not unusually large by ME standards we analyzed the data both with them in and with them out. The remaining six observers produced functions that could reasonably be treated as linear (see Figure 1).

West et al reviewed 14 studies that provided individual observer results for ME and CMM, and calculated the standard deviation divided by the mean for the individual exponent values from each study. As a basis for comparison we took these values and calculated the mean, which was 0.333, the standard deviation, which was 0.080, and the 0.05 confidence interval, which was plus or minus 0.042. Even with the four linearly deviant observers included, the mean of the individual exponent values divided by the standard deviation was 0.190 for the Fraser clip and 0.150 for the children running clip, significantly lower than what would be expected with ME or CMM. Without the four deviants included, the mean divided by the standard deviation was 0.076 for the Fraser clip and 0.047 for the children running clip. These values were similar to the mean divided by standard deviation values found by West et al (2000) using constrained scaling (these values were 0.045, 0.066, and 0.152).

Also, because of the low variability we were able to detect a small but significant difference in exponent values both with ($P < 0.01$) and without ($P = 0.01$) the four linearly deviant observers, indicating that the exponent values for the slower video were higher than the exponent values for the faster video. This finding illustrates the advantage of having more precise ways of measuring perceived magnitudes (note, since the purpose of this paper is to examine the cognitive aspects of scaling, we will not discuss why this difference might exist).

Discussion

These findings replicate the West et al (2000) finding that augmenting the cognitive abilities of the observer can significantly reduce inter-observer variability and, more generally, supports the four theoretical assumptions behind constrained scaling (see above). The results also support the hypothesis that people can maintain scales in memory by memorizing a limited number of S/R pairs. By providing support to remember five S/R pairs we significantly reduced inter-observer variability to a level comparable to that found in West et al (2000), who provided feedback for a large number of responses. Other strategies may also be possible but, at the very least, this result shows that providing support for remembering a small number of S/R pairs can provide a significant advantage.

In terms of strategy, examining the actual responses that the observers made revealed that they took a category scaling approach. Two observers used the five R values they had learned almost exclusively. The other observers added only a few new R values and some stopped using one or two of the learned R values. The new R values also tended to be used as categories, that is, they were used repeatedly. This was quite different from the West et al (2000) observers who responded with a wide range of R values. From this it would appear that observers prefer to continue using a response strategy that resembles the one they were trained on. This may be due to observers inferring that the number of test stimuli will be similar to the number of learning stimuli, or it may be that teaching them to respond in a particular way creates cognitive structures that are not amenable for doing the task in other ways.

The fact that observers were able to respond accurately using a category scaling strategy, on a scale that was determined using ME, suggests that training and providing feedback to observers eliminates the factors that cause category scaling to produce different results from ME. This result is quite promising as it suggests that providing external support for the scaling process can wipe out methodologically induced biases.

Conclusions

These results provide compelling evidence that cognitively augmenting observers can substantially increase the reliability of psychophysical scaling, which is particularly important for measuring and studying individual differences and small group differences (as in this study). We also believe that this approach will eventually provide a means for assessing the validity of the scales as well. This is based the assumption that the further a learned scale is from the natural scale, the more cognitive resources will be required to maintain the mapping (C) from P to R (for some evidence of this see Marks, Galanter, & Baird, 1995; West et al, 2000). To improve further we need to better understand the strategies available to observers, and how to more effectively intervene to support the scaling process. Eventually, we hope that this approach will lead to psychophysical measurement techniques that have the same unambiguous status as physical measuring techniques.

References

Algom, D., & Marks, L. E. (1984). Individual differences in loudness processing and loudness scales. *Journal of Experimental Psychology: General*, *113*, 571-593.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berglund, M.B. (1991). Quality assurance in environmental psychophysics. In S.J. Bolanowski Jr.& G.A. Gescheider (Eds.) *Ratio Scaling of Psychological Magnitude: In Honor of the Memory of S. S. Stevens.* Hillsdale, New Jersy: Lawrence Erlbaum Associates, Publishers.

Bolanowski, S. J., & Gescheider, G. A. (1991). *Ratio Scaling of Psychological Magnitude: In Honor of the Memory of S. S. Stevens*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Boring, R.L., West, R.L., & Dillon, R.F. (2000). Evaluation of framerate quality for different video content types. Poster presented at the CITO Digital Media Research Review, Toronto, Ontario, February 15, 2000.

International Organization for Standardization (1959). *Expression of physical and subjective magnitudes of*

sound [ISO/R-131-1959(E)]. Geneva: International Organization for Standardization.

King, M. C., & Lockhead, G. R. (1981). Response scales and sequential effects in judgement. *Perception & Psychophysics*, *30(6)*, 599-603.

Koh, K. (1993). Induction of combination rules in two dimensional function learning. *Memory and Cognition*, *21(5)*, 573-590.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17(5)*, 811-836.

Logue, A. W. (1976). Individual differences in magnitude estimation of loudness. *Perception & Psychophysics*, *19(3)*, 279-280.

Luce, D. R., & Mo, S. S. (1965). Magnitude estimation of heaviness and loudness by individual observers: A test of a probabilistic response theory. *The British Journal of Mathematical and Statistical Psychology*, *18(2)*, 159-174.

Marks, L. E. (1974). On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as science. *Perception & Psychophysics*, *16(2)*, 358-376.

Marks L. E. (1991). Reliability of magnitude matching, *Perception & Psychophysics*, *49(1)*, 31-37.

Marks, L. E., Galanter, E., & Baird, J. C. (1995). Binaural summation after learning psychophysical functions for loudness. *Perception & Psychophysics, 57*, 1209-1216.

Marks, L. E., & Stevens, J. C. (1965). Individual brightness functions. *Perception & Psychophysics*, *1*, 17-24.

Petrov, A. (2001). Fitting the ANCHOR model to individual data: A case study in Bayesian methodology. Fourth international conference on Cognitive modeling (pp. 175-180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Petrov, A. & Anderson, J. R. (2000) ANCHOR: A memory based model of category rating. *Proceedings of the 22nd annual confernce of the cognitive science society* (pp. 369-374). Hillsdale, NJ: Lawrence Erlbaum Associates.

Poulton, E. C. (1989). *Bias in quantifying judgements*. London: Lawrence Erlbaum Associates, Publishers.

Rule, S. J., & Markely, R. P. (1971). Subject differences in cross-modality matching. *Perception & Psychophysics*, *9*, 115-117.

Schneider, B. (1980). Individual loudness functions determined from direct comparisons of loudness intervals. *Perception & Psychophysics*, *28*, 493-503.

Schneider, B. (1988). The additivity of loudness across critical bands: A conjoint measurement approach. *Perception & Psychophysics, 43,* 211-222.

Stevens, J. C. (& Marks, L. E. (1980). Cross-modality matching functions generated by magnitude estimation. *Perception & Psychophysics*, 27, 379-389.

Stevens, S. S. (1956). The direct measurement of sensory magnitudes – loudness. *American Journal of Psychology, 69*, 1-25.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social Prospects*. New York: A Wiley-Interscience Publication.

Stevens, S. S. (1971) Issues in psychophysical measurement. *Psychological Review, 78, 5*, 426-450.

Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception & Psychophysics, 33*, 203-214.

Ward, L. M. (1991). Associative measurement of psychological magnitude. In S. J. Bolanowski & G. A. Gescheider (Eds.), *Ratio Scaling of Psychological Magnitude: In Honor of the Memory of S. S. Stevens* (pp. 79-100). Hillsdale, New Jersy: Lawrence Erlbaum Associates, Publishers.

Ward, L. M., & West, R. L. (1998). Modelling human chaotic behaviour: Non-linear forecasting analysis of logistic iteration. Nonlinear Dynamics, Psychology, and Life Sciences, 2, 4, 261-281.

West, R. L., & Ward, L. M. (1994). Constrained Scaling. In L. M. Ward (Ed.) *Fechner Day 94*. Vancouver: International Society for Psychophysics.

West, R. L., & Ward, L. M. (1998). The value of money: Constrained scaling and individual differences. Fechner Day: Proceedings of the Fourteenth Annual Meeting of the International Society for Psychophysics.

West, R. L., Ward, L. M., & Khosla, R. (2000). Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. Perception & Psychophysics, 62(1), 137-151.

Wanschura R. G., & Dawson, W. E. (1974). Regression effect and individual power functions over sessions. *Journal of Experimental Psychology, 102(5)*, 806-812.

West, R. L., & Ward, L. M. (1994). Constrained Scaling. In L. M. Ward (Ed.) *Fechner Day 94*. Vancouver: International Society for Psychophysics.

Zwislocki, J. J., & Goodman, D. A. (1980). Absolute scaling of sensory magnitudes: A validation. *Perception & Psychophysics, 28*, 28-38.

Adapting to a Response Deadline in Categorization

A. J. Wills (a.j.wills@ex.ac.uk)
School of Psychology, University of Exeter,
Perry Road, Exeter. EX4 4QG. England

Abstract

The effect of a response deadline on categorical decisions was investigated. Time available for response was manipulated in the test phase, along with stimulus difficulty. Effects of these manipulations were observed in response accuracy, and in the mean, standard deviation and skew of the reaction times. The effects observed demonstrate that participants responded to the deadline in an adaptive manner - reducing their reaction time to long-latency decisions whilst leaving short latency decisions relatively unaffected. A simple connectionist model of categorical decisions (Wills & McLaren, 1997) is shown to account for this behavior.

Introduction

Categorization is a basic and essential cognitive function. Our ability to engage it has been well studied, and a number of different theories of the underlying processes have been proposed (e.g. Ashby & Gott, 1988; Gluck, 1991; Nosofsky, 1986; Nosofsky, Palmeri & McKinley, 1994). At first, attempts to quantitatively fit models of categorization to empirical data concentrated on categorization accuracy. However, in recent years, models which have the potential to predict reaction time distributions in categorization have been developed and evaluated (e.g. Ashby, 2000; Maddox, Ashby & Gottlob, 1998; Lamberts, 2000; Nosofsky & Palmeri, 1997; Wills & McLaren, 1997).

This paper focuses on the effects of imposing a response deadline on a) participants' response accuracy and b) the nature of their reaction time distributions. It has been known for some time that categorical decisions made under time pressure may be different to those made without time pressure (eg. Smith & Kemler Nelson, 1984). More recently, this avenue of research has been developed by investigation of the effects of time pressure with more complex stimuli (e.g. Lamberts, 1995; Palmeri & Blalock, 2000) coupled with formal modeling of the results found (e.g. Lamberts, 1995).

It is worth considering Lambert's (1995) study in a little more detail as it provides one motivation for the current work. At one level, the results found are intuitive. In these experiments, Lamberts employed a simple deadline procedure. Participants first learned, in the absence of time pressure, to categorize artificial stimuli (schematic faces) into two categories. Following this training, participants had to categorize test stimuli before a given deadline (e.g. 1600ms from stimulus onset). Failure to respond in time resulted in an error tone, followed by the presentation of the next stimulus. Participants were informed about the time available for response, which changed at regular intervals. In one experiment, the deadlines employed were 600ms, 1100ms, 1600ms and no deadline. Participants were less accurate at shorter deadlines. Interestingly, the effect was stimulus specific, with some stimuli being considerably more affected by time pressure than others. Lamberts proposed a particular formal model of this effect (the "Extended Generalized Context Model" or EGCM, Lamberts, 1995) and showed that it provided a good fit to the accuracy data.

Time pressure and reaction time

Lamberts' experiments reveal another result. In his experiments, categorization in the absence of a response deadline takes approximately 1500ms (Lamberts, 1995, experiment 2). As the stringency of the deadline increases, so the mean reaction times decrease, with categorization under a 600ms deadline taking about 450ms. In other words, categorical decisions appear to take considerably less time when there is time pressure than when there is not. This is, of course, intuitively obvious. The interest, from the perspective of the current paper, is that there seem to be at least three distinct reasons why it might happen. When considering the following, it is important to remember that the descriptions relate to observed reaction time distributions - they are not statements about underlying process:

Non-selective adaptation: The participant reacts to the imposition of the deadline in a manner that decreases all reaction times in the distribution by a fixed amount. As a consequence, mean of the distribution will drop, but the standard deviation and skew will be unaffected.
Selective, linear adaptation: The participant reacts to the imposition of the deadline in a manner which decreases all reaction times in the distribution by a fixed factor (i.e. $RT_{deadline} = f \times RT_{no\ deadline}$). As a result, the mean and standard deviation of the distribution will drop, but the skew will be unaffected.
Selective, non-linear adaptation: The participant reacts to the imposition of the deadline in a way that cannot be characterized as non-selective, or selective, non-linear, by the definitions above. Changes in the mean, standard deviation, and skew of the distribution may all be observed.

Demonstrating adaptation to a deadline

It therefore seems clear that to distinguish between these explanations, one must estimate changes in the mean, standard deviation and skew of the reaction time distribution produced by imposition of a deadline. Whilst the experiment reported in this paper is by no means the first to investigate the effects of a deadline on categorization accuracy and reaction time, previous work has had at least one of the two following limitations:

Missing data artifact

In a number of studies (e.g. Lamberts, 1995; Lamberts & Brockdorff, 1997; Palmeri & Blalock, 2000) it is possible that the changes observed are an artifact of the data collection procedure. In a response deadline procedure, longer-than-deadline responses typically result in a "time out" error and hence no data about reaction time is available for that trial. As a direct consequence, mean response time is lower than it would have been without a deadline. The same problem applies to studies that compare two different deadlines. In experiments where percentage of time-outs is reported by condition, they can been seen to increase as the response deadline becomes more stringent.

One solution to this problem is to use a "response signal" procedure (e.g. Lamberts, 1998) where participants are instructed to respond as soon as possible after they get a signal to do so. Another solution (see e.g. van Zandt, Colonius & Proctor, 2000) is to provide a "too slow" signal after the response has been made.

A third possibility is to use the standard response deadline procedure, but only evaluate responses that fall below a certain percentile of the reaction time distribution (with time-outs being considered as the slowest trials). The largest number of time-outs made at any level of time pressure, by any participant, to any of the test stimuli, determines this percentile. For all conditions and stimuli, only responses that fall below that fixed percentile are considered. It is therefore important to keep the percentage of time-outs low so a reasonable amount of data is still available for analysis. It is this final possibility that is employed in the current study.

Insufficient information

The three possibilities for adaptation outlined above can only be distinguished if one has estimates for the mean, standard deviation, and skew of the reaction time distributions. Recently, many studies of categorization have begun to report reaction time distributions in detail (e.g. Maddox & Ashby, 1996). However, categorization studies that employ time pressure as a manipulation tend to concentrate on categorization accuracy, and may also report mean reaction times. Data from different tasks, such as perceptual matching, show that the mean, standard deviation, and positive skew all reduce in response to increasing time pressure (van Zandt et al., 2000).

Given the absence of appropriate information, it was decided to perform a short empirical study that would have the potential to discriminate between the three types of adaptation to a response deadline which have been outlined. This is followed by a demonstration that a particular model of categorical decisions (Wills & McLaren, 1997) can mimic the results found. Implications of both the empirical and the theoretical investigations for categorization research are then discussed.

Experiment

The current experiment had two phases. In the training phase, participants were presented with novel, abstract stimuli paired with either the category label "A" or the category label "B". In the test phase that followed, participants had to decide the category membership (A or B) of unlabelled stimuli either without time pressure, with a 2500ms time limit for each decision or with a 1000ms time limit for each decision (a between-participants manipulation).

Whilst these deadlines may appear relatively lax compared to the reaction times observed in some classification tasks, previous work (e.g. Wills & McLaren, 1997) indicates they represent a fairly high level of time pressure for participants with relatively little experience of the complex stimuli employed.

Method

Participants and apparatus

The participants were 44 adults, mainly undergraduate students. The experiment was in two different, quiet cubicles on two Acorn RISC PC computers, with 14" color monitors. Participants sat 1 meter from the screen.

Stimuli

Each stimulus was a collection of twelve different small pictures (hereafter "elements") in a 4.5cm by 3.5cm rectangle outline, arranged on an invisible four-by-three grid (see Figure 1 for an example stimulus). Every stimulus contained twelve elements drawn from the pool of thirty-six elements we have used in previous experiments (see Jones, Wills & McLaren, 1998, p.37). At the beginning of the experiment, and separately for each participant, 12 elements from the pool were randomly designated as category A elements, and a different 12 as category B elements. The remaining 12 elements were not used for that participant.

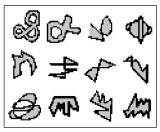


Figure 1: An example stimulus

Training stimuli

Sixty training stimuli (thirty from each category) were created for each participant. Each training stimulus was constructed by starting with all 12 elements characteristic of a particular category (e.g. category A elements for a category A training stimulus). Then, each element in the training stimulus underwent a 10% chance of being replaced by an element chosen from the other set (e.g. replaced by a category B element in the case of a category A training stimulus). Choice of replacement elements was random within the constraint that no element could occur more than once in any given stimulus. The position of elements within a stimulus was randomly determined for each stimulus presented, with the constraint that exactly one element occurred at each location in the four-by-three grid.

This method of stimulus construction produces training examples which are composed predominately of elements characteristic of a particular category but which also exhibit considerable variability.

Test stimuli

Test stimuli were designed to vary in difficulty of categorization. Given the nature of the training stimuli, the correct response to a test stimulus is to categorize it as an "A" if it contains more A elements than B elements, and as a "B" otherwise. A number of previous experiments have demonstrated that as the difference between the number of A elements and the number of B elements increases in a stimulus of this type, the probability of a correct classification also increases (Jones et al, 1998; Wills & McLaren, 1997). Test stimuli in this experiment are therefore described in terms of their difference scores (the absolute value of the number of A elements minus the number of B elements).

All stimuli contained twelve elements, so there are seven possible difference scores and hence seven levels of difficulty. The seven differences scores are 12, 10, 8, 6, 4, 2 and 0, which are denoted as having a difficulty level of 1,2,3,4,5,6 and 7 respectively. Twenty examples at each of the first six levels of difficulty were created for each participant. The specific elements used to create each test stimulus were chosen randomly within the constraint provided by the difference score, and the constraint that stimuli in which category A elements were more numerous that category B elements should occur with the same frequency as stimuli in which category B elements were more numerous than category A elements. As in the construction of the training stimuli, the position of elements within a test stimulus was randomly determined, and no element was allowed to occur more than once in any given stimulus.

Ten examples of stimuli with a zero difference (difficulty level 7) were also generated for each participant. However, as there is no correct answer for such stimuli, performance on them is not analyzed in this paper.

Procedure

Participants were allocated to one of three groups that differed only in the time allowed for decision in the test phase. These groups are referred to hereafter as the *1000ms, 2500ms* and *No-deadline* groups. Sixteen participants were allocated to the 1000ms group, sixteen to the 2500ms group, and twelve to the no-deadline group.

The sixty training stimuli were presented sequentially and in a random order. Each example was presented for five seconds in the center of the monitor accompanied by the appropriate category label (presented as a large capital A or B in an outline rectangle immediately to the right of the stimulus). The stimulus and the category label were then replaced with mid-gray rectangles that stayed on the screen for two seconds and were followed by the next example. Participants were not required to respond in any way in this first phase of the experiment but were asked to concentrate on the examples shown as they would later be asked to classify new, unlabelled examples. This training procedure has proved effective for stimuli of this type in a number of other experiments (Jones et al, 1998; Wills & McLaren, 1997).

The training phase was followed immediately by the test phase. There were 130 stimuli in this phase (see "Stimuli" section) which, again, were presented sequentially and in a random order. Participants classified each stimulus as an "A" or a "B" by pressing either the "X" or ">" key on the computer keyboard. The allocation of keys to responses was counter-balanced across participants.

In the 1000ms and 2500ms conditions, participants were told that they only had 1 second or 2.5 seconds to make each decision. If they did not respond within this time interval, the stimulus was replaced by the phrase "TIME OUT!" in 2cm high letters. After a five second count-down and a two-second pause, the next stimulus was presented. This time-out procedure was designed to be as salient as possible in order to keep the total number of time-outs low.

Results

Accuracy and mean reaction time data from the no-deadline condition have been reported previously (Wills & McLaren, 1997). All other data are novel.

In the 2500ms condition, 2.87% of trials were timed-out. The figure was 4.84% in the 1000ms condition.

Whilst both rates are relatively low, there were significantly more time-outs in the 1000ms condition, $t(30) = 2.41$, $p < 0.05$. All participants in this experiment made at least sixteen responses before the deadline at each level of stimulus difficulty. Therefore the four slowest responses made by each participant at each level of stimulus difficulty were disregarded in the following analyses (see Introduction for an explanation of this procedure). Time-out trials were counted as the slowest possible responses. For the remaining data, the accuracy, and the mean, standard deviation and skew, for each level of stimulus difficulty and for each participant were calculated.

This data set was subjected to a series of mixed-model ANOVAs, with one within-participants variable (Difficulty, 6 levels) and one between-participants variable (Deadline, 3 levels). A significance level of .05 was set for all analyses. Figures 2 and 3 summarize the data set by providing across-participant averages.

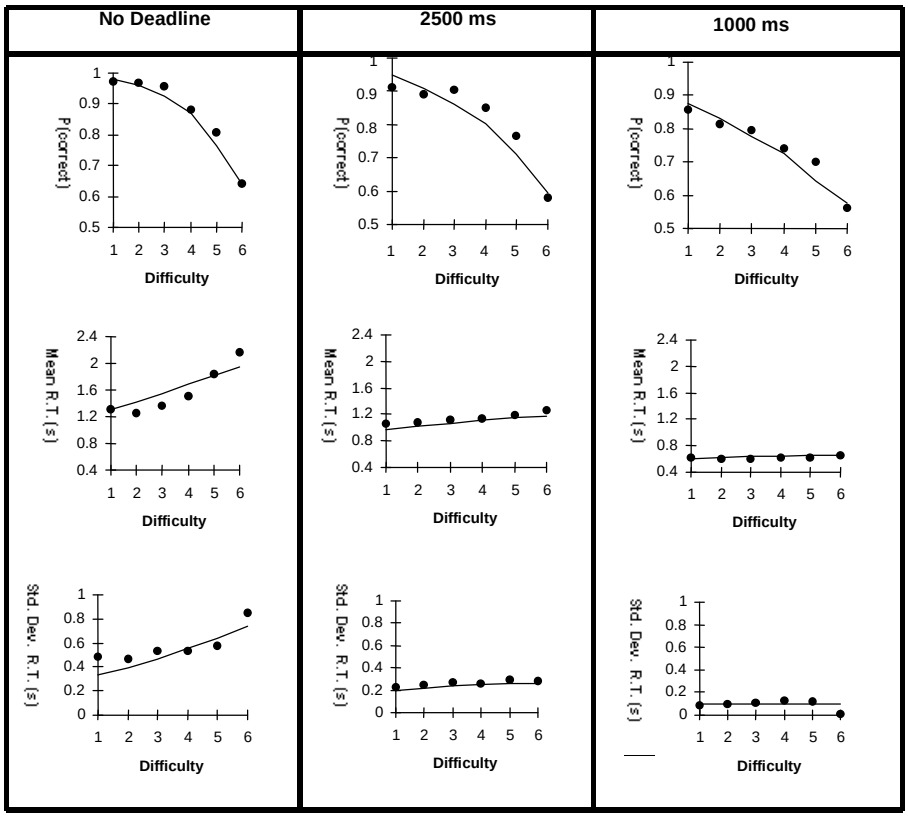


Figure 2: Accuracy, mean reaction time, and standard deviation of reaction time as a function of stimulus difficulty (arbitrary units) and response deadline (milliseconds). Filled markers indicate empirical data. Lines indicate modeling results.

Figure 3 also averages across stimulus difficulty because (as will be seen in a moment) there was no significant effect of stimulus difficulty on skew.

Response accuracy was adversely affected by both stimulus difficulty, $F(5, 205) = 48.1$, and deadline $F(2, 41) = 6.45$. These two factors did not interact significantly, $F(10, 205) < 1$. Mean reaction time increased with stimulus difficulty, $F(5, 205) = 20.19$, and decreased with time pressure, $F(2, 41) = 12.92$. The effect of stimulus difficulty was less pronounced with increasing time pressure, as evidenced by a significant interaction term, $F(10, 205) = 9.32$.

The standard deviation of reaction times increased with stimulus difficulty, $F(5, 205) = 5.77$, and decreased with increasing time pressure, $F(2, 41) = 17.36$. The effect of stimulus difficulty was less pronounced with increasing time pressure, as evidenced by a significant interaction term, $F(10, 205) = 8.45$.

The skew of reaction times decreased with increasing time pressure, $F(2, 41) = 19.52$. However, stimulus difficulty had no significant effect, $F(5, 205) = 1.31$, and the interaction term was non-significant also, $F(10, 205) < 1$. The no-deadline condition shows significantly positive skew, $t(11) = 5.06$, whilst the 1000ms condition shows significantly negative skew, $t(15) = 2.76$. The 2500ms condition showed no significant skew, $t(15) = 0.98$.

Modeling

Wills & McLaren's winner-take-all (WTA) model, like many process models of categorization, assumes that the evidence a presented stimulus is the member of a particular category is represented by a single number or *magnitude term.* In this simulation, the magnitude term for category x (denoted v_x) is $M \times c$, where c is the number of category x elements the presented stimulus contains, and M is a free parameter. Such a relationship sufficiently describes the output of a feature-based, single-layer, delta-rule network taught to classify the stimuli (see Wills & McLaren, 1997 for more details).

The model is illustrated in Figure 4. A single unit represents each category. The magnitude terms for each category are passed to these units as input activation. The output activity of each unit is a function of the total

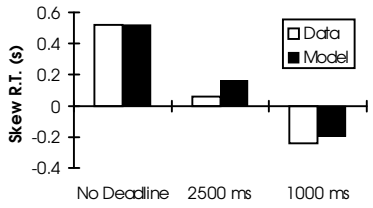


Figure 3: Skew of reaction times as a function of response deadline (milliseconds).

input it receives. Specifically, the output activation of unit *i* on update *c* is determined by

$$o_{i,c} = \frac{o_{i,c-1} + En_{i,c}}{1 + En_{i,c} + D} \quad \text{(where } n_{i,c} > 0\text{)}$$

$$o_{i,c} = \frac{o_{i,c-1}}{1 - En_{i,c} + D} \quad \text{(where } n_{i,c} \leq 0\text{)}$$

$n_{i,c}$ is the total input to unit *i* on update *c* and *E* and *D* are constants representing the rate of excitation and decay within the unit.

In addition to the magnitude-term inputs, each unit has a fixed excitatory connection to itself and fixed inhibitory connections to other units. These connections cause the units to "compete" with one another until only one has non-zero activation. Grossberg (1976), and many others since, have employed similar, neurally-inspired decision-making systems.

The total input to a unit *i* on update *c* is given by

$$n_{i,c} = r_{i,c} + o_{i,c-1} - \sum_{j \neq i} o_{j,c-1}$$

where $r_{i,c}$ is the noisy input produced by the magnitude term v_i. The noise in these particular simulations had a range of $+N$ to $-N$, and a rectangular distribution. Superimposed on this noise is the constraint that $r_{i,c}$ cannot exceed one or fall below zero.

The first unit to produce an activation greater than *S* is assumed to cause the execution of its corresponding response. The number of cycles the unit takes to exceed *S* represents decision latency, with each cycle representing exactly *T* seconds.

The model employed includes a number of simplifications, including the assumption that noise is rectangular and that non-decisional components of the categorization process take a fixed T_{res} seconds. Neither simplification is central to the operation of the model - similar predictions can be derived from a model with a variable T_{res} and Gaussian noise. However these simplifications have the advantage of considerably speeding the search of parameter space.

The model described above has seven parameters - *N*, *E*, *D*, *M*, *S*, *T* and $T_{res.}$ The basis of the model's predictions is that time pressure reduces the value of *S*, so *S* was assigned a different value for each of the three between-participant conditions of the experiment. In all previous applications of the model, it has been assumed that $E = 2D$, and this assumption is continued in the current simulation. *T* is not a parameter of the model in any important sense, as its only purpose is to convert from one arbitrary unit of time (cycles) to another (seconds). Hence, model fitting involves the manipulation of seven free parameters, from which predictions for 57 data points are to be derived.

Model fitting proceeded via a grid-search procedure. The range and steps of the parameters were *N* (0.1→3, step 0.1), *E* (0.01→0.05, step 0.01 and 0.05→0.5, step 0.05), *M* (0.01→0.08, step 0.01), and *S* (0.3→0.7, step 0.05) for each of the three *S* parameters, with the constraint that *S* did not increase as response deadline decreased. 10,000 decisions were simulated for each permutation of parameters and for each stimulus difficulty level. The cycles-to-decision in each set of 10,000 decisions were then place in rank order, and the 2,000 slowest decisions discarded (in order to mimic the data deletion performed on the empirical data).

This collection of simulated decisions was then employed to produce a set of predictions for each of the permutations of the parameters *N, E, M* and the three *S* parameters, S_{ND}, S_{2500} and S_{1000}. The relationship between cycles-to-decision and seconds was then estimated for each set of decisions via linear regression

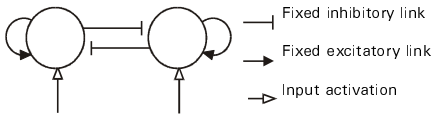


Figure 4: The winner-take-all model

of the 18 empirical mean reaction times to the 18 mean cycles-to-decision. The gradient of the line gives the conversion factor *T* whilst the intercept provides the value for T_{res}.

Once *T* and T_{res} were determined for a set of predictions, they were employed to convert each of the 180,000 simulated decision latencies into seconds. Calculations of mean reaction time, reaction time standard deviation and reaction time skew for each stimulus difficulty level in each of the three conditions were then performed using standard formulae.

Scaled root mean square deviations (SRMSD) were used to assess closeness of fit. Scaling was performed by multiplying each empirical data point and each prediction by a factor *s*. SRMSD was calculated separately for accuracy, mean RT, RT standard deviation, and RT skew predictions. The scaling factors employed were 2, 0.5, 1 and 1 respectively. Total SRMSD was taken to be the sum of these four SRMSDs. The set of parameters providing the best overall fit were as follows *N:* 2.6, *E:* 0.03, *M:* 0.04, S_{ND}*:* 0.55, S_{2500}*:* 0.50, S_{1000}*:* 0.40, T_{res}*:* 0.033. This is the fit shown in Figures 2 and 3. One cycle of the model was estimated by linear regression to be 0.014 of a

second. Cycles-to-decision predicted over 95% of the variance of mean reaction times in this regression (r^2 = 0.953). The SRMSD for accuracy predictions was 0.061, for mean reaction time it was 0.047, for reaction time standard deviation it was 0.062 and for reaction time skew it was 0.063.

Discussion

Imposition of a response deadline decreased the mean, standard deviation and skew of reaction times in a categorization task. From this information about the distribution, one can conclude that these participants adapted to the response deadline in a selective, non-linear manner (as defined in the Introduction). This is a result which, if found to be general, would need to be accommodated by formal models of categorical decisions. The fact that one of the reaction time distributions to be fit is negatively skewed might be considered as a particular source of concern, as categorization have almost uniformly been fit to distributions with some degree of positive skew in the past.

In the space available, it was not possible to evaluate whether all current models of categorical decision have the potential to accommodate the results found. Instead, it was demonstrated that one particular model of categorical decisions (Wills & McLaren, 1997) can mimic the pattern of results found. Wills & McLaren's model is (in approximate terms) a connectionist implementation of a random-walk process (e.g. Laming, 1968). As such, its follows the same basic principles as a variety of other accounts of categorical decision, including stochastic forms of decision-bound theory (Ashby, 2000), extensions of EGCM that can model reaction times (Lamberts, 2000), and Nosofsky & Palmeri's (1997) exemplar-based random walk model. It therefore seems likely that many contemporary models of categorization are capable of accounting for the sort of adaptation to a response deadline observed in this study.

References

Ashby, F. G. (2000). A stochastic version of general recognition theory. *J. Math. Psych., 44*, 310-329.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *JEP: LMC , 14*(1), 33-53.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2*, 50-55.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. *Biological Cybernetics, 23*, 121-134.

Jones, F. W., Wills, A. J., & McLaren, I. P. L. (1998). Perceptual categorization: Connectionist modelling and decision rules. *Quart. J. Exp. Psy., 51B*(3), 33-58.

Lamberts, K. (1995). Categorization under time pressure. *JEP: General, 124*(2), 161-180.

Lamberts, K. (1998). The time course of categorization. *JEP: LMC, 24*(3), 695-711.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review., 107*(2), 227-260.

Lamberts, K., & Brockdorff, N. (1997). Fast categorization of stimuli with multivalued dimensions. *Memory & Cognition, 25*(3), 296-304.

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* London: Academic Press

Maddox, W. T., & Ashby, F. G. (1996). Perceptual separability, decisional separability and the identification-speed classification relationship. *JEP: HPP, 22*(4), 795-817.

Maddox, W. T., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Percept. & Psychophys., 60*(4), 620-637.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorisation relationship. *JEP: Genral, 115*(1), 39-57.

Nosofksy, R. M., & Palmeri, T. J. (1997a). An exemplar-based random walk model of speeded classification. *Psych. Review, 104(*2), 266-300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*(1), 53-79.

Palmeri, T. J., & Blalock, C. (2000). The role of background knowledge in speeded perceptual categorization. *Cognition, 77*, B45-B57.

Smith, J. D., & Kemler Nelson, D. G. (1984). Overall similarity in adults' classification: The child in all of us. *JEP: General, 113*, 137-159.

van Zandt, T., Colonius, H., & Proctor, R. (2000). A comparison of two response time models applied to perceptual matching. *Psych. Bull. & Rev., 7*(2), 208-256.

Wills, A. J., & McLaren, I. P. L. (1997). Generalization in human category learning. *Quart. J. Exp. Psy., 50A*(3), 607-630.

Acknowledgments

Financial support for this research was provided by Emmanuel College, Cambridge, and by the ESRC. Thanks to Ian McLaren, Stephen Monsell, Stian Reimers, Thomas Palmeri and Koen Lamberts for helpful comments.

A Vector Model of Causal Meaning

Phillip Wolff (pwolff@memphis.edu)
Department of Psychology, 202 Psychology Bldg.
Memphis, TN 38152 USA

Matthew Zettergren (mzttrgrn@memphis.edu)
Department of Electrical and Computer Engineering, Engineering Bldg.
Memphis, TN 38152 USA

Abstract

This paper proposes a new model of causal meaning, the Vector Model, which formalizes a model of causation based on Talmy's notions of force dynamics (Wolff, Song, & Driscoll, 2002). In the Vector Model, the concepts of CAUSE, ENABLE and PREVENT are distinguished from one another in terms of force vectors, their resultant and the relationship of each force vector to a target vector. The predictions of the model were tested in two experiments in which participants saw realistic 3D-animations of an inflatable boat moving through a pool of water. The boat's movements were completely determined by the force vectors entered into a physics simulator. Participants' linguistic descriptions of the animations were closely matched by those predicted by the model given the same force vectors as those used to produce the animations. Our model may have implications for the semantics of causal verbs as well as the perception of causal events.

Introduction

This research investigates people's notions of causation as reflected in their use of causal verbs. We approach this problem by formulating a model of causal meaning that defines causal concepts in terms of relationships between force vectors, their resultant and a target position vector.

We begin by noting two key problems for models of causal meaning. First, such models must be able to distinguish the concept of CAUSE from the concept of ENABLE. We say, for example, the wave (and not the keel) *caused* the sailboat to rock, while the keel (and not the wave) *enabled* the sailboat to rock. The precise way in which these two notions differ has been difficult to specify. Contributing to this difficulty is the fact that the two concepts cannot be distinguished in terms of necessity or sufficiency (Cheng & Novick, 1991; Goldvarg & Johnson-Laird, 2001). In the above example, neither the wave nor the keel alone is sufficient, but both may be necessary for the boat's rocking to occur. Several solutions to this challenge have been proposed, but most have not escaped criticism (see Cheng & Novick, 1991; Goldvarg & Johnson-Laird, 2001; Wolff, Song, & Driscoll, 2002).

A second key problem for models of causal meaning concerns how the concept of CAUSE is represented in expressions that refer to specific instances of causation. Many models of causation define causation in terms of probabilities (e.g., Cheng, 1997; Cheng & Novick, 1991; Glymour, 2001). Such models are well suited for explaining the meaning of generic statements of causation, that is, statements about what is typically the case in multiple occurrences of a particular event, as in *Heavy snowmelt causes rivers to flood*. What these theories do not handle well are expressions that refer to a single instance of causation, as in *The heavy snowmelt caused the Colorado to flood*. Sentences describing single instances express what is definitely true of a particular event, not what is typically true of many. Moreover, such sentences are incompatible with the non-occurrence of the result (e.g., flooding), but if causation is inherently probabilistic, such non-occurrences cannot be strictly ruled out (Goldvarg & Johnson-Laird, 2001).

In some theories of causation, the concept of CAUSE is defined in such a way that it can be used in descriptions of singular causation. For example, according to Michotte (1963), causation is inferred from the perception of a transfer of motion from one ball to another–an "ampliation of motion" (p. 143). A related proposal is that CAUSE is inherently based on the idea of force and that the occurrence of CAUSE involves a mechanism by which this force is transmitted (Ahn & Kalish, 2000; Shultz, 1982). While these theories specify properties that could be predicated of a single event (and are highly related to the proposal we make in this paper), they do not provide us with a clear solution to the first problem of causal meaning: how the notion of CAUSE might be distinguished from the notion of ENABLE.[1] Both CAUSE and ENABLE presumably involve the transference of force.

In this paper, we propose a model of causal meaning that addresses these two problems. This model represents a formalization of the Force Dynamic Model described in Wolff, Song and Driscoll (2002; also Wolff & Song, 2001). In the next section, we describe

[1] Counterfactual theories of causation face related problems (see Spellman & Mandel, 1999)

the Force Dynamic Model as well as some of the empirical evidence in support of it. We then turn to a description of its formalization.

The Force Dynamic Model of Causation

A theory of force dynamics was first proposed by Talmy (1988), and has been elaborated by several other researchers (Jackendoff, 1991; Kemmer & Verhagen, 1994; Pinker, 1989; Robertson & Glenberg, 1998; Siskind, 2000; Verhagen & Kemmer, 1997). From a force dynamic perspective, the concept of CAUSE is one member of a family of concepts that include the concepts of ENABLE and PREVENT, among others. With each of these concepts, there are two key players: an affector and a patient.[2] Differences among the concepts are captured in terms of various patterns of tendency, relative strength, rest, and motion.

The Force Dynamic Model specified in Wolff, Song and Driscoll (2002) combines two of Talmy's (1988) core dimensions (Tendency & Result) with a dimension suggested by Jackendoff (1991).[3]

Table 1: The Force Dynamic Model's representations of CAUSE, ENABLE, & PREVENT

	Patient Tendency for Result	Affector-Patient Opposition	Occurrence of Result
CAUSE	N	Y	Y
ENABLE	Y	N	Y
PREVENT	Y	Y	N

As shown in Table 1, this model specifies that the concepts of CAUSE, ENABLE, and PREVENT can be captured in terms of 1) the tendency of the patient for the result, 2) the presence of opposition between the affector and patient, and 3) the occurrence of the result. In causing situations (see 1a), for example, the tendency of the patient, the boat, is not for the result, heeling. But because the tendency is opposed by the affector, the result, i.e., heeling, occurs.

(1) a. The blast caused the boat to heel.
b. Vitamin B enables the body to digest food.
c. The rain prevented the tar from bonding.

In enabling situations, as in (1b), the tendency of the patient, the body, is for the result, to digest food. This tendency is not opposed by vitamin B. Rather, vitamin B assists in the realization of this tendency, which leads to the occurrence of a result. In situations involving preventing, as in (1c), the tendency of the patient, the tar, is towards the occurrence of the result, bonding, but this tendency is opposed and blocked by the affector, and as a consequence, the result does not occur.

Evidence in support of the Force Dynamic Model

As indicated in Table 1, the Force Dynamic Model predicts that each concept shares one feature in common with each other concept: ENABLE and PREVENT both involve patients with a tendency for the result; CAUSE and PREVENT both involve opposition; and CAUSE and ENABLE both lead to results. The model implies, then, that the three concepts should be equally similar in meaning. Therefore, if we were to plot these concepts in a similarity space in terms of the verbs that encode them, they should reside roughly equally distant from one another. In fact, this is exactly what we found when we asked people to sort 48 sentences from the British National Corpus that contained 23 periphrastic causative verbs (i.e. verbs that pattern syntactically and semantically like the verb *cause*, e.g., *make*, *enable* and *prevent*) and submitted their sorts to a multidimensional scaling program[4] (Wolff et al., 2002).

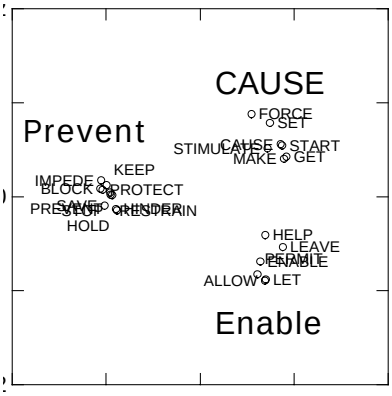


Figure 1: MDS solution of periphrastic causative verbs

As Figure 1 shows, the periphrastic causative verbs in English fall into three categories: a CAUSE category that includes the verbs *cause*, *force*, *get*, *make*, *set* and *stimulate*, an ENABLE category that includes the verbs *allow*, *enable*, *help*, *leave*, *let*, and *permit*, and a PREVENT category that includes the verbs *block*, *hinder*, *hold*, *impede*, *keep*, *prevent*, *protect*, *restrain*, *stop*. Importantly, the clusters associated with these three concepts reside roughly equally distant from one another, just as predicted by the Force Dynamic Model. We have replicated these results for specific and generic statements of causation. These results, along with several rating studies, lead us to believe that the Force Dynamic Model captures the primary semantic dimensions underlying the periphrastic causative verbs, and the verb *cause* in particular.

[2] We use the more familiar terms affector and patient instead of antagonist and agonist as originally used in Talmy (1988).
[3] In Talmy (1988) nearly all interactions involve opposition while in Jackendoff (1991) this parameter is allowed to vary.
[4] Multidimensional scaling is a procedure that locates items in space so that their distances in that space reflect as closely as possible their measured inter-item (dis)similarities.

The Vector Model of Causation

In the Vector Model, the notions of tendency, opposition (here, concordance), and result are represented as force vectors, their resultant and the relationship of each force vector to a target position vector. The model is described below for physical interactions in which the patient has no initial velocity. However, it is assumed that it could be extended to situations in which the patient does have an initial velocity (and, hence, momentum). It is also assumed that the model could be extended to cover non-physical kinds of causation (e.g., social, psychological).

In our description, all vectors are typed in boldface font; **P•T** denotes the dot product of the vectors **P** and **T**; ||**P**|| denotes the magnitude of **P**.

In the case of physical causation, **A** represents a vector that specifies the force exerted on the patient by the affector; **P**, any force produced by the patient to move itself, or in the absence of such a force, its weight (e.g., force pulling it toward the earth) and/or resistance to motion due to frictional forces; **O**, the vector representing the summation of the remaining *other* forces acting on the patient [5] and **R**, the resultant force acting on the patient based on the vector addition of **A**, **P** and **O**. An example configuration is shown in Figure 2.

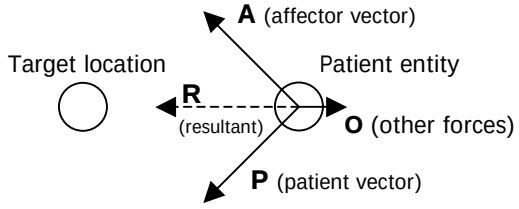


Figure 2: Forces associated with the affector, **A**, patient, **P**, and other forces, **O**, combine to produce a resultant force, **R**, in the direction of a target.

The target's location is specified in terms of a position vector, **T**. When the target and patient are points, **T** simply begins at the patient and ends at the target, as shown in Figure 3.

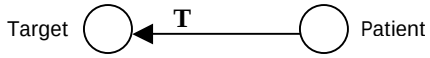


Figure 3: The target's location is specified by a position vector **T**.

In the more general case in which the target is represented by an area, the target's location would be specified by a set of real 1- or 2-dimensional position vectors, such that every vector from the patient's position to a point that could be considered a part of the target would be an element of that set. [6]

For this particular version of the Vector Model, we assume that all of the forces are constant with respect to time and space (i.e., $\partial/\partial t[\mathbf{Z}(x,y,t)] = \partial/\partial x[\mathbf{Z}(x,y,t)] = \partial/\partial y[\mathbf{Z}(x,y,t)] = \mathbf{0}$ where **Z** is any force in this model), the patient has no initial velocity, and ||**P**|| and ||**T**|| > 0.

The main dimensions of the Vector Model are defined in Table 2 for patients and targets that can be construed as points, and where ||**A**|| > 0.

Table 2: Dimensions underlying the Vector Model

Dimension	Formal Definition
Tendency (of patient for the target)	Angle between **P** and **T** = 0°
Concordance (of affector & patient)	Angle between **A** and **P** = 0°
Result	Angle between **R** and **T** = 0°

Rationale for the definitions *Tendency* - If the patient has a tendency for the target, then the direction of its force vector will coincide with the direction of the position vector **T**. Thus, the angle between the vector **P** and **T** will be 0°. A test for this possibility can stated with respect to the dot product of **P** and **T**. Specifically, when the patient has a tendency for the target, **P•T** = ||**P**||*||**T**||[7], and when it does not, **P•T** < ||**P**||*||**T**||.

Concordance – Concordance concerns the similarity of the force vectors associated with the affector and the patient. If the affector and patient exert forces (on the patient) in the same direction, then they are considered to be in concordance. In a similar fashion to tendency, concordance can be defined with respect to the dot product, but this time between **P** and **A**. Specifically, when the affector and patient are in concordance, **P•A** = ||**P**||*||**A**||, and when they are not, **P•A** < ||**P**||*||**A**||.[8]

Result – As with tendency and concordance, occurrence of a result can be defined in terms of the similarity between two vectors, but this time between **R** and **T**. When the angle between **R** and **T** is 0°, the result will occur, assuming all of the forces acting on

[5] The contribution of other forces, **O**, might include forces whose entities could serve as affectors or patients in other interactions, as well as forces that might be used to distinguish between periphrastic causative verbs within a subcategory (e.g., *help* vs. *enable* vs. *allow* vs. *let*).

[6] In the more general case in which the target is other than a point, we expect the definition of concordance must be changed to include a certain level of angular tolerance that would be based, in part, upon the relative size of the target and its proximity to the patient.

[7] By definition of the dot product, **P•T** = ||**P**||*||**T**||*cos(θ), where θ is the angle between vectors **P** and **T**. In the case where θ is 0°, the equation becomes **P•T** = ||**P**||*||**A**||*cos(0°), which reduces to **P•T** = ||**P**||*||**T**||.

[8] When concordance is defined in terms of the dot product, it allows for a special type of concordance in which ||**A**|| = 0. When ||**A**|| = 0, the equality **P•A** = ||**P**||*||**A**|| would hold, which may be representative of the kinds of situations referred to by the verbs *let*, *allow*, and *permit*.

the patient are constant with respect to time and space, as specified formally above. In terms of the dot product, the result will occur if $\mathbf{R} \bullet \mathbf{T} = \|\mathbf{R}\| * \|\mathbf{T}\|$ and will not occur if $\mathbf{R} \bullet \mathbf{T} < \|\mathbf{R}\| * \|\mathbf{T}\|$.

As with the Force Dynamic Model, CAUSE, ENABLE, and PREVENT are defined with respect to values along three dimensions, specified in Table 3, and share one feature with each other concept. Thus, both models predict that the three concepts should be equally similar to one another.

Table 3: The Vector Model's representations of CAUSE, ENABLE, & PREVENT

	Tendency of Patient for Target	Concordance of Affector & Patient	Result
CAUSE	N	N	Y
ENABLE	Y	Y	Y
PREVENT	Y	N	N

Testing the Vector Model of Causation

Beyond similarity, the Vector Model makes predictions about the vector configurations underlying verbs of causation. These predictions were tested in two experiments. Participants viewed 3D animations of an inflatable boat, the patient, moving across a shallow pool in relationship to a half-submerged cone, the target (see Figure 4). Each animation had two main parts. In the first, the boat moved from the side of the pool to the center. This part was included to establish the boat's tendency. In the second part, a bank of fans (i.e., the affector) started blowing. Thus, in the second part of every animation, the force produced by the boat itself was combined with the force exerted on the boat by the fans to give rise to a resultant force that determined the boat's direction and speed.

After watching an animation, participants chose among several possible linguistic descriptions. We predicted, per the Vector Model (and its computer implementation), that participants would choose a description containing the verb *cause* when the boat started moving away from the cone (Tendency = N), but was moved to the cone (Result = Y) by the fans blowing in a direction different from the direction of the boat (Concordance = N). We predicted that participants would choose a description containing the verb *help* (a type of ENABLE verb, see Figure 1), when the boat moved towards the cone (Tendency = Y) and ultimately reached it (Result = Y) when the fans blew in the same direction as the boat's direction of motion (Concordance = Y). We also predicted that participants would choose a description containing the verb *prevent* when the boat started towards the cone (Tendency = Y) but did not hit it (Result = N) because the fans blew it back or away from the cone (Concordance = N). Finally, we predicted that when none of the above configurations were instantiated, participants would choose "none of the above." These predictions were tested for one- and two-dimensional interactions in Experiments 1 and 2, respectively.

Figure 4: Sample frame from an animation used in Experiment 2 that instantiated a "cause" interaction

Experiment 1

Method

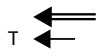
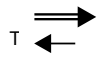
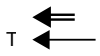
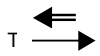
Participants The participants were 18 University of Memphis undergraduates.

Materials Eight 3D animations were made from an animation package called Discreet 3ds max 4. The direction and speed of the boat was calculated by a physics simulator called Havok Reactor. In each animation the boat was initially located four boat-lengths away from the center of the pool. In the first half of the animation, the boat moved towards the center, ostensibly under its own power. Once the boat reached the center, the fans started blowing. The animation ended when the boat hit the cone or neared the side of the pool (~4 seconds total).

The top of Table 4 shows the direction and relative magnitudes of the force vectors associated with the affector and patient that were entered into the physics simulator. The affector, **A**, and patient, **P**, vectors were either in the direction of the target or in the opposite direction. The magnitude of the other forces vector, **O**, was set to 0. In half of the interactions, the affector vector was 1.7 times stronger than the patient (configurations 1-4), while in the remaining interactions the strengths were reversed (configurations 5-8).

Procedure The animations were presented in random order on Windows-based computers. After each animation, participants chose a sentence that best described the occurrence. All of the sentences were the same ("The fans ____ the boat to [from] hit[ting] the cone") except for the verb, which was either *caused*, *helped* or *prevented*. Another option was "none of the above." Participants indicated their answers by clicking a radio button next to their choice.

Table 4. The vectors configurations used in Experiment 1, along with associated predictions and results

Configuration	1	2	3	4	5	6	7	8
Affector (⟹) Patient (⟶) Target (T)								
Predictions	*Help*	*Cause*	*Prevent*	No verb	*Help*	No verb	No verb	No verb
Results								
Cause	11%	94%	-	-	6%	6%	-	-
Help	89%	6%	-	-	94%	-	11%	-
Prevent	-	-	100%	-	-	-	6%	6%
No verb	-	-	-	100%	-	94%	83%	94%

Results and Discussion

The predictions of the Vector Model were fully borne out by the results. The bottom of Table 4 shows the percentage of times people chose each of the four possible options for each of the vector configurations. Participants chose *cause*—as opposed to the other possible options—for the animation in which the boat first moved away from the cone but was later pushed back against it by the fans (configuration 2), a N-N-Y type of occurrence in terms of tendency, concordance and result (see Table 3), $\chi^2(3, N=18) = 62$, $p < .001$. Participants chose *help* when the direction of the boat and the fans was the same (1, 5), a Y-Y-Y type of occurrence, $\chi^2(3, N=18) = 116$, $p < .001$. Participants chose *prevent* when the boat moved towards the cone but was then kept from hitting it by the fans (3), a Y-N-N occurrence, $\chi^2(3, N=18) = 72$, $p < .001$. Finally, participants chose "none of the above" when the vector configurations did not map onto any one of the three main kinds of configurations, $\chi^2(3, N=18) = 237$, $p < .001$. Importantly, participants did not choose *prevent* whenever the boat missed the cone (4, 6, 8). Instead, *prevent* was restricted to those situations in which the boat had an initial tendency for the target (3). Likewise, participants did not choose *cause* or *enable* when the boat simply hit the cone (7), but only when the vector configurations matched those defined by the model. Thus, the Vector Model is capable of not only specifying distinct types of causal concepts, but also distinguishing between causation and non-causation.

The results strongly support the Vector Model, but only in the case of interactions occurring within a single dimension. In Experiment 2 we examine the ability of the model to handle two-dimensional interactions.

Experiment 2

Method

Participants. The participants were 18 University of Memphis undergraduates.

Materials. Ten 3D animations were made in the same way as in Experiment 1 except that the affector and patient force vectors were oriented in several directions other than directly towards or away from the target, and the magnitudes of the affector and patient vectors were always the same. The ten vector combinations at the top of Table 5 depict five combinations in which the patient vector is oriented away from the target by 45° (1-5) and five combinations in which the patient vector is oriented towards the target (6-10). The affector vector was oriented from 180° to 360° at 45° intervals.

Procedure The procedure was as in Experiment 1.

Results

The predictions of the Vector Model were supported once again. The bottom of Table 5 shows the percentage of times people chose each of the four possible options for each of the vector configurations. Participants chose *cause* for the animation in which the

Table 5. The vectors configurations used in Experiment 2, along with associated predictions and results

Configuration	1	2	3	4	5	6	7	8	9	10
Affector (⟹) Patient (⟶) Target (T)										
Predictions	No verb	*Cause*	No verb	No verb	No verb	*Help*	*Prevent*	*Prevent*	*Prevent*	*Prevent*
Results										
Cause	-	89%	-	-	-	11%	-	-	-	-
Help	-	11%	-	-	-	83%	-	-	-	-
Prevent	-	-	17%	-	11%	-	94%	94%	89%	89%
No verb	100%	-	83%	100%	89%	6%	6%	6%	11%	11%

boat was not headed for the cone but hit it because of the fans (2, a N-N-Y occurrence), $\chi^2(3, N=18) = 53$, $p < .001$. Participants chose *help* when the boat was headed for the cone and then was assisted in hitting it by the fans (6, a Y-Y-Y occurrence), $\chi^2(3, N=18) = 44$, $p < .001$. Participants chose *prevent* when the boat was initially headed toward the cone but was later blown away from it (7, 8, 9, 10, a Y-N-N occurrence), $\chi^2(3, N=18) = 229$, $p < .001$. Finally, participants chose "none of the above" when the vector configurations did not map onto any one of the three main kinds of configurations (1, 3, 4, 5), $\chi^2(3, N=18) = 238$, $p < .001$.

Conclusions

In this research we proposed a new model of causal meaning. We also provided empirical support for this model by showing that people's linguistic descriptions of animations are well accounted for by the model and its computer implementation given the same force vectors as those used to produce the animations.

According to the Vector Model, each kind of causal relation is associated with a range of spatial geometries in addition to a particular temporal organization. As a consequence, the model is able to handle causal relations that are highly problematic for probabilistic models, in particular, those in which the cause and effect occur simultaneously (*The sun's gravity causes the earth to revolve around the it*). In such situations, it is difficult to count the causing and resulting events for the purposes of calculating probabilities. In contrast, for the Vector Model, such situations are not problematic since they give rise to readily identifiable vector configurations.

The model provides a new explanation for why billiard-ball events, like the ones studied by Michotte (1963), are construed as causal. Traditionally, this was explained in terms of the spatial-temporal contiguity of the causing and resulting events. Clearly, spatial-temporal contiguity is important: without it, there can be no interaction of (contact-type) forces. But spatial-temporal contiguity is not particular to causal interactions alone. According to the Vector Model, what leads people to describe billiard-ball events as causal is that the patient resists moving (Tendency=N), the affector opposes this tendency (Concordance=N), and the patient ends up moving (Result=Y).

In sum, the Vector Model is able to address several important problems in the causation literature in addition to the two problems discussed in the introduction: the distinction between CAUSE and ENABLE and the expression of singular causation. It also takes us a step closer towards understanding how physical interactions may be construed for the purposes of language.

Acknowledgments

This work was supported by grants award to the first author from The University of Memphis Faculty Research Fund and ONR (N00014-01-1-0917). The conclusions of this research do not necessarily reflect those held by these funding sources. We give special thanks to Tanya Vassilieva, Bianca Klettke, and Derek Wong for their help in the research.

References

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.) *Explanation and Cognition* (pp. 199-225). Cambridge, MA: MIT Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psych. Review, 104*, 367-405.

Cheng, P. W. & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40*, 83-120.

Goldvarg, E., & Johnson-Laird, P. (2001). Naive Causality: a mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565-610.

Jackendoff, R. (1991). *Semantic Structures*. Cambridge, MA: The MIT Press.

Kemmer, S. & Verhagen, A. (1994). The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics, 5*, 115-56.

Michotte, A. (1963). *The perception of causality*. London: Methuen. (Originally published 1946.)

Glymour, C. (2001). *The mind's arrows*. MIT Press.

Pinker, S. (1989). *Learnability and Cognition*. Cambridge, MA: MIT Press.

Robertson, D. A., & Glenberg, A. M. (1998). Force dynamics in language and cognition. *Proceedings of the 20th annual meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*, 1-51.

Siskind, J. M. (2000). Visual event classification via force dynamics. *Proceedings AAAI-2000*, 149-55.

Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science, 8*, 120-123.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

Verhagen, A., & Kemmer, S. (1997). Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics, 27*, 61-82.

Wolff, P., Song, G., & Driscoll, D. (2002). Models of causation and causal verbs. In *Papers from the 37th Meeting of the Chicago Linguistics Society, Main Session, Vol.1*. Chicago: Chicago Linguistics Society.

Wolff, P. & Song, G. (2001). Models of causation and the semantics of causal verbs. The 42th Annual Meeting of the Psychonomic Society, Orlando, FL.

A Self-Organizing Connectionist Model of Character Acquisition in Chinese

Hongbing Xing (xinghb@blcu.edu.cn)
Center for Studies of Chinese as a Second Language
Beijing Language and Culture University
Beijing, 100083, China

Hua Shu (shuh@bnu.edu.cn)
Department of Psychology, Beijing Normal University
Beijing, 100875, China

Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23713, USA

Abstract

Despite growing interests in the acquisition of Chinese orthography, few studies have modeled the acquisition process using connectionist networks. This study uses a self-organizing connectionist model to simulate children's learning of Chinese characters. There are two major goals of our study: (1) To evaluate the degree to which connectionist models can inform us of the complex structural and processing properties of the Chinese orthography. One of the most difficult tasks in achieving this goal is how to faithfully capture the orthographic similarities of Chinese characters. We derived our character representations on the basis of analyzing a large-scale character database that can be readily mapped to school children's orthographic acquisition. (2) To test the utility of self-organizing neural networks in orthographic acquisition. Most previous connectionist models of orthographic processing have relied on the use of feed-forward networks. Results from our simulations present positive evidence for both of our goals. In particular, we show that our model demonstrates early regularity effects and frequency effects in the acquisition of Chinese characters, matching up with acquisition patterns from empirical research.

Introduction

In recent years there have been growing interests in the psycholinguistic study of orthographic acquisition in Chinese (see Yang & Peng, 1997; Shu & Anderson, 1998; Shu, Anderson, & Wu, 2000). A unique feature of the Chinese orthography is that it uses characters rather than alphabetic letters as the basic writing unit, in square configurations that map mostly onto meaningful morphemes rather than spoken phonemes. Processing or acquisition within this "fractal" organization of characters may differ in important ways from that of English and other alphabetic languages (Shu & Anderson, 1998). There are four major types of Chinese characters: pictographic, referential, associative compounds, and ideophonetic compounds. The last type, also known as the semantic-phonetic compounds or, simply, phonetic compounds, is the most interesting and important. In the *Dictionary of Modern Chinese Frequent Characters* (National Language Commission, 1988), there are 5,631 ideophonetic characters, accounting for 81% of the total 7,000 frequent characters in the dictionary (Li & Kang, 1993). Shu, Chen, Anderson, Wu, and Xuan (in press) collected 2,570 characters listed in the Elementary School Textbooks used in Beijing to establish the "School Chinese Corpus". They categorized and labeled every character in this corpus, on dimensions such as phonetic part, phonetic type, position of the phonetic part in the character, age at which the character is taught, and frequency of the character. Shu et al.'s analyses reveal that most of the Chinese characters taught in elementary schools are ideophonetics, as shown in Table 1.

Table 1 Ratios of ideophonetics in each grade
(Shu, Chen, Anderson, Wu, & Xuan, in press)

Grade	1	2	3	4	5	6	Mean
Ratio	.45	.70	.76	.84	.86	.81	.74

Given the prominence of ideophonetics in Chinese orthography, it is thus important for us to understand the functions of these characters. Ideophonetics consist of two major components: the semantic part (often called a radical) that gives information about the character's meaning, and the phonetic part that gives partial information about the whole character's pronunciation. We say "partial", because the phonetic radical may or may not indicate the true pronunciation of the whole character, in one of three ways: (a) Regular: the whole character is pronounced the same as the phonetic radical in isolation – that is, the same as the phonetic radical when it is being used as a simple character; for example, "清/qing1/ and "青/qing1/". (b) Semi-regular: the whole character is pronounced partly as the phonetic radical, with a different

tone (e.g., "请/qing3/ and "青/qing1/"), a different onset (e.g., "晴/jing1/ and "青/qing1/"), or a different final (e.g., "沙/sha1/ and "少/shao3/"). (c) Irregular: the whole character is pronounced completely differently from the phonetic radical (e.g., "猜/cai1/ and "青/qing1/"). These patterns of (ir)regularities in the pronunciations of ideophonetics influence the recognition and processing of Chinese characters, a phenomenon known as the *regularity effect* in the literature.

Previous studies have examined regularity effects in children's acquisition of Chinese characters. Shu, Anderson, and Wu (2000) showed that children display regularity effects when they are required to write down the pronunciations of Chinese characters: they perform better on regular characters (type a discussed above) than on irregular characters (types b and c). When children see unfamiliar characters, they often exploit the pronunciation of the phonetic radical as a possible reading of the whole character, and this ability increases with school grade. Yang and Peng (1997) also found regularity effects in children's speed of naming characters: children in Grade 3 name regular characters more rapidly than irregular characters, but by Grade 6, they name both types of characters equally quickly. Frequency also plays an important role in interacting with the regularity of characters; for example, Shu and Wu (1996) showed that children in Grade 3 display no regularity effects on characters of low frequency, while children in Grades 4 and 6 do. Finally, Shu, Zhou and Wu (2000) found that young children develop from early on phonological awareness of the structures of characters and the functions of the phonetic and semantic radicals. Some 4th graders already start to acquire the awareness of the consistency of phonetic radicals, and by Grade 6 this awareness becomes more transparent.

The above-mentioned properties of Chinese characters and the acquisition profiles therein lend themselves naturally to connectionist modeling. Given the discrepancy between regular and irregular characters, do we need to assume dual mechanisms to handle the two types of characters in acquisition (as symbolic theorists would like to argue)? Or rather, can we assume a connectionist learning mechanism that can capture the acquisition of both types of characters? Previous research in English has examined these issues in language acquisition and orthographic processing (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). However, due to the difficulty in representing the complex structure of the Chinese orthography, there has been very little research in this domain in Chinese. In this study, we make an initial attempt to model the acquisition of Chinese orthography, in particular, the regularity effect in acquisition (as reported in empirical studies) with a neural network.

Method

Architecture

Most previous connectionist models of orthographic processing have relied on the use of feed-forward networks, typically with the back-propagation learning algorithm (e.g., Seidenberg & McClelland, 1989). Recently, a number of studies have explored self-organizing neural networks as viable models of language processing and language acquisition (Anderson, 1999; Miikkulainen, 1993, 1997; Li, 1999, 2000). Self-organizing networks are particularly well suited for the study of language acquisition, due to their biological plausibility, unsupervised learning, and the ability to develop semantic structures (Li, 2002).

In this study, we use a self-organizing feature map model developed by Miikkulainen (1997), originally for modeling disordered lexicons (DISLEX). DISLEX relies on principles of self-organization and Hebbian learning. In this model, different feature maps dedicated to different types of linguistic information (orthography, phonology, or semantics) are connected through associative links via Hebbian learning. To model orthographic processing, an input pattern activates a group of units on the orthographic input map, and the resulting bubble of activity propagates through the associative links and causes an activity bubble to form in the other map (semantic or phonological). Fig. 1 presents a diagrammatic sketch of the model's reading process from seeing the orthographic representation of *dog* to the comprehension of the word's meaning.

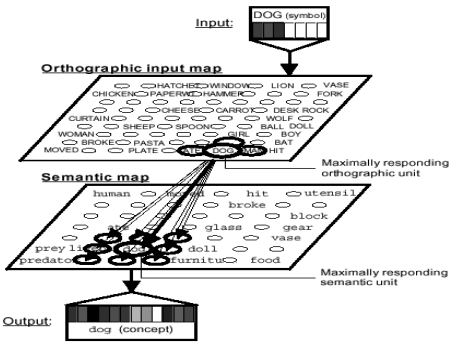


Fig.1 Reading comprehension of *dog* in DISLEX (Miikkulainen, 1997; reproduced with author's permission)

If the direction of the associative propagation is from orthography to semantics, comprehension is modeled, as shown in Fig. 1; production is modeled if it goes from semantics to orthography. In our simulations, we examine the associative propagation from orthography to phonology to model the character naming process in the acquisition of ideophonetics by Chinese-speaking children. At this stage our model has not yet included semantic information.

Input Representation

There are a few general properties of Chinese characters that are important for us to consider for their accurate representations. (1) They are structurally complex.

The basic units of characters are strokes and components. A few simple strokes can make up a component, or a character; (2) They are combinatorially complex. Compound characters have two to over ten components, and these components combine to form a character according different rules in a hierarchically organized structure; (3) The majority of Chinese characters are ideophonetic compounds, as discussed in *Introduction*; and (4) One character corresponds to one monosyllable in spoken language.

Phonological Representation. According to traditions in Chinese linguistics, the monosyllable of each character consists of three parts: initial, final, and tone (see Table 2). Initial is usually a consonant. Final consists of at least the nucleus vowel, sometimes with or without a head vowel or a tail vowel. The nucleus vowel may be one single phoneme or a diphthong (two phonemes). Lexical tones are supra-segmental, imposed on the initial and the final. In our representation scheme, we represent each phoneme (consonant or vowel) by 5 dimensions or features, and each feature by the phoneme's articulatory properties on a continuous scale from 0 to 1 (Table 3). The overall method of representation is similar to PatPho, a phonological representation scheme for English described by Li and MacWhinney (2002).

Table 2 Structure of the syllable and representation

Initial	Final			Tone
	Head Vowel	*Nucleus Vowel*	*Tail Vowel*	
5 dim.	5 dim	5 +5dims	5 dim	5 dim.

With this method we can represent all Chinese monosyllables with tone (a total of 1,335), each of which on a 30-dimensional feature vector. Table 3 lists the articulatory features we used to represent the Chinese phonemes.

*Table 3 Articulatory features on 5 dimensions (D1-D5) for the representation of Chinese phonemes**

D1	vowel	voiced	voiceless				
	0.1	0.75	1.0				
D2	bilabial	Labio-dental	front	central	back	palatal	velar
	0.143	0.286	0.429	0.572	0.715	0.858	1.0
D3	round	nasal	stop	fricative	affricate	retroflex	lateral
	0.143	0.286	0.429	0.572	0.715	0.858	1.0
D4	high	mid	low				
	0.333	0.666	1.0				
D5	front	central	central-back	back			
	0.25	0.5	0.75	1.0			

* Numbers indicate dimensional values for each feature

Orthographic Representation. To accurately represent Chinese orthography in feature vectors has proven a very challenging task, and this may have been the primary reason for the lack of modeling research in this domain, as we pointed out earlier. The only large-scale attempt in this respect was Chen & Peng (1994). They used 30 feature units to represent the various components of about 1,108 Chinese characters. Their representation scheme, however, is still insufficient for our purposes of modeling children's development in character acquisition. To overcome this bottleneck problem, we did a detailed analysis of all the characters in the *UCS Chinese Character Database* (Standards Press, 1994) and examined the strokes, components, and structures of these characters.

The *UCS Chinese Character Database* contains information about the structure and components for each of the 20,902 Chinese characters used in China, Japan, and Korea. This information includes the hierarchically ordered sequences of each component when characters are decomposed into smaller units of strokes. Other information includes pronunciation of the character, first-level categorization of the character, number of components, number of strokes, and frequency of usage. The database lists 560 basic components for the 20,902 Chinese characters, including the character's structural features, shape features, position of components, number of component strokes, etc. Most relevant for our study is the information about phonetic radicals in ideophonetic characters. This includes the position of the phonetic radical in the character, whether the position of the radical is fixed, and the relationship between the pronunciation of the phonetic radical and that of the character. Finally, the database contains information about the frequency of each character in elementary school texts, as well as some of the original texts.

On the basis of our analyses of this database, we represented each ideophonetic Chinese character with a 60-unit feature vector, along the dimensions as depicted in Fig.2.

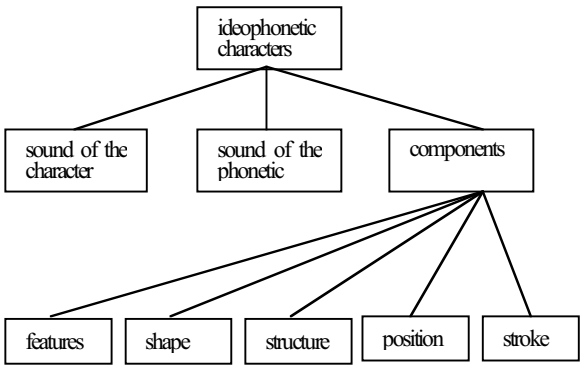


Fig. 2 Orthographic representation of characters

The first 6 units represent the sound of the character, while the second 6 units represent the sound of the phonetic radicals. The purpose of these phonological units is to see how much overlap there is between the

pronunciation of the phonetic radical and that of the whole character. The next 48 units are used to represent component features, shapes, stroke structures, position of radicals, and stroke numbers. For example, component features include single, separate, crossing, and connecting; position of radicals includes top, bottom, left-side, right-side, middle, and inner, etc. The last unit is used for stroke numbers, and to determine the value of this unit, we analyzed the number of strokes of characters in our database. Given that most of the characters are between 1 to 10 strokes, characters with 10 or more strokes are represented as 1.0, and the rest as values (in decrements of .1) corresponding to the number of strokes (i.e., 0.9 for characters with 9 strokes, 0.8 for characters with 8 strokes, and so on).

Materials and Procedure

Materials. The basic training materials consist of groups or families of Chinese characters – characters in the same family have the same phonetic radicals, sometimes including the radical itself as a character. Because we are modeling elementary school children's acquisition of characters, the amount of character families differ for different grades, and the same family may also contain different numbers of family members, according to our analysis of the School Chinese Corpus (Shu et al., in press). We selected families of characters from the elementary textbooks for Grades 1, 3, and 5 as the basic materials in our training. Characters are selected as our input materials (a) if they have been learned in or before this grade, or (b) if the family includes all ideophonetic characters that have been learned before. Table 4 shows the composition of our training materials, based on depth of learning in school grades.

Table 4 Selected characters and family compositions

Grade	Total characters	# of families	Mean members of a family
One	306	214	2.35
Three	305	139	4.33
Five	300	113	5.64

Training. Each batch of characters corresponding to each grade was submitted to the model, trained for 350 epochs for the self-organization of phonological representations and of orthographic representations. Upon training of the network, a phonological representation of a character was inputted to the network, and simultaneously, the orthographic representation of the same character was also presented to the network. By way of self-organization, the network formed an activity on the phonological map in response to the phonological input, and an activity on the orthographic map in response to the orthographic input. The phonological representation of the character was also co-activated with its orthographic representation. As the network received input and continued to self-organize on each map, it simultaneously learned associative connections between maps through Hebbian learning: initially, all units on the phonological map were fully connected to all units on the orthographic map; as learning continued, only the units that were co-activated in response to the inputs were associated. As the end of learning, the network should have created compressed new representations in the corresponding maps for all the inputs and linked the phonological representation to its orthographic pattern. All simulations were conducted with the DISLEX simulator (Miikkulainen, 1999).

Testing. Once the network has completed self-organizing on the phonological and orthographic inputs and has learned the associative connections, we tested the model's performance by presenting the network with 16 ideophonetic characters. We inputted the orthographic and phonological patterns of these 16 characters to the trained and well-settled network to test the output pronunciations of the characters in the model (see Fig. 1). No learning takes place at this stage. For each grade, a total of 48 characters was tested in the model, in three batches: 16 high frequency characters, 16 low frequency characters, and 16 new characters that have not be learned by the grade being tested.

Results and Discussion

Table 5 shows the overall performance of the network after it was trained for 350 epochs on all characters corresponding to each of the three grades being considered. These results show that the network achieved an average of 76% accuracy for orthographic representations; for phonological representations, it reached an average of 79% accuracy, and for the associative connections from orthography to phonology it achieved an average of 93% accuracy, a highly successful naming ability.

Table 5 Percent accuracy on orthography, phonology, and associative connections in the model after training

Grade	Ortho. map	Phono. map	Associative ortho.->phono.
One	75.6	78.4	90.5
Three	71.7	76.2	93.7
Five	80.2	82	95

Thus, after training, the model developed clearly structured representations for both the phonological and orthographic input patterns. Fig. 3 shows an example from the orthographic map trained on Grade 5 characters (only a portion of the entire map is shown here due to space limit).

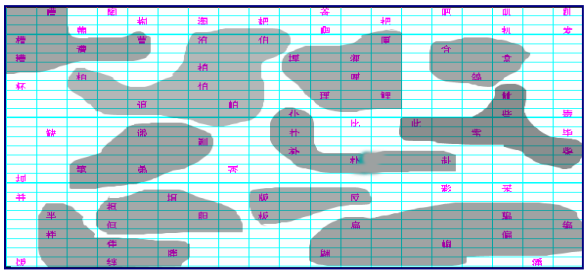


Fig. 3 Orthographic map trained on Grade 5 characters

In Fig. 3, it can be seen that clear families of characters emerged in the map after the network was trained for 300 epochs on the Grade 5 characters. For example, one group of characters on the lower right-hand corner represents the "扁/bian3" family with "骗/pian4" ,"编/bian1", "偏/pian1", "蝙/bian1", "翩/pian", while the other group on the upper right hand of the map represents the "合/he2" family with "盒/he2" ,"鸽/ge1", etc.

To see the model's ability in character naming, we tested the accuracy of its naming of regular and irregular characters for Grades 1, 3, and 5, with regular character being one that has exactly same pronunciation as its phonetic radical (see Introduction). The ratios of naming accuracy are presented in Fig. 4, on which we can make several observations: (a) the model's naming accuracy increases over time for both regular and irregular characters; (b) the difference in naming accuracy between regular and irregular characters also increases across grade; and (c) regularity effect does not exist for Grade 1 but becomes transparent for Grades 3 and 5. These results match up well with the empirical patterns observed by Shu et al (2000).

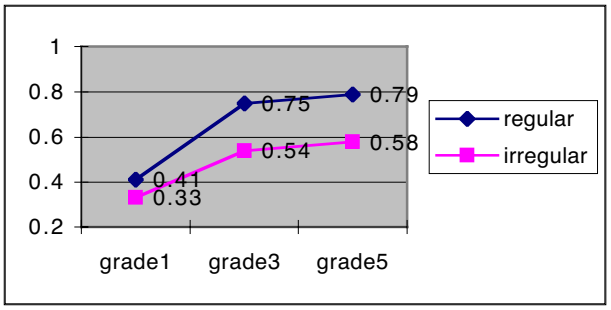


Fig. 4 Naming accuracy for regular and irregular characters

Interestingly, the regularity effect shown in Fig. 4 is modulated by character frequency. Ample empirical evidence suggests that frequency effect interacts with regularity effect in character acquisition (Shu, Anderson, & Wu, 2000). Table 6 shows that regularity effect in the model is only marginal for high frequency characters, but is much clearer for low frequency characters and novel characters (whose frequency is unknown to the network).

Finally, we analyzed the network's error types in naming irregular characters. In naming ideophonetic characters, children as well as our network could use a variety of methods to get at the pronunciation of the irregular character. These methods allow us to discern regularity effects in reading acquisition. There are basically three major methods they could use: (1) reading the irregular character as the pronunciation of its phonetic radical (e.g., "橙/cheng2" as "登/deng1"); (2) reading the character as another character having a similar orthography/radical in the same family (e.g., "蝙/bian1" as "偏/pian1"); and (3) reading the character as other irrelevant characters (e.g., "枞/zong1" as "凯/kai3"). Table 7 shows the ratio of the network's erroneous naming of irregular characters for each grade, as a function of naming methods (M1 = Method (1); M2 = Method (2); and M3 = Method (3)).

Table 6 Interaction between regularity and frequency

Frequency	Regular	Irregular
	.88	.83
Low frequency	.75	.46
New characters	.33	.17

Table 7 Network's naming for irregular characters

Grade	Irregular Character Naming Errors		
	M1	*M2*	*M3*
One	.06	.25	.69
Three	.36	.46	.18
Five	.30	.50	.20

Table 7 shows several interesting patterns. First, for Grade 1 the network's errors are mainly based on Method 3, i.e., reading characters as irrelevant characters. This shows that regularity effect has not played much of a role yet in the naming of irregular characters. For Grades 3 and 5, however, the error types shift more toward Methods 1 and 2, showing that the model is exploring orthographic and phonological similarities of the radical to give possible pronunciations of the irregular character. These developmental patterns of regularity effect are consistent with empirical data from Shu et al (2000), according to which children, although in principle can utilize ideophonetic information early on, show regularity effect only after they have learned a relatively large number of items in the ideophonetic families.

Conclusion

This study uses a self-organizing connectionist model to simulate children's acquisition of Chinese characters. There are two major goals of our study. First, we wanted to see if connectionist models can be applied successfully to model the learning process in the acquisition of Chinese characters, a topic that has not been touched on in the literature. Given the complex structural and processing properties of the Chinese orthography, it is only natural that we examine this

domain with systematically varying modeling parameters. One of the most difficult tasks in achieving this goal is how to faithfully capture the orthographic similarities of Chinese characters, as discussed in *Method*. We derived our character representations on the basis of analyses of a large-scale character database that can be readily mapped to school children's orthographic development.

The second goal of our study is to test the utility of self-organizing neural networks. Most previous connectionist models in this domain have relied on the use of feed-forward networks, typically with the back-propagation learning algorithm. In previous research, Miikkulainen (1993, 1997) explored self-organizing neural networks as plausible models of language and memory processing, and Li (1999, 2000, 2002) showed these networks as viable models of language acquisition. We wanted to see if such models can be used successfully to examine orthographic acquisition. Our initial simulations as presented here seem to provide positive evidence in this respect. In particular, we showed that our self-organizing network demonstrates regularity effect and frequency effect in the acquisition of Chinese characters, and that these results match up with developmental patterns observed in empirical research. In future studies, we will continue these lines of experiments to examine frequency effect, phonological consistency effect, and their interaction with regularity effect in character acquisition, addressing other relevant theoretical issues in connectionist language acquisition.

Acknowledgments

This research was supported by a grant from Natural Science Foundation of China #60083005 to H.S., and in part by an NSF grant #9975249 to P.L. while the first author was visiting the Cognitive Science Lab at the University of Richmond. We would like to thank Risto Miikkulainen for making available the source code of the DISLEX program, and Igor Farkas for helping with the set-up of the simulations.

References

Anderson, B. (1999). Kohonen neural networks and language. *Brain and Language*, *70*, 86-94.

Chen, Y., & Peng, D. (1994). A connectionist model of recognition and naming of Chinese characters. In H-W. Chang, J-T. Huang, C-W Hue, & O. Tzeng (eds.), *Advances in the study of Chinese language processing* (Vol.1, pp. 211-240). Taipei: National Taiwan University Press.

Li, P. (1999). Generalization, representation, and recovery in a self-organizing feature-map model of language acquisition. In M. Hahn & S.C. Stoness (eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp.308-313). Mahwah, NJ: Lawrence Erlbaum

Li, P. (2000). The acquisition of tense-aspect morphology in a self-organizing feature map model. In L. Gleitman & A.K. Joshi (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp.304-309). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (2002). Language acquisition in a self-organizing neural network model. In P. Quinlan (ed.), *Connectionist models of development*. Philadelphia and Briton: Psychology Press.

Li, P., & MacWhinney, B. (2002). *PatPho*: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers*. (in press)

Li, Y. & Kang, J. S. (1993). Analysis of phonetics of the ideophonetic characters in Modern Chinese. In Y. Chen (ed.), *Information analysis of usage of characters in Modern Chinese* (pp. 84-98). Shanghai: Shanghai Education Publisher. (in Chinese)

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. Cambridge, MA: MIT Press.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, *59*, 334-366.

Miikkulainen, R. (1999). The DISLEX simulator (new version). Available on-line at http://www.cs.utexas.edu/users/nn/pages/software/.

National Language Commission of China (1988). *Dictionary of Frequent Characters in Modern Chinese*. Beijing: Yuwen Press.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56-115.

Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Phonological Review*, *96*, 523-568.

Shu, H., & Anderson, R. C. (1998). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. W. Inhoff, H.-C. Chen (eds.). *Reading Chinese script: A cognitive analysis* (pp. 1-18). Mahwah, NJ: Lawrence Erlbaum.

Shu, H., Anderson, R. C., & Wu, N. (2000). Phonetic awareness: Knowledge of orthography-phonology relationships in the character acquisition by Chinese children. *Journal of Educational Psychology*, *92*, 56-62.

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (in press). Properties of School Chinese: Implications for learning to read. *Child Development*.

Shu, H., Zhou, X., & Wu, N. (2000). Utilizing phonological cues in Chinese characters: A developmental study. *Acta Psychologica Sinica*, *32*, 164-169. (in Chinese)

Standards Press of China (1994). *Information Technology – UCS: Universal Multiple-Octet Coded Character Set* (Part 1 : Architecture and Basic Multilingual Plane). Beijing.

Yang H, & Peng, D. L. (1997). How are Chinese characters represented by children? The regularity and consistency effects in naming. In H. C. Chen (ed.). *The cognitive processing of Chinese and related Asian Languages*. Hong Kong: The Chinese University Press.

Uncertainty in Causal and Counterfactual Inference

Daniel Yarlett (dany@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW

Michael Ramscar (michael@dai.ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW

Abstract

We report 4 studies which show that there are systematic quantitative patterns in the way we reason with uncertainty during causal and counterfactual inference. Two specific type of uncertainty – uncertainty about facts and about causal relations – are explored, and used to model people's causal inferences (Studies 1-3). We then consider the relationship between causal and counterfactual reasoning, and propose that counterfactual inference can be regarded as a form of causal inference in which factual uncertainty is eradicated. On this basis we present evidence that there are also systematic quantitative patterns underlying counterfactual, as well as causal, inference (Study 4). We conclude by considering the consequences of these results for future research into causal inference.

Introduction

The ability to make causal inferences is of central importance to cognitive agents wishing to control or predict events in the world. However, many of our beliefs are held with less than perfect certainty. Given this, it is natural to enquire about the way in which uncertainty affects the process of reasoning about the world. In this paper we examine the way in which two types of uncertainty – uncertainty about facts and uncertainty about causal relations – are assimilated during the process of causal and counterfactual inference. Studies 1-3 reveal that there are systematic quantitative patterns in our treatment of uncertainty in causal inference, suggesting that our understanding of causality is not inherently deterministic as has recently been proposed (Goldvarg & Johnson-Laird, 2001). We then consider the relationship between causal and counterfactual inference, and show that modified forms of the models which can be used to predict causal inference can be used to predict counterfactual inference (Study 4), a result consistent with theories which treat counterfactuals as supervenient on causal knowledge (Pearl, 2000; Yarlett & Ramscar, in press; Jackson, 1977). We conclude by considering the consequences of these results for future research into causal and counterfactual inference.

Causal Inference and Uncertainty

We make a causal inference when we acquire some new evidence about a cause, and on this basis update our beliefs about effects related to that cause. For example, imagine you meet Tom at a party. During your brief conversation, he says and does certain things that make you think he is an Army Officer, although you're not completely certain about this. As a result of this suspicion you might now think it more likely that Tom is able to fire a pistol and abseil, compared to when you first met him. Your beliefs about Tom have changed as a result of causal inference.

When it comes to making causal inferences, two types of uncertainty are especially important: *factual uncertainty* and *causal uncertainty*. Factual uncertainty arises simply because our experience of the world is in many cases insufficient to allow us to be completely certain about our beliefs. For example, Tom's extensive knowledge of firearms and military strategy, as displayed in your conversation, might make you suspect that he is in the Army. But you are nevertheless aware that you could be wrong about this. Therefore there is some factual uncertainty in your belief that Tom is in the Army.

The second type of uncertainty relevant to making inferences about the world is *causal uncertainty*. This arises because although there are systematic regularities in the world, these rarely obtain without exception. For example, we all agree that clouds cause rain, even though rain does not invariably fall when it is cloudy; and we would probably also concur that smoking causes cancer, although we know that not all smokers contract cancer. Causal uncertainty, then, arises because of our awareness that although events of type A may tend to produce events of type B, it is not the case that As are *always* or *invariably* followed by Bs.

Although it seems intuitively plausible that both factual and causal uncertainty should play a role in determining our causal inferences, to our knowledge very little empirical work has explored this issue. Some previous work has found an effect of factual uncertainty in both deductive (Stevenson & Over, 1995; Byrne, 1989) and causal (Cummins *et al.*, 1991) settings, but

none of these studies examined the systematic effects of factual uncertainty from a quantitative perspective. And although it seems reasonable to assume that causal uncertainty plays a role in causal inference and reasoning – and indeed, many recently proposed theories (e.g., Cheng, 1997; Pearl, 2000) and models (Rehder, 1999; Yarlett & Ramscar, in press) concerned with causal reasoning successfully make this assumption – it is by no means uncontentious. Goldvarg & Johnson-Laird (2001) have recently argued that the meaning of causal statements is inherently deterministic, and more generally, theories of reasoning which invoke mental models do not easily permit the accommodation of less than certain inferences (but see Johnson-Laird, 1994, and Stevenson & Over, 1995). The present series of studies therefore set out to investigate whether factual and causal uncertainty play a role in the process of causal inference and, if so, whether they do so in a systematic fashion.

Study 1

Study 1 was designed in order to get ratings about the causal uncertainty attaching to a specific set of cause-effect pairs, in order to explore the structure of the information with which people relate causes and effects, and also to investigate the information that might be used in causal inference. People were asked to rate the causal uncertainty attaching to a range of cause-effect pairs on a range of scales which measured: (i) how strongly the cause causes the effect; (ii) how strongly the effect depends on the cause; (iii) the conditional probability of the effect given the presence of the cause; and (iv) the conditional probability of the effect given the absence of the cause. In addition to the ratings collected, the following ratings were derived from the conditional probability ratings:

$$\Delta P \text{ Contingency} = P(e|c) - P(e|\sim c)$$

$$\text{Power PC} = \frac{P(e \mid c) - P(e \mid \sim c)}{1 - P(e \mid \sim c)}$$

These quantities have variously been proposed as measures of the strength of a cause (e.g. Cheng & Novick, 1992; Cheng, 1997).

Materials and Design. The materials used described 10 different cause-effect pairs. They were selected in order to cover a wide variety of domains, and included the following pairs: smoking and cancer; cars and pollution; stress and insomnia; sunbathing and suntanning; weight-training and muscle-growth; cholesterol and heart attacks.

Subjects were asked directly about the strength of relation that they thought held between the pairs in question. For example, for the smoking-cancer pair, the following questions were used: (i) "How strongly do you think smoking causes cancer?"; (ii) "How strongly do you think whether someone gets cancer depends on whether they smoke?"; (iii) "How likely do you think someone would be to get cancer given that they smoke?"; and (iv) "How likely do you think someone would be to get cancer given that they do not smoke?" All ratings were collected on a 0-100 scale. For the causal ratings the scale was anchored by 'does not cause at all' and 'always causes'; for the dependency ratings 'does not depend at all' and 'perfectly depends'; and for the subjective probability ratings 'completely unlikely' and 'completely certain'.

	Factor 1 (Causal Power)	Factor 2 (Baserate)
Causal	0.969	-0.042
P(e|c)	0.927	-0.326
Power PC	0.873	-0.481
Dependency	0.699	-0.600
ΔP	0.664	-0.738
P(e|~c)	-0.106	0.982

Table 1: Factor loadings from Study 1.

Three groups were asked to rate the causal, dependency, and conditional probability ratings. A within-subjects design was not used because of concerns that this would artificially homogenise what might in reality be different ratings (e.g. being asked to rate causal, dependency and conditional probability ratings consecutively might encourage subjects to simply return similar ratings on all scales).

Participants. 49 students from the University of Edinburgh participated voluntarily.

Results. A factor analysis (principal-components analysis with rotated axes) was conducted on the ratings in order to examine their structure. Only the first two rotated factors had eigenvalues greater than 1, and these together accounted for 95.16% of the variance in the data. The factor loadings are shown in Table 1.

Discussion

The two factors extracted in the factor analysis successfully explained a large proportion of the variance in the causal uncertainty ratings collected in Study 1. Moreover, the extracted factors are readily interpretable because the causal ratings load very highly on the first factor (0.969) and negligibly on the second factor (-0.042), while the P(e|~c) ratings load very highly on the second factor (0.982) and negligibly on the first factor (-0.106). Study 1 therefore suggests that two factors are especially important in accounting for our representation of causal uncertainty: the causal strength with which a cause produces its effect ('Causal Power'), and the base rate of the effect in the absence of the cause ('Baserate'). This suggests that models of

Model	Definition
Probabilistic	P(e|c)P(c) + P(e|~c)P(~c)
Linear	P(e|~c) + causes(c,e)P(c)
Noisy-OR	1 – [1 – P(e|~c)][1 – causes(c,e)P(c)]
Causal	causes(c,e)P(c)
Dependency	depends(e,c)P(c)

Table 2: The models of causal inference.

causal inference should incorporate these two parameters. This proposal was investigated in Study 2.

Study 2

Study 2 examined the degree to which causal inference can be modelled using information about factual and causal uncertainty. The ratings from Study 1 were used in order to provide information about the degree of causal uncertainty attaching to the 10 causal pairs, while new data was acquired concerning their factual uncertainty. Short scenarios centring around each of the 10 causal pairs were designed, in which it was deliberately made unclear whether the cause was present or absent. These were used to induce factual uncertainty in participants in the study. For example, the scenario for the smoking-cancer causal pair ran as follows:

> "Imagine you're introduced to Bill, a friend of a friend, one day. You ask Bill for a lighter but he doesn't carry one. However, it does look a little as though he might have tobacco stains under his nails."

After reading each description, participants were requested to rate their factual uncertainty, on a 0-100 scale, by being asked how likely they thought it was that the cause was present given what they had read (i.e., in this case how likely they thought it was that Bill was a smoker). They were then asked to make a causal inference by judging, given their confidence that Bill may or may not be a smoker, how likely they thought he would be to contract cancer at some point in his life. The information collected about factual and causal uncertainty was then used to parameterise various models of causal inference, in order to see if the inferences participants made could be predicted.

Models of Causal Inference

The models of causal inference investigated are listed in Table 2. The probabilistic model defines the normative method of inferring the probability of an effect given information about a related cause. The linear model, in contrast, states that one's belief in an effect is the combination of a base rate of belief – the belief that the effect is present in the absence of the cause – plus the extra support that the cause provides for belief in the effect, which is defined as the product of one's belief that the cause is present and the degree to which the cause and effect are causally related. The noisy-OR model (Pearl, 1988) treats causes as mechanisms that operate independently and additively to produce a common effect. The probability of an effect in this framework is thus given as the probability that not all the causes fail to generate the effect.[1] Finally, the causal model predicts that people's belief in the cause is a product of the degree to which the cause and effect are causally related, and the degree to which the cause is believed to be present. And the dependency model is similar to the causal model, except that it measures causal uncertainty using dependency, instead of causal strength, ratings.

Materials and Design. The cause-effect pairs and connection ratings from Study 1 were used. In addition, scenarios for each causal pair were designed in order to embed the causal relation in a specific context, and deliberately induce factual uncertainty as to whether the cause in question was present or not.

A within-subjects design was deliberately eschewed in Study 2 because of concerns that it could artificially bring people's causal inferences in line with the predictions of the probabilistic model. Many people are familiar with basic probability theory, and our concern was that being asked to rate the conditional probability of the effect given the cause before making their causal inference (as a within-subjects design would have required), could force people to reason about the effect arithmetically, in opposition to their natural style of reasoning. Accordingly, a between-subjects design was adopted, in which the causal uncertainty ratings used were those collected in Study 1, while the factual uncertainty ratings and the causal inferences themselves, were collected in the present study.

Participants. Participants were 21 students at the University of Edinburgh. All participants were volunteers, and no reward was offered for participation.

Results. The performance of the models of causal inference is shown in Figure 1. The probabilistic ($r = 0.665$, $p < 0.05$, one-tailed), linear ($r = 0.621$, $p < 0.05$, one-tailed), and noisy-OR ($r = 0.711$, $p < 0.05$, one-tailed) models were all significant predictors of people's causal inferences. The causal ($r = 0.495$, $p > 0.05$, one-tailed) and dependency ($r = 0.268$, $p > 0.05$, one-tailed) models, however, failed to significantly predict people's inferences. A further analysis was also

[1] Interestingly, the linear and the Noisy-OR models of causal inference find their counterparts in the ΔP and Power PC theories of causal induction respectively (they can be derived as the maximum likelihood estimates of causal strength parameters in causal graphs appropriately parameterised; see Glymour, 1998; Tenenbaum & Griffiths, 2000).

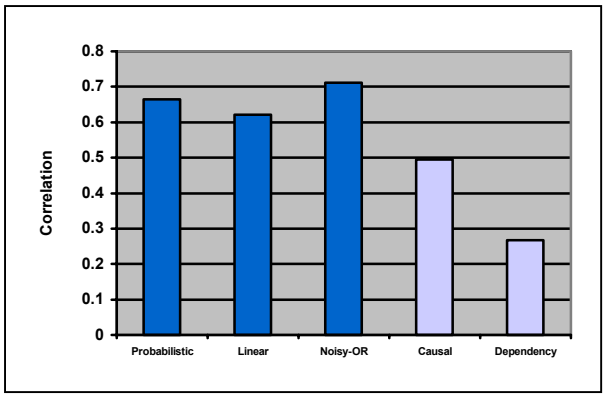


Figure 1: Results of Study 2.

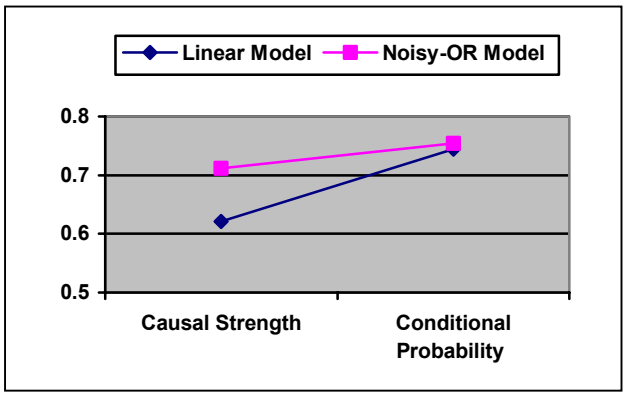


Figure 2: Effect of changing parameterisation of Linear and Noisy-OR models.

conducted on the linear and noisy-OR models, because they could also be parameterised with conditional probability instead of causal strength ratings. In both cases, parameterising the models with the conditional probability of the effect given the presence of the cause, instead of the causal strength ratings, served to increase their empirical performance (see Figure 2).

Discussion

The fact that the linear, noisy-OR and probabilistic models were significantly correlated with the strength of people's causal inferences suggests that information about factual and causal uncertainty plays an important role in the inference process, and also that there seem to be domain-general quantitative patterns in the way we reason from cause to effect. However, the factual and causal uncertainty ratings and inferences predicted in Study 2 were between-subjects aggregates. It is therefore possible that the success of the proposed models is merely an artefact of the experimental design, and that the models would prove unable to predict causal inferences on a within-subjects basis. Study 3 investigated this issue, while also allowing us to examine how much of the residual error in the causal models could be attributed to idiosyncratic use of the rating scales.[2]

Study 3

Study 3 used a within-subjects design in which people estimated the factual and causal uncertainty attaching to each of the 10 cause-effect pairs, and then made a causal inference about the effect. Because the causal and dependency models failed to significantly predict people's inferences they were dropped from consideration. Instead we focused on the performance of just the probabilistic, linear and noisy-OR models.

Materials and Design. The causal pairs and materials as used in Studies 1 and 2 were again used in this study. Each subject saw all 10 scenarios, in one of two reverse-orderings. The linear and noisy-OR models were parameterised using only conditional probability, and not causal strength ratings, because of the better performance of this form of the models in Study 2.

Participants. Participants were 15 students enrolled at the Division of Informatics, University of Edinburgh. All participants were volunteers, and no reward was offered for participation.

Results. The performance of the causal models is shown in Figure 3. Both the linear ($t = 2.280$, $df = 14$, $p = 0.038$, two-tailed) and the noisy-OR model ($t = 2.379$, $df = 14$, $p = 0.032$, two-tailed) performed significantly better than the probabilistic model, although there was no significant difference between the linear and noisy-OR model ($t = 1.302$, $df = 14$, $p = 0.214$, two-tailed). The degree of variance explained in the inference process by just taking into account either the amount of factual uncertainty, in the form of the p(c) ratings, or the amount of causal uncertainty, in the form of the p(e|c) ratings, is also shown in Figure 3 for comparison. These two models performed significantly worse than all the other models.

To confirm that both the factual and causal uncertainty parameters added to the models' predictive validity the performance of the linear and noisy-OR models was compared to modified versions of them in which (i) factual uncertainty was ignored; and (ii) causal uncertainty was ignored. The linear model performed significantly better than its counterpart which ignored factual uncertainty ($t = 2.358$, $df = 14$, $p = 0.017$, one-tailed), and marginally better than its counterpart which ignored causal uncertainty ($t = 1.546$, $df = 14$, $p = 0.072$, one-tailed). The noisy-OR model performed significantly better than both its modified

[2] Because Study 2 had shown that the probabilistic model predicted causal inference with some level of success in a context in which patterns of causal inference consistent with the predictions of the probabilistic model could not have been artificially induced, the use of a within-subjects design was now appropriate.

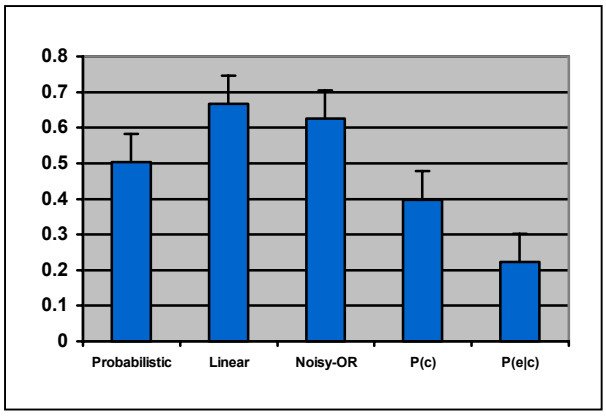


Figure 3: Results from Study 3.

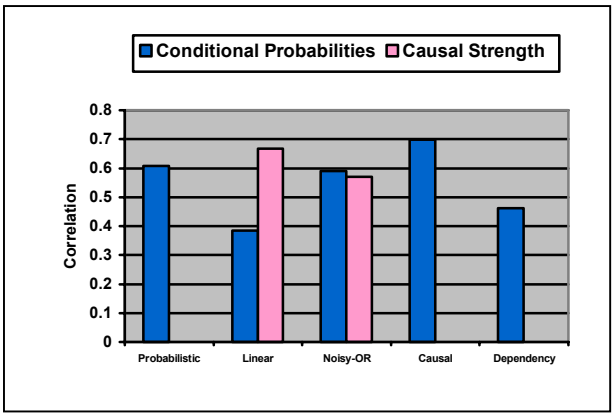


Figure 4: Results from Study 4.

versions which ignored factual ($t = 4.199$, $df = 14$, $p < 0.001$, one-tailed) and causal ($t = 2.049$, $df = 14$, $p = 0.030$, one-tailed) uncertainty.

Discussion

The results of Study 3 show that quantitative models – particularly the linear and noisy-OR model – can successfully predict people's causal inferences with some degree of success. Moreover, the results of Study 3 also show that removing information about either factual or causal uncertainty from these models significantly decreases their performance, thus showing that these factors do seem to play an important role in causal inference.

Causes and Counterfactuals

The studies reported so far examined the role of uncertainty in causal reasoning. However, there is an intimate connection between causal and counterfactual reasoning (c.f. Lewis, 1973b; Jackson, 1977; Pearl, 2000; Yarlett & Ramscar, 2001). In the light of this it is interesting to examine whether the findings concerning causality in Studies 1-3 can also be applied to counterfactual reasoning.

The proposal we examined is that, at least in the present context, counterfactual reasoning can be treated as a form of causal reasoning in which residual factual uncertainty is eliminated (for treatments of counterfactual reasoning in more complex systems see Yarlett & Ramscar, in press, and Pearl, 2000). For example, imagine that you are fairly sure that Bill is not a smoker, but that I ask you how likely you think he would be to contract cancer if (counterfactually) he were a smoker. Even though there may be some factual uncertainty in your belief that Bill is not *actually* a smoker, there should be no factual uncertainty attaching to the counterfactual scenario because the counterfactual asks you to assume, unequivocally, that he *is* a smoker. Study 4 investigated this proposal.

Study 4

Study 4 investigated whether quantitative patterns could be found underlying counterfactual, as well as causal, inference. The scenarios used in Studies 2 and 3 were altered so that instead of engendering uncertainty they were perfectly unambiguous that the cause in question was absent. Then, instead of being asked to make a straightforward causal inference, subjects were asked to consider how strongly they would believe in the effect *if the cause were present*.

Materials and Design. The materials used were adapted forms of the scenarios used in Studies 2-3. Here is the smoking scenario, with the added information shown in italics:

> "Imagine you're introduced to Bill, a friend of a friend, one day. You ask Bill for a lighter but he doesn't carry one. However, it does look a little as though he might have tobacco stains under his nails. *It later turns out that Bill is not a smoker; in fact he's never even smoked a cigarette in his life.*"

Subjects were then asked to rate "But if Bill were a smoker, how likely do you think he would be to get cancer at some point in his life?". Data was collected using a between-subjects design, as used in Study 2.

Participants. Participants were 23 students at the University of Edinburgh.

Results. The results of Study 4 are shown in Figure 4. The causal model ($r = 0.699$, $df = 8$, $p < 0.05$, one-tailed), linear model parameterised with causal strength ratings ($r = 0.667$, $df = 8$, $p < 0.05$, one-tailed), and noisy-OR model parameterised with either conditional probabilities ($r = 0.589$, $df = 8$, $p < 0.05$, one-tailed) or causal strengths ($r = 0.571$, $df = 8$, $p < 0.05$, one-tailed), significantly predicted people's counterfactual inferences.

Discussion

The results of this study show that modified forms of the models used to predict causal inferences can also be employed in the prediction of counterfactual inferences, and also that counterfactual inference can be profitably regarded as a special case of causal inference in which factual uncertainty has been eradicated. This result is both consistent with theories which hold that counterfactuals supervene on causal relations (e.g., Jackson, 1977; Pearl, 2000; Yarlett & Ramscar, in press), and at tension with theories that treat counterfactual judgements as propositions assigned binary truth-values (e.g., Byrne, 1997; Byrne & Tasso, 1999; Lewis, 1973). However, given the success of multiple models at capturing the quantitative patterns in counterfactual inference exhibited in Study 4, clearly further work is required to tease the models apart, and determine whether patterns in both causal and counterfactual inference can be successfully captured by the same models.

General Discussion

The 4 studies reported here suggest that both factual and causal uncertainty play an important role in determining causal and counterfactual inference, and furthermore that counterfactual inference can profitably be regarded as a form of causal inference in which factual uncertainty is eradicated. However, one potential cause for concern is the often considerable amount of variance left unexplained by the sort of quantitative models described in this paper. Clearly more work needs to be done before the role of such models in describing causal and counterfactual inference is fully understood. In particular, in future work we intend to examine whether alternative ways of measuring causal and factual uncertainty can increase the explanatory power of the quantitative models, and also whether additional factors can be imported into the models to improve their empirical fit (e.g. how many alternative or preventative causes exist for a specific cause effect pair being reasoned about; see Cummins *et al.*, 1991).

Acknowledgements

The authors would like to thank attendees of the workshop on *Causal Learning and Inference in Humans and Machines* at NIPS*2001.

References

Byrne R.M.J. (1989). Suppressing Valid Inferences with Conditionals, *Cognition*, **31**, 61-83.

Byrne R.M.J (1997). Cognitive Processes in Counterfactual Thinking About What Might Have Been, *Psychology of Learning and Motivation*, **37**, 105-154.

Byrne R.M.J. and Tasso A. (1999). Deductive Reasoning with Factual, Possible, and Counterfactual Conditionals, *Memory & Cognition*, **27**(4), 726-740.

Cheng P.W. (1997). From Covariation to Causation: A Causal Power Theory, *Psychological Review*, **104**(2), 367-405.

Cheng P.W. and Novick L.R. (1992). Covariation in Natural Induction, *Psychological Review*, **99**(2), 365-382.

Cummins D.D., Lubart T., Alksnis O. and Rist R. (1991). Conditional Reasoning and Causation, *Memory & Cognition*, **19**(3), 274-282.

Glymour C. (1998). Learning Causes: Psychological Explanations of Causal Explanation, *Minds and Machines*, **8**, 39-60.

Goldvarg E., and Johnson-Laird P.N. (2001). Naïve Causality: A Mental Model Theory of Causal Meaning and Reasoning, *Cognitive Science*, **25**, 565-610.

Hadjichristidis C., Stevenson R.J., Over D.E., Sloman S.A., Evans J.St.B.T., Feeney A. (2001). On the Evaluation of *If p then q* Conditionals, *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 381-386). Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Jackson F. (1977). A Causal Theory of Counterfactuals, *Australasian Journal of Philosophy*, **55**, 3-21.

Johnson-Laird P.N. (1994). Mental Models and Probabilistic Thinking, *Cognition*, **50**, 189-209.

Lewis D.K. (1973a). *Counterfactuals*, Blackwell, Oxford, UK.

Lewis D.K. (1973b). Causation, *Journal of Philosophy*, **70**, 556-567.

Rehder B. (1999). A Causal Model Theory of Categorization, *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 595-600). Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Pearl J. (1988). Probabilistic Inference for Intelligent Systems, Morgan Kaufmann, San Mateo, California.

Pearl J. (2000). Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge, UK.

Stevenson R.J. and Over D.E. (1995). Deduction from Uncertain Premises, *The Quarterly Journal of Experimental Psychology*, **48A**(3), 613-643.

Tenenbaum J.B. and Griffiths T.L. (2000). Structure Learning in Human Causal Induction, *Advances in Neural Information Processing Systems*, **13**, MIT Press, Cambridge, Massachusetts.

Yarlett D.G. and Ramscar M.J.A. (in press). A Quantitative Model of Counterfactual Reasoning, *Advances in Neural Information Processing Systems*, **14**, MIT Press, Cambridge, Massachusetts.

Linguistic Cues Enhance the Learning of Perceptual Cues

Hanako Yoshida (hayoshid@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Psychology Department, Indiana University, 1101 E. 10th Street
Bloomington, IN 47405-7007 USA

Abstract

Past research on children's categorizations has centered on the mechanism of children's use of multiple cues in categorization. This paper examines correlations between perceptual cues and linguistic cues. The question asked is a classic one in learning theory: given two redundant cues, does the learner learn more about each than when one cue independently predicts the category? This question has special cogency in the context of children's language learning. We show that when linguistic cues correlated with perceptual cues, children learn more about perceptual cues.

Introduction

Two- and 3-year-old children learn new object names rapidly, often correctly determining the range of instances to which the name applies from just one experience hearing the word used in a single context. Children do this by exploiting multiple cues to meaning. Past research indicates they use both linguistic cues and perceptual cues to figure out the likely meaning of a novel noun. Much of the relevant evidence in this literature concerns the count-mass distinction in English. Count nouns refer to entities conceptualized as discrete objects and as countable. Count nouns obligatorily take the plural (e.g., cups, hopes). Mass nouns refer to entities conceptualized as continuous substances and do not take the plural, but rather mass quantifiers (e.g., some water, a lot of sand). Children use linguistic cues to the count/mass status of a noun to figure out the category to which a novel noun refers. For example, if an entity named with a novel name is presented in a frame that indicates it is a count noun (e.g., "This is a mel"), English-speaking children interpret the word as referring to a discrete entity and typically extend the object name to a class of similarly shaped things (Soja, Carey & Spelke, 1991; Soja, 1992; Landau, Smith and Jones, 1988; Landau, Smith and Jones, 1998; Imai & Gentner, 1997). When the same noun is presented in a frame indicating it is a mass noun (e.g., "This is some mel"), English-speaking children interpret the word as referring to a substance and extend its meaning to entities of the same material (Soja, Carey & Spelke, 1991; Soja, 1992).

Children also use perceptual cues. For example, children extend novel names to new instances by shape when the named entity is solid and rigid (e.g., made from wood) but extend the name to new instances by material when the named entity is nonsolid and non-rigidly shaped. Much previous research has explored which of these kinds of cues dominate by putting them in conflict. In this paper, we ask whether and how they might interact and support children's learning of object names. This is a relevant question for two reasons.

First, linguistic and perceptual cues are highly correlated. This was documented by Samuelson and Smith (1999) who studied the structure of the first 300 nouns commonly learned by English-speaking children. Among these 300 names for common categories, solid things tend to be named by count nouns that refer to things of the same shape, whereas nonsolid things tend to be named by mass nouns that refer to entities of the same material. For learners of English, then, there is a tight correlation between linguistic cues associated with count/mass distinction and perceptual cues that indicates the solidity or non-solidity of an entity.

Second, the evidence suggests that children learn the correlations among perceptual cues, linguistic cues, and category structure as they learn names for common object and substance categories. Specifically, the influence of perceptual cues on children's noun extensions emerges and grows stronger as vocabulary grows. Samuelson and Smith's (1999) data indicate that children learning English do not extend names for solid and nonsolid things differently until children have over 150 nouns. Similarly, English-speaking children's sensitivity to count/mass syntax in the novel noun extension task emerges during this same time period (Soja, 1992).

Two hypotheses

What is the relation between learning about perceptual cues to category organization and linguistic cues to category organization? One possibility is that they are completely independent. Cross-linguistic comparisons of English and Japanese speakers are consistent with this view. Japanese differs from English in that it has no obligatory plural and no counterpart to the count-mass distinction in English. Yet, several studies suggest that

Japanese-speaking children extend names for novel solids and non-solids in pretty much the same way as English-speaking children (e.g., Imai & Gentner, 1997). Thus, children's learning about perceptual correlations and their learning about syntactic cues to category structure (so-called syntactic bootstrapping) may proceed from different learning mechanisms. At the very least, learning about perceptual correlations does not require support from linguistic correlations.

The second contrasting possibility is that learning about perceptual and linguistic cues to category structure are mutually reinforcing. Imai & Gentner's (1997) comparisons of Japanese-speaking and English-speaking children's extensions of names for novel solids and non-solids suggest some subtle differences in the range of items treated as objects and substances, and also some differences in the developmental trend (see, also Yoshida & Smith, 2001.) Further, a number of learning models (Billman & Knutson, 1996; Medin, Altom, Edelson and Freko, 1982; Goldstone, 1998) suggest that the addition of correlated cues bolsters learning about each cue.

Rationale for the experiment

In the present experiment, we examine the role of syntax in children's learning about perceptual cues to category structure through a training study. We attempted to train the solid-nonsolid distinction in Japanese-speaking children who were too young to robustly make the distinction in their novel noun extensions (Shirai, 2000). The design of the four training conditions is shown in Table 1. The linguistic cues are *hitotsu* and *sukoshi*. In the specification of quantitative constructions, (e.g., There is one cup) *hitotsu* is used with objects and *sukoshi* is used with substances. This is thus a natural and salient lexical contrast in Japanese, yet it is one that is neither mandatory nor particularly common. This is in contrast to the count-mass distinction in English, which is mandatory and pervasive. In control conditions, we show that Japanese-speaking children are not sensitive to this contrast, and do not know its implications concerning objects and substances, prior to training.

During the test phase, half of the children in each condition were tested with the linguistic cues and half were not. Here, then, is the question: Will Japanese speaking children show a stronger distinction in their novel noun generalizations between solids and non-solids if trained with these correlated linguistic cues than if trained without them? Is this so even when the linguistic cues are not present during testing? Because our design involves using natural and thus potentially meaningful lexical cues, and because we attempted to accelerate the emergence of a distinction that children eventually learn, we also included two control conditions. Neither involved any training but tested children's sensitivity to the linguistic and perceptual cues.

Experiment

Method

Participants Forty monolingual 2 _ year-old Japanese-speaking children residing in Niigata, Japan, were randomly assigned to the two training conditions and one control condition. Half of the children in each condition participated in the novel noun generalization test at the end of training, either with linguistic cues or without linguistic cues in the tasks, for a total of 6 conditions.

Stimulus Training stimuli consisted of 4 training pairs, two solids and two non-solids, as shown Figure 1. The items in the solid pairs were the same shape but differed substantially in material and color. The items in each nonsolid pair were the same material but differed substantially in color and shape. Stimuli for the test trials consisted of novel solid items made of wood, clay, or sponge, and novel nonsolid items made of hair gel, hand cream, or toothpaste. During test, children were queried about 6 unique test sets, 3 times each for a total of 18 trials per a participant. Each of these test sets contained one exemplar and 3 choice objects unique to that set. One test choice object matched the exemplar in shape only, one matched in material only, and one matched in color only (See Figure 2.) During test, the exemplars were named with a novel name with and without the lexical cues *hitotsu* and *sukoshi* corresponding to the condition to which the participants were assigned.

Design and procedure Children participated in one of the 6 conditions that resulted from crossing the 3 levels of training (training with correlated linguistic cues, training without correlated linguistic cues, and absence of training), 2 levels of linguistic cues (with/without linguistic cues in the task) with 2 levels of solidity (sold/non-solid) for each condition.

Children in the Training condition participated in 10 training sessions over a period of 4 weeks. During each session, the child was presented with the training stimulus repeatedly with/without correlated linguistic cues depending on the child's condition. Each stimulus was shown and introduced by its own novel name, and then played with and repeatedly named for 5 minutes. Each training session took a place for approximately 30 minutes every other day. Notice, this is an implicit category-learning task. During training, children are not required to discriminate between category instances, but attend to the linguistic cues as predictive of category membership.

All children participated in the same test trials where the child was shown an exemplar of a test stimulus set

and told its unique novel name with and without corresponding linguistic cues; "This is (*hitotu/sukoshi*) kochi" or "This is (*hitotu/sukoshi*) kochi". The child was then presented with 3 test objects and was asked to hand the item that can be considered as the name of the exemplar with and without corresponding linguistic cues; "Where is (*hitotu/sukoshi*) kochi?" or "Where is kochi?" Feedback was not provided on these test trials. Since these are novel objects and novel names, success requires knowledge of some general principle---that solids are named by shape and nonsolids are named by material. Are children more likely to notice this regularity when there are correlated linguistic cues?

Results

Each graph in Figure 3 shows the percentage of children's "correct novel word generalization" for solid and non-solid items where "correct" was considered to be shape based for the solids and material based for the non-solids.

Two graphs in the top row represent the performance of children without training sessions. Children generalized all names the same---by shape---treating solids and non-solids equally. This shows that, prior to training, children are not sensitive to the linguistic distinction, nor given these stimuli, to the solid-nonsolid distinction. The two graphs in the bottom represent the performance of children participated in the training sessions with corresponding syntactic cues. The graph on the left shows children's performance without the corresponding syntactic cues in the test trials and one on the right shows children's performance with the corresponding syntactic cues.

Overall, children who had training sessions with the linguistic cues generalize novel names correctly more often than of children without the training sessions, $F(1, 28)=27.2$, $p< .01$. The results suggest that the presence of correlated linguistic cues enhances learning about perceptual cues;

Discussion

The training study revealed the importance of correlated cues in category learning by demonstrating how correlations between linguistic cues and perceptual cues mutually reinforce attention to relevant perceptual cues in the name extension task. The findings fit the traditional idea of how language influences thought.

Whorf. (1956, p.252) wrote

> *And every language is a vast pattern-system, different from others, in which are culturally ordained the forms and categories by which the personality not only communicates, but also analyzes, notices or neglects types of relationship and phenomena….*

If correlated linguistic cues influence what is learned about perceptual cues, Whorf will be right: the language one learns will influence what one notices or neglects to notice about the world.

Table

Table 1: 4 key conditions

	Test with novel stimuli	
	Syntax	**No-syntax**
Trained with syntax	**Correlated cues**	**No in task correlated cues**
No-Trained	**Only in task correlated cues**	**No correlated cues**

Figures

Figure 1: Stimulus items used for the training sessions.

Figure 2: Stimulus items used for the test trials.

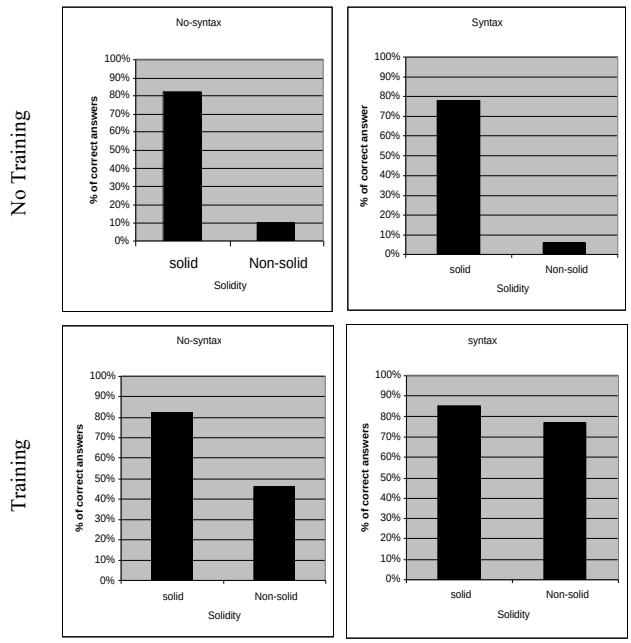


Figure 3: Mean percentage of correct answers.

References

Bauer, P.J. & Mandler, J.M. (1989) Taxonomies and triads: Conceptual organization in one- to two-year olds. *Cognitive Psychology, 21*, 156-184.

Imai, M., & Gentner, D. (1997) A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition, 62* 169-200.

Landau, B., Smith, L. & Jones, S. (1998) Object perception and object naming in early development. *Trends in cognitive science, 2*, 19-24.

Lucy, J. A. (1992) Language diversity and thought: A reformulation of the linguistic relativity hypothesis. Cambridge: Cambridge University Press.

Mandler, J., M., Bauer, P. J., and McDonough, L. (1991) Separating the sheep from the goats; Differentiating global categories. *Cognitive Development*, *3*, 247-264

Markman, E. M. (1991). The whole object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In J. P. Byrnes & S. A. Gelman (Eds.), Perspectives on language and cognition: Interrelations in development (pp-72-106). Cambridge: Cambridge University Press

Quinn, P.C., & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of Experimental Child Psychology, 63*, 189-211,

Rosch, E., & Mervis, C. B.; Gray, W. D.; Johnson, D. M., & Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439

Samuelson, L. & Smith, L. B. (1999) Early noun vocabularies: Do ontology, category structure, and syntax correspond? *Cognition*. *73,* 1-33

Smith, L. B. (1995). Self-organizing processes in learning to learn words: Development is not induction. The Minnesota Symposia on Child Psychology. Volume 28. Basic and applied perspectives on learning, cognition, and development (pp. 1-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, L. B. (2000) Learning how to learn words: An associative crane In Golinkoff, Roberta Michnick; Hirsh-Pasek, Kathy; Bloom, et al. (Eds.) Becoming a word learner: A debate on lexical acquisition. New York: Oxford University Press.

Smith, L. B., Colunga, E., & Yoshida, H. (in press) Making Ontology: Cross-linguistic evidence In Oakes, L. Rakison, D. Category Development

Soja, N. (1992). Inferences about the meanings of nouns: the relationship between perception and syntax. *Cognitive Development, 7*, 29-46.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition, 38,* 179-211.

Yoshida, H. & Smith, L. B. (2001) Early noun lexicons in English and Japanese. *Cognition*, 82, 63-74

Whorf, B. (1956). Language, thought and reality: Selected writings of Benjamin Lee Whorf (J.B. Carroll, Ed.). Cambridge, MA: MIT Press.

Whitman, J & Shirai (2000) Introduction *Journal of East Asian Linguistics, 9*, 315-324

How are Speech and Gesture Related?

Hanako Yoshida (hayoshid@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Raedy M. Ping (rping@indiana.edu)
Elizabeth L. Davis (elldavis@indiana.edu)
Psychology Department, Indiana University, 1101 E. 10th Street
Bloomington, IN 47405-7007 USA

Abstract

People gesture when they speak. Despite considerable attention from a variety of disciplines, the precise nature of the relation between gesture, speech and thought has remained elusive. The research reported here considers two very different hypotheses about the fundamental relationship. By one account, gesture is a consequence of physiological arousal. By another account, gesture use reflects more cognitive processes and is strongly linked to mental representation. This paper seeks the mechanism underlying the link between gesture and speech by showing that both mental representation and physiological arousal are reflected in our gesture use.

Introduction

Communication is the activity of transmitting information about things in the world. In everyday life, we refer to things by employing a number of communicative tools. Language, gestures, facial expressions, and non-linguistic vocalizations are often considered communicative tools. Interestingly, we integrate multiple modalities in our communication and indeed, past research suggests a tight link between modalities, and particularly between speech and gesture.

This study is motivated by two approaches to the study of gesture. McNeill (1992), Kita (2000) and McNeill & Duncan (2000) consider the relationship between the specific forms of gesture and speech, specifically between iconic gesture and speech. An iconic gesture is one in which the speed, motion, or shape of the gesturing hand resembles the meaning being conveyed. McNeil and colleagues consider the emergence of iconicity in gesture as natural, and due to the embodied and imagistic nature of thought. Another approach, however, considers gesture as a consequence not of meaning, but of arousal. Schwarts & Black, (1996) and Iverson & Thelen (1999) particularly suggest that gestures can be explained as an overflow of one's excitement, which is physiologically produced by speech. The idea here is that speech is a motor program, and as the speaker is aroused—by topic or by the very act of speaking—that energy overflows and is evident in other bodily movements.

This present research is concerned with one flashpoint in these accounts: the idea that gesture use fluctuates. At present, the variability in gesture use and in kind of gesturers is not well explained. Yet this very variability should be the key to underlying mechanism. We demonstrate the bi-directional relationship between gesture, thoughts and arousal by confirming that both mental representation and arousal are reflected in and influence iconic gesture use. We experimentally manipulate speech rate as a means of increasing the "motor overflow" from speech. Although speech rate is at best an indirect marker of arousal, we believe it is a good starting point for a mechanistic explanation of variability in gesture use.

Iconicity

Communicative tools in general range from conventional forms that are more arbitrary, to less conventional forms that are more iconic. An example of the conventional forms in traditional verbal languages would be the word "dog" which can hold the meaning only when both senders and receivers are knowledgeable about the rule indicating the label—'a dog' means a dog. This form can also be seen in written language when one spells out letters, d, o, g to refer to a dog. Again, receivers can make sense out of the particular combination of letters only by knowing the convention, that the order and the combination of letters are signifying a dog. People also refer iconically. For example, in verbal language use, one might imitate a sound of a dog to refer to a dog (e.g., "woof, woof!"), and in American Sign Language the word for a dog consists of a pat on one's leg and a snap of one's fingers as if calling a dog.

People also often gesture iconically (Kita, 2000). For example, a speaker might make a circle with one's hands to describe shape and/or the size of a plate to which the person refers. Indeed, even nonverbal primates may gesture to convey meaning. For example, Tanner and Byrne (1996) recently reported that several observed gestures of gorillas are iconic. Another example that highlights both natural emergence and ease of iconic gesture use is an observation made by Goldin-Meadow and Feldman (1977). They observed

communication formation of linguistically deprived deaf children who were unable to acquire oral language naturally and who were not exposed to a standard manual language. They reported that children spontaneously refer to things by using iconic gestures. All these results suggest a tight and natural relationship between gesture and thought. If this is so, then experimental manipulations of the content of thought should have direct and measurable effects on gesture.

We show experimental evidence for this connection in Experiment 1. Participants were asked to read and re-tell the story to their children. A story was about either a carp, which was expected to induce more gesture that engages a hand held in a vertical position or a stingray, which was expected to induce more gesture that engages a hand held in a horizontal position. If mental representations are reflected in our gesture use, then we should expect these different gesture patterns across participants who read the same story about the two different fishes. Moreover, Experiment 1 also serves as baseline measure of gesture production with a normal speech rate. This baseline was used to manipulate participants' speech rate in Experiment 2.

Gesture as motor overflow

The second idea that gestures result from an overflow in the motor program for speech is not one that has received much empirical attention, nor one that has been well specified theoretically. However, this idea suggests that amount of motor activity in speaking should have direct influences on amount of gestures, and this should be thought so independently of the content of speech. Accordingly, to test this idea, in Experiment 2, we entrained speech to either a slow or fast metronome. Do gestures decrease when people speak at slower than normal rates? Do gestures increase given the faster rate of speaking? They should if gestures reflect motor overflow. Finally, does rate of speech influence rate of gesture independent of content? An affirmative answer would suggest two independent driving forces behind gesture: the meaningful content of thought and arousal.

Experiment 1

Method

Participants Twenty mothers and their children whose ages ranged from 26 to 61 months participated in the study.

Procedure Participants were assigned to either the Carp condition or the Stingray condition. The difference in the manipulation between the conditions was a kind of fish used as a hero of the story. Parents were presented with a sheet of paper, which contained either a carp story or a stingray story and were asked to read and retell the story from memory to their child. Participants were also asked to have the child tell the story back to them. While participants were retelling a story, the aspects of their story telling were recorded by a video camera for a later coding. The story used in the study is provided in Table 1.

Result

Each parent's session was coded for several variables: all gestures, two target hand positions and speech rates. All variables were measured as rates, that is, production per minute. In order to code for these variables, the video was played in slow motion; the exact times that a hand movement began and ended were coded. Two target hand positions were counted. The "Carp-swimming" gesture consisted of holding the hand flat palm perpendicular to the horizontal table surface and the "Stingray-swimming" gesture was defined as a gesture in which the hand was flat palm down parallel to the horizontal surface of the table. These gestures were coded as either corresponding or non-corresponding gestures depending on whether the hand position corresponds to the fish in the assigned story (e.g., a carp gesture given a carp story was corresponding gesture and a stingray gesture given a carp story was a non-corresponding gesture.) All gestures were independently scored by two observers.

Participants produced gestures (target gestures plus all other gestures) at a rate of 4.1 gestures per a minute. However, as shown in Figure 1, participants produced more target gestures that corresponded to the story than gestures that did not correspond. In addition, storytellers held their hands in a horizontal position more often when telling the story about the stingray than when telling one about a carp. Finally, the average speech rate was 141.0 words per a minute.

Experiment 2

Method

Participants Twenty mothers and their children whose ages ranged from 23 to 60 months participated in the study.

Procedure The procedure was the same as that of Experiment 1 with an exception. Participants were assigned to one of four conditions: Carp condition, fast metronome, Carp condition with slow metronome, Stingray condition with fast metronome and Stingray condition with slow metronome. The speeds of the metronome were selected based on the average speech rate observed in Experiment 1. Since the normal speech rate in Experiment 1 was 140 words per a minute, we selected 150 beats per a minute for the Fast condition and 80 beats per a minute for the Slow condition. We were conservative in attempting to entrain speech to a

faster rate in this first entraining experiment in an effort to maximize our success in entraining speech. The parent storytellers were not told about the purpose of the metronomes, but were instead told to ignore them. The task of individual participants was identical to that of Experiment 2.

Table

Table 1: Story used for the study: The Adventures of (Carl the Carp / Steve the Stingray)

(Carl the Carp / Steve the Stingray) was a happy, friendly (carp with a round, silver body, big, round eyes, and a flowing, fan-like tail / stingray with a flat, gray body, beady black eyes, and a long, whip-like tail) who loved swimming with his other (carp / stingray) friends. They especially liked swimming into the depths of the water and returning to the surface. One day, when (Carl the Carp / Steve the Stingray) and his friends were enjoying another adventure in the depths of the water, they sensed some danger. They were eating lunch when, all of a sudden, there was a giant shadow looming over them. (Carl the Carp / Steve the Stingray) turned around, but it was too late! He found himself inside the belly of a big fish. He thought to himself, "This is the end of (Carl the Carp / Steve the Stingray). I'll never be able to swim around with my (carp / stingray) friends and have fun again." Just as he was thinking this, he heard a loud rumbling and was pushed out from the fish's stomach. His friends, who had been hiding in the rocks, swam cautiously towards him to make sure he was okay. "(Carl / Steve)," one of his friends said, "that big fish just burped you up! Are you all right?" "Yes," a relieved (Carl the Carp / Steve the Stingray) said, "I'm fine. Let's finish our adventure." Then, (Carl the Carp / Steve the Stingray) and all of his (carp / stingray) friends swam happily off for another fun adventure.

Results

First, we did successfully manipulate speech rate, $t = -2.44$, $p < .05$. The average number of words per minute was 145.5 in the Fast condition and 120.0 in the Slow condition. Second, as can be seen in Figure 2, participants produced more gestures per a minute in the Fast condition than they did in the Slow condition. This indicates that gesture production is tied to speech rate.

As shown in Figures 3 and 4, content effects are seen in the both Fast and Slow conditions, but the effect is reliable only in the Fast condition (for Fast condition; $t = 2.198$, $p < .05$, and for Slow condition; $t = 1.405$, $p = .18$.) Participants in both Fast and Slow conditions produced more gestures that corresponded to the story that they retold than gestures that did not correspond. The fact that speech entrained to a slow metronome, 80 beats per a minute, decreases gesture use and weakens the content effect on iconic gestures suggests a direct link between arousal and content.

Discussion

The two experiments yield 3 main results. First, the content of thought directly influences hand position. Thinking about carps leads one to holds one's hand perpendicular to the tabletop. Thinking about stingrays leads to hands held in a horizontal position over the table. By experimentally manipulating the content of thought, one manipulates hand position. This is consistent with the proposal that gestures iconically represent meaning. Second, speech and gesture are readily entrained to a rhythmic beat. And, faster rates of speech lead to more gestures and slow rates of speech lead to fewer gestures. This is consistent with the idea of gestures emerging as a consequence of arousal. Amount of energy appears to be is a driving force in gesture. Third, arousal (or rate of speech) appears linked to the frequency of iconic gestures. This suggests, contrary to the motor overflow hypothesis, that the relevant activation concerns meaningful content and not just the speech motor plan. High activation leads to more iconically related gestures as if high arousal leads to more highly activated meanings that emerge in both spoken words and in meaning related gestures.

Figures

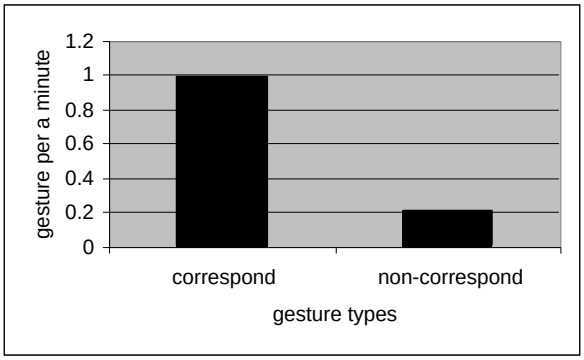


Figure 1: correspond and non-correspond gesture

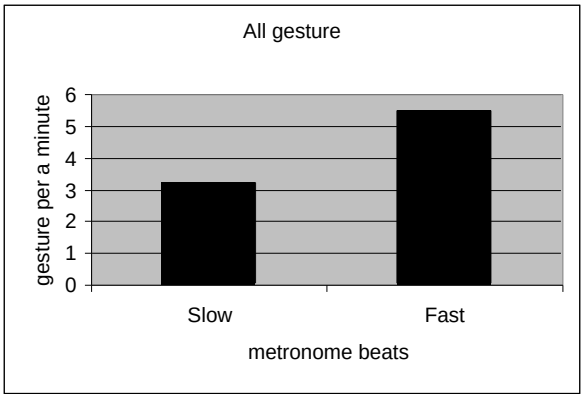


Figure 2: all gestures per minute.

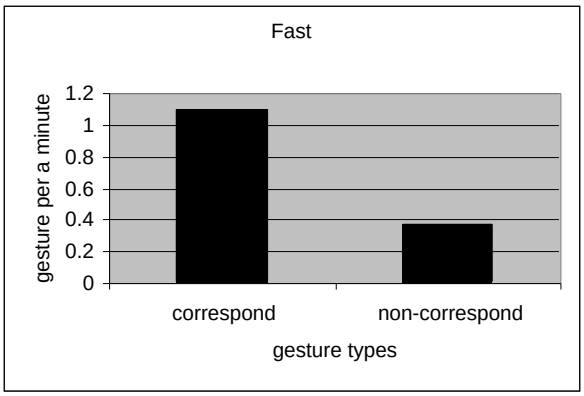


Figure 3: correspond and non-correspond gesture in Fast condition

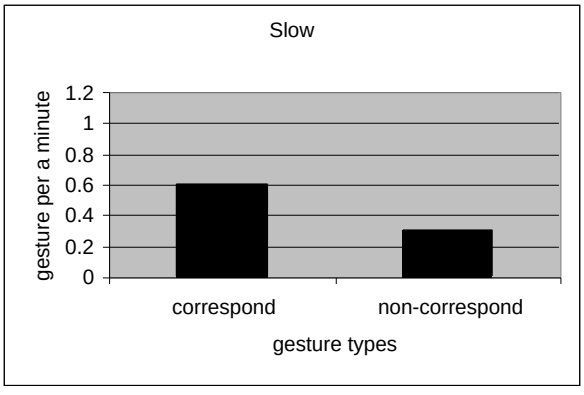


Figure 4: correspond and non-correspond gesture in Slow condition

References

Goldin-Meadow, Susan & Feldman, Heidi (1977) The development of language-like communication without a language model. Science. 197(4301): 401-403. US: American Assn. for the Advancement of Science.

Iverson, J.M. & Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence of speech and gesture. Journal of Consciousness Studies, 6 (11-12), 19-40.

Kita S. (2000). Why Do People Gesture? Cognitive Studies, 7 (1), 9-21

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and gesture, pp. 162-185. Cambridge, UK: Cambridge University Press.

McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.

McNeill D. & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), Language and Gesture: Window into Thought and Action, 141-161. Cambridge: Cambridge University Press, in press.

Schwartz, D. L., & Black, J. B. (1996) Shutting between depictive models and abstract rules: Induction and fallback. Cognitive Science, 20, 457-498

Tanner, J. E. & Byrne, R. W. (1996) Representation of action through iconic gesture in a captive Lowland gorilla. Current Anthropology, 37, 162-173

Toward An Action Based Taxonomy of Human Errors in Medicine

Jiajie Zhang[1], Vimla L. Patel[2], Todd R. Johnson[1], & Edward H. Shortliffe[2]

[1] School of Health Information Sciences
University of Texas at Houston
7000 Fannin, Houston, TX 77030
{Jiajie.Zhang, Todd.R.Johnson}@uth.tmc.edu

[2] Department of Medical Informatics
Columbia University
622 West 168th Street, New York, NY 10032
{Patel, Shortliffe}@dmi.columbia.edu

Abstract

One critical step in addressing and resolving the problems associated with human errors is the development of a cognitive taxonomy of such errors. In the case of errors, such a taxonomy may be developed (1) to categorize all types of errors along cognitive dimensions, (2) to associate each type of error with a specific underlying cognitive mechanism, (3) to explain why, and even predict when and where, a specific error will occur, and (4) to generate intervention strategies for each type of error. Based on Reason's (1992) definition of human errors and Norman's (1986) cognitive theory of human action, we have developed a preliminary action-based cognitive taxonomy of errors that largely satisfies these four criteria in the domain of medicine. We discuss initial steps for applying this taxonomy to develop an online medical error reporting system that not only categorizes errors but also identifies problems and generates solutions.

1. Introduction

The medical error report from the Institute of Medicine (Kohn, Corrigan, & Donaldson, 1999) has greatly increased people's awareness of the frequency, magnitude, complexity, and seriousness of medical errors. As the 8th leading cause of death in the US with 98,000 preventable deaths per year, ahead of motor vehicle accidents, breast cancer, or AIDS, medical errors need immediate attention from academic, healthcare, and government institutions and organizations. To achieve the goal of reducing medical errors by 50% in five years set by the former Clinton Administration, we need to understand the fundamental causes of medical errors such that medical errors can be prevented or greatly reduced systematically at a large scale. In our opinion, cognitive factors are fundamental in medical errors. This can be seen from the view of the healthcare system hierarchy and the view of action chains.

Cognitive factors are critical at various levels of the healthcare system hierarchy of medical errors (Figure 1). At the lowest core level, it is individuals who trigger errors. Cognitive factors of individuals play the most critical role here (Reason, 1992). At the next level, errors can occur due to interactions between an individual and technology. This is an issue of human-computer interaction where cognitive properties of interactions between human and technology affect and sometimes determine human behavior (Helander, Landauer, & Prabhu, 1997; Zhang, 1997; Zhang & Norman, 1994). At the next level, errors can be attributed to the social dynamics of interactions between groups of people who interact with complex technology in a distributed cognitive system. This is the issue of distributed cognition and computer-supported cooperative work (Baecker, 1993; Hutchins, 1995a, 1995b; Zhang, 1997). At the next few levels up, errors can be attributed to factors of organizational structures (e.g., coordination, communications, standardization of work process), institutional functions (e.g., policies and guidelines), and national regulations. At these higher levels, cognitive factors also play some roles. Although the properties at the six levels can be to some extent studied independently, a cognitive foundation for the system is essential for a complete and in-depth understanding of medical errors.

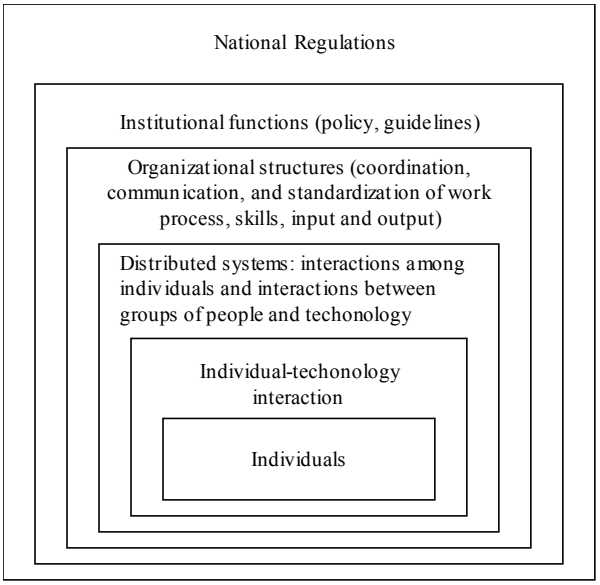


Figure 1. The system hierarchy of human errors in medicine

From the view of action chains, the critical roles of cognitive factors in medical errors are also clear. Figure 2 shows the chain of events and factors that lead to an error in a system. It is clear that individuals are at the last stage of the chain, although the individuals may not be the root cause of the error. If the chain of events can be stopped at the in-

dividual's stage through cognitive interventions, errors could be potentially prevented.

Medical errors are human errors in healthcare. By definition (Kohn et al., 1999; Reason, 1992), human errors are errors in human actions. Human actions are primarily cognitive activities. It is not surprising to see that human errors occur primarily due to inadequate information processing in cognitive tasks (Bogner, 1994; Norman, 1981; Reason, 1992; Woods, Johannesen, Cook, & Sarter, 1994). In order to prevent or greatly reduce medical errors, it is critical to understand the underlying cognitive mechanisms of medical errors.

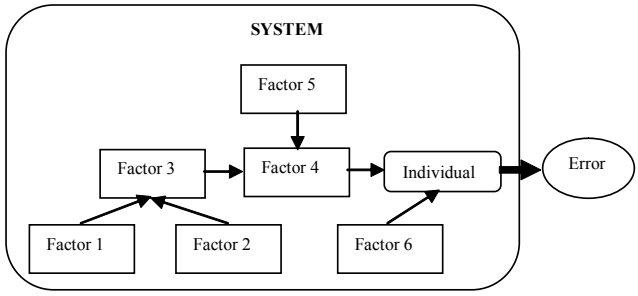


Figure 2. The chain of events leading to an error

2. Theoretical Background

To understand the cognitive mechanisms underlying medical errors, we first need to develop a cognitive taxonomy of medical errors that can (1) categorize all types of medical errors along cognitive dimensions, (2) associate each type of medical error to a specific underlying cognitive mechanism, (3) explain why and even predict when and where a specific error will occur, and (4) generate intervention strategies for each type of error.

The purpose of this paper is to develop an action based cognitive taxonomy that can be potentially expanded to include all four features listed above.

2.1. Reason's definition of human error

Reason's (Reason, 1992) definition of human error is the most widely accepted: an error is a failure of achieving the intended outcome in a planned sequence of mental or physical activities. According to Reason, human errors are divided into two major categories: (1) slips that result from the incorrect execution of a correct action sequence and (2) mistakes that result from the correct execution of an incorrect action sequence. In comparison with mistakes, slips have been extensively studied and better understood (for reviews, see Norman, 1986; Reason, 1992).

2.2. Norman's action theory

To be comprehensive, descriptive, predictive, and generalizable, a cognitive taxonomy should be based on a sound cognitive theory that has explanatory and predictive power. Since human errors are defined as errors in human actions, a cognitive theory of human actions can provide the theoretical foundation for the cognitive taxonomy. In our opinion, the cognitive theory of human action most appropriate for medical errors is the seven-stage action theory developed by Norman (Norman, 1986, 1988) and refined by Zhang and colleagues (Zhang, 1987; Zhang, Patel, & Johnson, in press). The seven-stage action theory is shown in Figure 3, with a demonstration showing the action of deleting a file on a DOS system. According to this theory, any action has seven stages of activities: (1) establishing the goal (e.g., "delete file"); (2) forming the intention (e.g., "use remove command"); (3) specifying the action specification (e.g., "remove ../../home/paper/talk_old.ver1"); (4) executing the action (e.g., "typing command text, hit return"); (5) perceiving the system state (e.g., "prompt symbol :\>, no feedback"); (6) interpreting the state (e.g., "nothing happened"); and (7) evaluating the system state with respect to the goals and intentions (e.g., "form sub-goal to find out current state of the system").

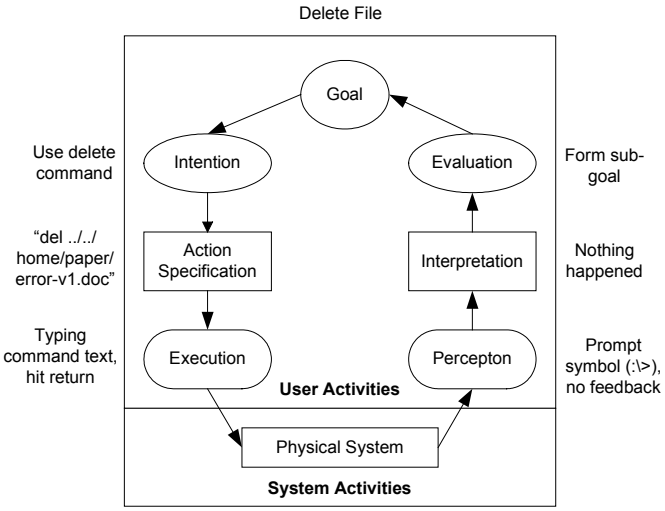


Figure 3. Norman's seven-stage theory of action.

3. The Cognitive Taxonomy

Reason developed one taxonomy of human errors (Reason, 1992); however, it was not based on a systematic theory of human action; it was primarily for slips, not for mistakes; and it has not been systematically applied to medical settings. Norman's (Norman, 1986) seven-stage action theory was developed for the study of human-computer interaction and the design of user interfaces--it has not been applied to the study of errors.

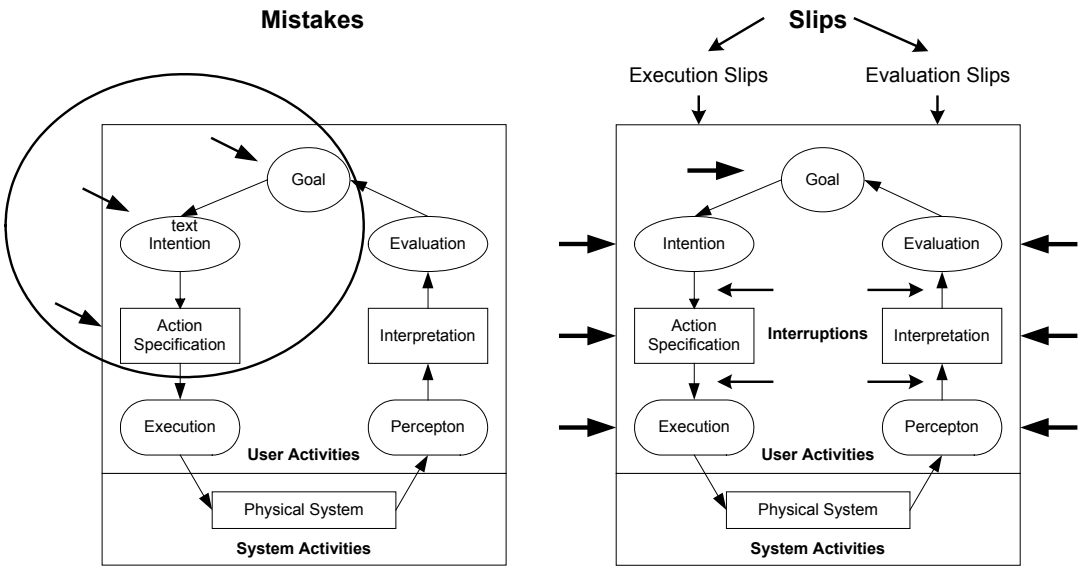


Figure 4. Slips can occur at all stages, whereas mistakes can only occur at the first three stages.

The cognitive taxonomy we develop here is an application and extension of Norman's action theory to the categorization of medical errors. It is an action-based cognitive taxonomy. This taxonomy covers all types of human errors, because a human error is an error in an action and any action has to go through the seven stages. According to our taxonomy, errors can occur at any of the seven stages of action and between any two adjacent stages: due to incorrect translation from goals to intentions, incorrect action specifications from intentions, incorrect execution of actions, misperception of system state, misinterpretation of data perceived, and misevaluation of interpreted information with regard to the goal of the task. Unlike other taxonomies, our taxonomy specifies the places where mistakes and slips may occur (Figure 4). A slip is the incorrect execution of a correct action sequence. Slips can occur at all seven stages of action and between stages. Mistakes, however, can only occur at the first three stages of action because a mistake is the correct execution of an incorrect action sequence and only the first three stages can contribute to the formation of an incorrect action sequence.

3.1. Slips

Under our cognitive taxonomy, slips can be divided into execution slips and evaluation slips (see Figure 4 and Table 1).

Execution slips are associated with the execution of an action. They occur at stages of Goal, Intention, Action Specification, and Execution. For the slips at each stage, there are corresponding cognitive mechanisms. A correct goal could be distorted due to its strongly shared schema with another irrelevant goal. A correct intention could be deactivated due to memory decay or swapped by another irrelevant intention due to similarity of schemas. A correct action specification could be distorted due to many factors such as attention shift, situational stimulation, etc. The execution of an action sequence could misfire due to memory and attention problems or various environmental factors. Table 4 shows a list of possible cognitive mechanisms for slips at each of the stages.

Similarly, evaluation slips are associated with the evaluation of the outcomes of an action. They occur at the stages of Perception, Interpretation, and Evaluation. There are also corresponding cognitive mechanisms associated with the slips at each of these stages. The outcome of an action might be impossible to perceive, hard to perceive, or perceived in an incorrect way. The interpretation stage may also induce errors due to prior knowledge, lack of context, or as a direct result of misperception. The evaluation stage may fail due to insufficient feedback, delayed feedback, information overload, memory failure, and other factors.

Table 1 shows not just the types of slips under the cognitive taxonomy but also examples of slips in each category and potentials solutions that can prevent the slips from happening.

3.2. Mistakes

Under our cognitive taxonomy, mistakes are categorized into goal mistakes, intention mistakes, and action specification mistakes. These correspond to the first three stages in the action cycle where mistake can occur. Goal mistakes and intention mistakes are mostly knowledge-based mistakes, such as faulty conceptual knowledge, incomplete knowledge, biases and faulty heuristics, incorrect selection of knowledge, information overload, etc. Action specification mistakes are

mostly rule-based mistakes, such as misapplication of good rules, encoding deficiencies in rules, action deficiencies in rules, dissociation between knowledge and rules, etc.

Table 1. An Action Based Cognitive Taxonomy: Slips

	Stage in Action Cycle	Examples	Cognitive mechanisms	Potential solutions
Execution Slips	Goal slips	*A doctor was called out of the room to answer an urgent call and afterwards he went to the room of a different patient who was next in the queue*. (Loss of activation)	•Loss of activation •Cross talk (concurrent) •Cross talk (sequential) •Altered goal •Delayed activation •Overflow of goal stacks	•Provide memory aids •Reduce multitasking •Reduce interruptions •Reduce goal stacks •Train users
	Intention Slips	*"I went into my bedroom intending to fetch a book. I took off my rings, looked in the mirror and came out again—without the book."* (Loss of activation)	•Loss of activation •Cross talk (concurrent) •Cross talk (sequential) •Reversal of schema •Activation of incorrect schema	•Provide memory aids •Reduce multitasking •Situated actions •Reduce interruptions
	Action Specification Slips	*IL-11 (Oprelvekin , or Interleukin-eleven) was misinterpreted as IL-2 (Aldesleukin, or Interleukin-two). 11 was read as the Roman numeral two.* (Associative activation)	•Associative activation •Failure of retrieval •Sequence mutation •Situated activation •Description •Cross talks	•Automation •Decision support •Situated actions •Train users •Direct action
	Execution slips	*"I meant to turn off the antibiotics IV only, but turned off the infusion pump completely*." (Double capture)	•Capture •Double capture •Perceptual confusion •Deviation of motor skills •Misfiring •Omission	•Automation •Visualization •Display design •Reduce interruption •Memory aids
Evaluation Slips	Perception slips	*A patient died of liquid aspiration Because the water trap connected with a tube had no mechanism to protect against reflux to patient's trachea, and there was no feedback in the system.* (Lack of perception)	•Lack of perception •Misperception •Mis-anticipation	•Direct perception •Immediate feedback
	Interpretation slips	*A yellow flashing light on a medical device was interpreated as non-critical when it really meant critical.* (Misinterpretation)	•Misinterpretation •Default schema •Confirmation bias •Information overload •Loss of memory	•Display design •Decision support •User training •Memory aids •Situation awareness
	Evaluation slips	*A nurse repeated radiation therapy to a patient three times in a row, due to poor feedback. The patient died three months later*. (Lack of feedback)	•Lost goal •Insufficient information •Evaluating different goal •Information overload •Lack of feedback	•Memory aids •Display design •Action tracking •Information reduction

Table 2. An Action Based Cognitive Taxonomy: Mistakes

	Stage in Action Cycle	Examples	Cognitive Mechanisms	Potential solutions
Knowledge-based Mistakes	Goal mistakes	*Stick with a diagnosis that was generated through a large investment of time and effort even if there was evidence indicating other possibilities*. (Biases)	•Misdiagnosis •Faulty conceptual knowledge •Incomplete knowledge •Biases •Faulty heuristics	•Training •Education •Representational Aid •Decision support
	Intention mistakes	*A physician treating a patient with oxygen set the flow control knob between 1 and 2 liters per minute, not realizing that the scale numbers represented discrete, rather than continuous, settings.* (Incorrect knowledge)	•Incorrect selection of knowledge •Misapplication of knowledge •Information overload •Incorrect knowledge	•Training •Education •Decision support •Information reduction •Display design •Representational Aid
Rule-based Mistakes	Action Specification mistakes	*Strange burn scars appeared in post-operative patients in a hospital. The problem was caused by electric discharge of the device that was not grounded. The device has a blinking red to signal for the problem, but the device operators did not know the meaning of the signal.* (Incomplete knowledge)	•Misapplication of good rules •Encoding deficiencies in rules •Dissociation between knowledge and rules •Action deficiencies in rules •Incomplete knowledge	•Decision support •Automation •User training •Representational Aid

Table 2 shows not only the types of mistakes under the cognitive taxonomy but also examples of mistakes in each category and potentials solutions that can prevent the mistakes from happening. In comparison with slips, mistakes are more complex and less understood.

Most studies about mistakes in the past were byproducts of studies of reasoning biases and heuristics in decision-making tasks (Hogarth & Einhorn, 1992; Tversky & Kahneman, 1974). Recently there have been a growing number of studies that explicitly examine various types of mistakes in medicine (Patel & Kaufman, 2000; Patel, Lloyd, & Melanson, 2000; Patel & Ramoni, 1997). We expect to see more studies of this kind and we will expand our taxonomy to accommodate new data and theories.

4. Discussion and Conclusion

One critical step towards reducing medical errors in particular and human errors in general is a cognitive taxonomy of errors that can (1) categorize all types of medical errors along cognitive dimensions, (2) associate each type of medical errors to a specific underlying cognitive mechanism, (3) explain why and even predict when and where a specific error will occur, and (4) generate intervention strategies for each type of error. Based on Reason's (Reason, 1992) definition of human errors and Norman's (Norman, 1986) cognitive theory of human action, we developed a preliminary action-based cognitive taxonomy of medical errors that to some extent satisfy these four criteria. Our taxonomy can categorize all types of errors (slips and mistakes) according the stages of the action cycle. We have identified a set of cognitive mechanisms (though not exhaustive) that underlie each type of slip or mistake. Our taxonomy can also explain why a specific error occurs, although we have not developed the taxonomy in enough detail to make predications on when and where an error will occur. Finally, at a high and conceptual

level, we have generated a set of possible solutions addressing each type of errors.

One important practical implication of the cognitive taxonomy of medical errors is that it can provide systematic, principled methods for the design of medical error reporting systems. Current medical error reporting systems are mostly based on free text in an unstructured format. Medical error data collected in this way are rarely useful for the detection of patterns, discovery of underlying factors, and generation of solutions, because user entered free text do not contain the right types of information needed for interventions and is difficult to analyze in a systematic way. Medical error reporting systems should not be merely record keeping systems. They should be systems for the identification of problems and generation of solutions. We are currently developing an online medical error reporting system that is based on the cognitive taxonomy we have been developing. In this system, questions and inquiries are generated to encode cognitively relevant information; the categorization of errors is along relevant cognitive dimensions; and it is designed to generate immediate recommendations on possible intervention strategies.

5. References

Baecker, R. M. (Ed.). (1993). *Readings in groupware and computer-supported cooperative work: Assisting human-human collaboration*. San Francisco: Morgan Kaufmann.

Bogner, M. S. (Ed.). (1994). *Human error in medicine*. Hillsdale, NJ: Erlbaum.

Helander, M. G., Landauer, T. K., & Prabhu, P. V. (Eds.). (1997). *Handbook of human-computer interaction* (2nd ed.). New York: North-Holland.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Hutchins, E. (1995a). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speed. *Cognitive Science, 19*, 265-288.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (1999). *To err is human*. Washington, DC: National Academy Press.

Norman, D. A. (1981). Categorization of Action Slips. *Psychological Review, 88*, 1-15.

Norman, D. A. (1986). Cognitive engineering. In S. W. Draper (Ed.), *User centered system design*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.

Patel, V. L., & Kaufman, D. R. (2000). *Conceptual and procedural errors in medical decision-making*. Proceedings of the Cognitive Society Conference.

Patel, V. L., Lloyd, S. J., & Melanson, P. (2000). *Decision making in emergency care: The use of data and heuristics*: Centre for Medical Education, McGill University.

Patel, V. L., & Ramoni, M. (1997). Cognitive models of directional inference in expert medical reasoning. In R. Hoffman (Ed.), *Expertise in context*. Menlo Park, CA: AAAI Press.

Reason, J. (1992). *Human error*. Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Woods, D. D., Johannesen, L., Cook, R. I., & Sarter, N. (1994). *Behind human error: Cognitive systems, computers and hidsight*. Dayton, OH: Crew Systems Ergonomic Information and Analysis Center, WPAFB.

Zhang, J. (1987). *The effect of the timing of interruption on human action* (Unpublished Report). San Diego: University of California, Department of Psychology.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21*(2), 179-217.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*(1), 87-122.

Zhang, J., Patel, V. L., & Johnson, T. R. (in press). Medical error: Is the solution medical or cognitive? *Journal of American Medical Informatics Association*.

Why do metaphors seem deeper than similes?

Sergey S. Zharikov (s-zharikov@northwestern.edu)
Northwestern University, Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Dedre Gentner (gentner@northwestern.edu)
Northwestern University, Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Abstract

Figurative expressions in metaphor form (e.g., *Marriage is a journey*) seem stronger and deeper than expressions in simile form (e.g., *Marriage is like a journey*). We ran a study to examine the nature of these judgments. Participants read short paragraphs describing either object attributes or relational structure and then made a forced choice of the grammatical form of a figurative expression mentioning the target concept referred to in the passages. The results showed that the metaphor form was chosen more often (1) for expressions with conventional bases, and (2) when figurative statements followed contexts containing relational information. We speculate about a possible linkage between conventionalization and relationality.

Introduction

Nominal figurative statements can be expressed in two ways – in simile form (*X is like Y*) and in metaphor form (*X is Y*). Although the two grammatical forms largely serve the same purpose (showing that one entity is figuratively similar to another), people report that expressions in metaphor form feel more profound and express stronger claims than expressions in simile form. For example, saying *Her heart is a stone* feels deeper than *Her heart is like a stone*. Further, several studies (Gibb & Wales, 1990; Bowdle, 1998; Aisenman, 1999) have found that if people are asked to make a choice between an expression in metaphor form or the same expression in simile form, the simile form is chosen more often. It appears that people are more conservative in using metaphors than in using similes.

The greater force of the metaphoric form was noted by Glucksberg and Keysar (1990), who proposed that the metaphor form is the basic form of figurative statements and that similes are understood as variants of metaphors. Noting that the grammatical form of metaphors matches that of category inclusion statements, they suggested that metaphors in fact function as category inclusion statements, where the category is an abstraction that can be accessed or created from the metaphor's base concept. (We will use the terms *target* and *base*, respectively, for the X and Y terms, of a figurative expression *X is [like] Y*.) There has been debate concerning the processing implications of this theory, but for our purposes the key point is Glucksberg and Keysar's insight that the grammatical form of figurative statements has psychological force, with metaphor being the stronger, more categorical form. This paper examines the reasons for this phenomenon.

Two recent theories have proposed different explanations for the simile-metaphor difference. One account singles out the conventionality of the base term; the other, the relationality/attributionality of the metaphor's interpretation. The first account, the Career of Metaphor hypothesis (Bowdle & Gentner, 1999; Gentner & Bowdle, 2001) suggests that the difference lies in the conventionality of the base term: figuratives with conventional bases are expressed as metaphors, and those with novel bases are expressed as similes. The second account, Aisenman's (1999) Relational Precedence hypothesis, suggests that the difference is due to the kind of interpretation the expression receives: relational interpretations are stated as metaphors, and attributional interpretations are stated as similes.

In their research on metaphor processing, Gentner and Wolff (1999) proposed an important distinction between newly minted figuratives and conventionalized figuratives. According to the Career of Metaphor hypothesis, figuratives with novel bases, such as *An encyclopedia is (like) a uranium mine,* are processed by comparison between the target and the literal meaning of the base. In contrast, figuratives with conventional bases, such as *An encyclopedia is a goldmine,* can be processed by alignment with a conventional abstraction (e.g., *a source of something valuable*) associated with the base term. The key difference between novel and conventional bases is that the representations of conventional bases include a secondary metaphoric meaning along with the original literal meaning. They have become polysemous. In contrast, representations of novel bases contain only a literal meaning.

Gentner and Wolff (1997) proposed that conventional metaphoric meanings are created over time as a result of repeated comparisons of different

target terms with the same base. The idea is that through progressive alignments of the base, a set of properties or a relational schema belonging to the base emerges as a separable abstraction. This can become an additional word sense – a kind of metaphoric category associated with the base.

Bowdle's Grammatical Concordance principle links the Career of Metaphor hypothesis with the simile-metaphor distinction. It states that metaphoric expressions are interpreted by the process of structural alignment (Gentner & Markman, 1997), but the nature of the invited alignment differs for metaphors and similes. The simile form invites directly aligning the literal base and target concepts (e.g., *encyclopedia* and *gold mine* in the above example), whereas the metaphor form suggests that the listener should first access the abstraction associated with the base – e.g., *source of something valuable* -- and then align it with the target representation. Consistent with this explanation, Gentner and Bowdle (2001) found that novel metaphors are slow to process. This follows from the claim that such statements lead to a false start in processing. For example, hearing *That encyclopedia is a uranium mine* is infelicitous, because there is no conventional abstraction associated with uranium mines.

Thus, the claim is that (1) repeated alignments can lead to the formation of an abstraction, and (2) figurative statements can occur in metaphor form only when there is existing abstraction (or metaphorical category) associated with the base. Perhaps the most striking evidence for this claim is Bowdle's (1998) study showing 'in vitro' conventionalization. After seeing novel bases in parallel comparisons with three target terms in simile form, subjects preferred to express further statements involving that base in metaphor form. They also (mis) recalled the statements they had seen as having been in metaphor form. Gentner and Bowdle (2001) found that as figurative statements became increasingly conventional, there is a shift in people's preference from the simile form to the metaphor form.

A second explanation for the subjective differences in perception of similes and metaphors was recently offered by Aisenman (1999). She extended Gentner's (1988; Gentner & Clement, 1988) distinction between attributional and relational comparisons and suggested that people primarily use the metaphor form to highlight common relations between the base and target, and the simile form to highlight common attributes (Aisenman, 1999). Thus, the metaphor form is likely to convey a deep common system of relations. This theory fits well with the intuition that metaphors often seem more profound than similes. In her study, Aisenman presented subjects with base and target terms and asked whether they would be more likely to put sentences with those terms in simile or metaphor form. When the base and target shared mostly surface attributes (e.g., *The sun is (like) an orange* – both are round and orange), participants preferred to state sentences in simile form. When the base and target shared common relational structure (e.g., *Television is (like) a magnet* - both attract), participants were more likely to use the metaphor form. Aisenman's results suggest that the metaphor form is preferred for relational commonalities.

There are thus two accounts for form differences in figurative language: metaphors tend to be preferred over similes (a) when the base is conventional or (b) when the interpretation is relational. To compare these accounts, we varied both factors – conventionality of the base and the type of commonalities between the base and target – and obtained people's preferences for stating figurative expressions in simile or metaphor form.

Experiment 1. Context Priming

We selected 20 metaphors from prior metaphor studies (Ortony, 1979; Gentner & Clement, 1988; Aisenman, 1999). The metaphors used were classified as double metaphors (Gentner & Clement, 1988) in that they permitted both attributional and relational interpretations. We presented subjects with short paragraphs describing the target, focusing either on its attributes or on its relational structure. Examples of relational and attributional contexts are listed in Table 1. Then, participants were asked to choose which of the two figurative sentences they preferred. Both sentences featured the target coupled with the same base and differed only in that one of them was a simile and one was a metaphor. Half the bases were novel, and half were conventional. Conventionality of the base was operationalized as having the metaphoric meaning listed in the Merriam-Webster Collegiate dictionary. The base terms never appeared in the contexts preceding the simile and metaphor statements. Table 1 shows a sample stimulus with a conventional base.

The Career of Metaphor account predicts that people would be more likely to prefer the metaphor form for statements with conventional rather than novel bases. Aisenman's Relational Precedence hypothesis predicts that people would be more likely to prefer the metaphor form when given the paragraph priming the relational interpretation.

Method

Sixty-four Northwestern University undergraduates were presented with 20 short paragraphs. Each paragraph supported either an attributional or relational interpretation of a figurative expression. After reading the paragraph, participants chose between simile and metaphor forms as shown in Table 1 and were asked to choose the sentence they preferred by circling it. Four

random orders were used across participants. Whether the sentence in simile or metaphor form was presented on the left side of the page was counterbalanced.

Table 1: Example of attributional and relational contexts

Conventional base

Attributional interpretation:

Mr. White, a sociologist, is writing an article about poverty in urban America. He considers poverty a horrible blight on our society and argues that the government must intervene with a welfare reform. He thinks that

Poverty is a disease. Poverty is like a disease.

Relational interpretation:

Mr. White, a sociologist, is writing an article about poverty in urban America. He considers poverty to be increasing and argues that, unless the government intervenes with a welfare reform, poverty will spread further. He thinks that

Poverty is like a disease. Poverty is a disease.

Results

We computed the number of metaphor choices by coding preference for simile form as 0 and preference for metaphor form as 1. Analysis of variance performed with base conventionality and context type as between-subjects factors showed a significant effect of base conventionality ($F_{1, 39} = 7.50$, $MSE = 0.31$, $p < 0.01$). The proportion of metaphor form choices was significantly higher for statements with conventional bases ($M_C = 0.39$) than for statements with novel bases ($M_N = 0.22$). The number of metaphor preferences was significantly lower than chance for both novel and conventional bases ($p < 0.05$).

We also obtained a marginally significant effect of preceding context type ($F_{1, 39} = 3.66$, $MSE = 0.15$, $p = 0.06$). Statements following relational contexts were preferred in metaphor form more often than statements following attributional contexts ($M_R = 0.37$, $M_A = 0.24$). The number of metaphor preferences was significantly lower than chance for both relational and attributional contexts ($p < 0.05$). The results are summarized in Figure 1.

The preference for metaphoric form for relational information was only marginally significant. However, an item analysis indicated a disparity in the quality of the items used. Some items were strongly preferred in simile form (e.g., only one out of 64 participants chose to put *Titanium chips are (like) diamonds* in metaphor form). It thus seemed possible that not all the items were suitable as metaphors. To ensure that the Relational Precedence view was fairly tested, we removed items that were put in metaphor form by less than seven participants (2 with novel bases and 2 with conventional bases). An ANOVA performed on the remaining items yielded a significant effect of context type ($F_{1, 31} = 5.10$, $MSE = 0.16$, $p < 0.05$) in addition to the significant effect of base conventionality ($F_{1, 31} = 10.15$, $MSE = 0.32$, $p < 0.01$). The interaction between base conventionality and context type was not significant.

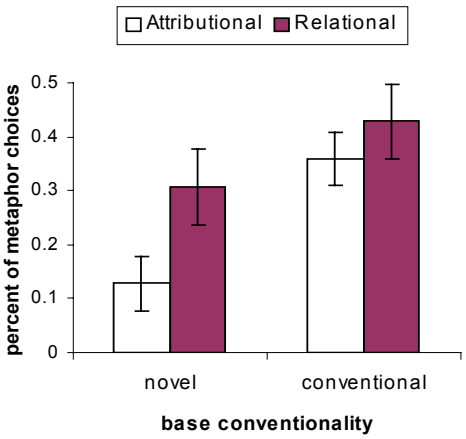


Figure 1: Proportion of metaphor choices for novel and conventional bases. Error bars show standard error.

Experiment 2

The results of the first study offered support for both the Career of Metaphor account – in which the metaphor-simile distinction is one of conventionality – and the Relational Precedence hypothesis. However, one concern here is to what extent the results simply reflect the nature of the materials. First, although the figurative expressions used in Experiment 1 allow both relational and attributional interpretations, it is possible that people may prefer one kind of interpretation over the other. Previous studies have demonstrated that people find relational interpretations of figurative expressions more interesting and apt (Clement & Gentner, 1988). Second, and more importantly, it is possible that the conventional metaphors we used were biased in favor of either relational or attributional interpretations, relative to the novel metaphors.

To calibrate the materials, we gave the figurative expressions used in Experiment 1 to a new group of participants, either in simile or metaphor form, and asked the participants to rate how much they agreed with the relational and attributional interpretations. Both interpretations were shown together for each figurative statement, but participants rated each

separately. Thus they were free to assign high or low ratings to both the relational and attributional interpretations if they chose.

Method

Thirty-two Northwestern University undergraduates were presented with 28 statements: 20 figurative statements taken from Experiment 1, and 8 fillers. The statements were blocked so that each participant saw either all statements in simile form or all statements in metaphor form. Two random orders were used. After each statement, a relational and attributional interpretations of the statement appeared. The order of the interpretations on the page was counterbalanced. The participants were asked to rate how much they agree with each of the interpretations on a 1 to 7 scale.

Results

We computed the scores for the relational and attributional interpretations for each item. Table 2 shows mean ratings for each of the four item categories, along with the number of relational and attributional interpretations that received ratings of 4 or greater (out of 7). Consistent with prior research, relational interpretations are preferred over attributional interpretations overall (Gentner, 1988; Gentner & Clement, 1988).

The key question for our purposes is whether the materials were skewed such that conventional metaphors had more or better relational interpretations than the other categories. This does not appear to be the case. Relational interpretations received high ratings (4 or above) for 8 out of 10 items in each of the four item categories -- conventional metaphors, novel metaphors, conventional similes, and novel similes. (Attributional interpretations were rated lower overall, as shown in Table 2.) It appears that the intended relational interpretations were highly apt for both metaphor and simile forms. These data offer some reassurance that the shift towards relationality in metaphor preference was not simply determined by disproportionate availability of relational interpretations for metaphors over similes.

We also created a relational preference score (Rpref), which was the difference between the relational rating and the attributional rating. An analysis of variance with base conventionality (novel or conventional) and grammatical form (simile or metaphor) as between-subjects factors revealed no significant differences in relational preference scores ($F_{3,\ 636} = 1.25$, $MSE = 13.07$, $p < 0.3$).

Table 2. Mean interpretation ratings and number of interpretations that received high ratings (in parentheses)

	Attributional	Relational
Conventional		
Metaphor	3.81 (5)	4.95 (8)
Simile	4.35 (6)	4.74 (8)
Novel		
Metaphor	3.68 (4)	4.51 (8)
Simile	3.81 (4)	4.78 (8)

Discussion

As predicted by the Career of Metaphor hypothesis, participants in Experiment 1 were likely to choose the metaphor form for figurative statements with conventional bases, and the simile form for those with novel bases. Aisenman's Relational Precedence hypothesis also received support: the metaphor form was chosen more often for relational meanings (i.e., following a relational context) than for attributional meanings (following an attributional context).

Might both claims be true? Some intriguing possibilities arise if we consider the implications of these two patterns taken together. Suppose that, as in the Career of Metaphor hypothesis, nominal figurative expressions are initially phrased as similes. As these expressions become conventionalized, the metaphor form becomes more felicitous. Suppose further that relational meanings of novel bases have more potential to get conventionalized. Then we would find a preponderance of relational meanings among conventional bases. An informal survey of the literature using conventional metaphors suggests that most of them do convey relational meanings. For example, the metaphors used by Ortony (1979) and by Glucksberg and Keysar (1990) are primarily relational (e.g., *Cigarettes are time bombs*; *Some jobs are jails*; *Sermons are sleeping pills*). Assuming that these stimuli are roughly typical of conventional metaphors, we might speculate that there is a preponderance of relational figuratives within the class of conventionalized metaphors. How might such a link between relationality and conventionality have come about?

One possibility is that different forms are used for conventionalized relational and attributional figurative statements. English has a special form for conventional bases that is often used for property attribution – "*as X as Y*," where X is the shared attribute, and Y is the base term – for example, *as white as snow*; *as strong as an ox* (Ortony, 1979). Perhaps conventional attributional meanings are siphoned off by this dedicated form.

However, relational adjectives can enter the *as X as Y* frame as well (e.g., *as delicious as an apple; as fierce as a tiger)*. The only requirement for the descriptor X seems to be that it be orderable on some dimension. Thus a possible special form for attributive figuratives does not seem like a viable explanation for the preponderance of conventional relational metaphors.

Another possibility is preemption by existing terms. Over the course of development, languages have developed names for attributes, which preempt the creation of new ones (Clark, 1992). On this account, creation of attributional metaphoric meanings might be less likely simply because we already have names for attributes. However, this explanation carries the hidden assumption that the number of attributes we want to express is smaller than the number of relations.

This brings us to the third and most speculative possibility. There is evidence that (1) people find shared relational structure more interesting or important than shared attributes; and that (2) relational meanings are relatively slow to emerge in cognitive development (Gentner & Rattermann, 1991; Halford, 1993) and arguably in the history of science. Applying this to the evolution of metaphor suggests that new relational abstractions are more likely to become entrenched than attribute meanings. Coherent relational systems are likely to be preserved in comparison processing, and this may carry over into the conventionalization of meanings and the formation of new categories (Gentner & Bowdle, 2001; Ramscar & Pain, 1996; Shen, 1992). On this account, a simile that expresses shared relational structure is more likely to give rise to parallels than one that expresses an attributional likeness. This would lead to differential likelihoods of conventionalization for relational and attributional figuratives.

Some evidence for this account can be obtained from studies of word meaning extension over time. One of the ways one can extend the meaning of a word is by analogy. For example, words like *bridge* and *sanctuary* initially had only concrete meanings, but now can denote metaphoric categories such as *something connecting two points* and *a safe place*, respectively. Table 3 shows the timeline of the first occurrences of the literal and figurative meanings of *sanctuary*, as listed in the Oxford English Dictionary, as well as other sample occurrences. (All senses are written exactly as in the OED.)

For *sanctuary*, the literal meaning of a holy building appears in 1340. Extensions to the church or the body of believers also appear in the 14th century. The first figurative usage appears two centuries later, in 1568. Interestingly, the first figurative use is signaled by an explicit comparison phrase "counted as a sanctuary". The first 'metaphorical' occurrence, unmarked by a comparison phrase, occurs considerably later, in 1685.

Table 3. Timeline of occurrences of literal and figurative meanings for *sanctuary*.

[Initial literal meaning]
I. a holy place – a building or place set apart for the worship of God or of one or more divinities: applied, e.g., to a Christian church, the Jewish temple and the Mosaic tabernacle, a heathen temple or site of local worship, and the like; also *fig*. To the church or the body of believers

1340…*in that sanctuary oure lord sall be kynge…*
1382 *And thei shulen make to me a seyntuarye, and Y shal dwelle in the myddil of hem.*
1530. *Sanctuarie, a place hallowed and dedicate vnto god.*

II.a – a church or other sacred place in which, by the law of the medieval church, a fugitive from justice, or a debtor, was entitled to immunity from arrest. Hence, in a wider sense, applied to any place in which by law or established custom a similar immunity is secured to fugitives.

1374 *To whiche Iugement they nolden nat obeye but defendedyn hem by the sikernesse of holy howses, that is to seyn fledden in to sentuarye.*
1463-4 *Eny persone..that shall dwelle or inhabit within the Sayntwarie and Procyncte of the same Chapell.*

[First figurative meaning]
1568 *Vsing alwaise soch discrete moderation, as the scholehouse should be counted a sanctuarie against feare.*

[First unmarked figurative meaning]
1685 *My house is your Sanctuary, and here to offer you violence, wou'd prejudice myself.*
1770 *The reformation was preceded by the discovery of America, as if the Almighty graciously meant to open a sanctuary to the persecuted in future years…*

Table 4. Timeline of occurrences of literal and figurative meanings for *bridge.*

[Initial literal meaning]
I. A structure forming or carrying a road over a river, a ravine, etc., or affording passage between two points at a height above the ground.

c1000 *theos brycg*
1131 *Men weorth on adrencte and brigges to brokene.*
c1449 *The brigge of Londoun.*
1660 *This was so severe a bill upon the Women, that, if a bridge was made from Dover to Calais, the women would all leave this kingdom.*

[Figurative]
1225 *The beoth ouer thisse worldes see, uppen the brugge of heouene.*
1742 *Faith builds a bridge from this world to the next.*
1863 *The bridge for thought to pass from one particular to the other.*
1874 *Gestures… forming the bridge by which we may pass over into spoken language.*

The pattern for *bridge*, shown in Table 4, is similar. The first literal meaning of *bridge* as a structure affording passage between two points above the ground goes back to the 11th century. However, the figurative

uses are not listed until the middle of the 18th century, except for a single reference to the bridge of heaven (which may have been meant literally) in 1225.

These patterns suggest that, at least in some cases, the more abstract, figurative meanings appear later in written language. In both cases, these figurative meanings are relational in nature. Interestingly, at least for *sanctuary*, the derived category no longer seems metaphoric; it has become a literal sense.

We suggest that the Relational Precedence account and the Career of Metaphor account may both be operative in the evolution of metaphor, and that they interact. Beginning with a pool of novel figuratives, the Career of Metaphor hypothesis states that for some of these the base term is repeatedly used in parallel comparisons, so that a conventional abstraction becomes associated with the base. What we suggest is that figurative expressions that yield coherent relational systems are most likely to be found novel and useful. Their bases are thus most likely to be reused and thereby conventionalized. For example, the simile *The cloud is like a marshmallow* elicits common attributes of the target and base, such as fluffy and white. But the potential abstraction 'white and fluffy' is unlikely to become a conventionalized word sense, both because of lexical preemption (we already have words for white and fluffy) and because the category it suggests is simply not very interesting. (Indeed, the conventional use of marshmallow as a metaphor is relational, as in *That boxer turned out to be a marshmallow.)*

Metaphors are a source of polysemy in language – they allow words with specific meanings to take on additional, related meanings (e.g., Glucksberg & Keysar, 1990; Lakoff, 1987; Lehrer, 1990; Miller, 1979; Murphy, 1996). We suggest that mappings that focus on relational structures are more likely to generate stable abstractions than mappings that focus on object attributes. In sum, conventionalization of relational meanings may fulfill an important cognitive function in creating new abstractions.

References

Aisenman, R. A. (1999). Structure-mapping and the simile-metaphor preference. *Metaphor and Symbol, 14*, 45-51.

Bowdle, B. F. (1998). *Conventionality, polysemy, and metaphor comprehension*. Doctoral dissertation, Department of Psychology, Northwestern University.

Bowdle, B. F., & Gentner, D. (1999). *Metaphor comprehension: From comparison to categorization.* In Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society (pp. 90-95), Vancouver, B.C.

Clark, E. V. (1992). Conventionality and contrast: Pragmatic principles with lexical consequences. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields and contrasts*. Hillsdale, NJ: Erlbaum.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59,* 47-59.

Gentner, D., & Bowdle, B. F. (2001). Convention, form, and figurative language processing. *Metaphor and Symbol, 16*, 223-247.

Gentner, D., & Clement, C. (1988). Evidence for relational selectivity in the interpretation of analogy and metaphor. In G. H. Bower (Ed.), *The psychology of learning and motivation, advances in research and theory*. New York: Academic Press.

Gentner, D., & Markman, A. B. (1997). Structure-mapping in analogy and similarity. *American Psychologist, 52,* 45-56.

Gentner, D., & Rattermann, M. J. (1991a). Language and the career of similarity. In S. A. Gelman & J. P. Brynes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225-277). London: Cambridge University Press.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language, 37,* 331-355.

Gibbs, H., & Wales, R. (1990). Metaphor or simile: Psychological determinants of the differential use of each sentence form. *Metaphor and Symbolic Activity, 5*, 199-213.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review, 97,* 3-18.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Erlbaum.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago.

Lehrer, A. (1990). Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics, 1,* 207-246.

Miller, G. A. (1979). Images and models, similes and metaphors. In A. Ortony (Ed.), *Metaphor and thought (1st ed.).* Cambridge: Cambridge University.

Murphy, G. L. (1996). On metaphoric representation. *Cognition, 60*, 173-204.

Ortony, A. (1979). Beyond literal similarity. *Psychological Review, 86,* 161-180.

Oxford English Dictionary, online edition http://www.oed.com/.

Ramscar, M., & Pain, H. (1996). *Can a real distinction be made between cognitive theories of analogy and categorization?* In Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Shen, Y. (1992). Metaphors and categories. *Poetics Today, 13*, 771-794.

Is Competitive Learning an Adequate Account of Free Classification?

Jan Zwickel (jzwickel@ix.urz.uni-hd.de)
Department of Psychology, Hauptstraße 47-51
69117 Heidelberg, Germany

A.J. Wills (a.j.wills@ex.ac.uk)
School of Psychology, University of Exeter
Perry Road, Exeter. EX4 4QG. England.

Abstract

Rumelhart & Zipser's (1986) competitive learning algorithm is an account of unsupervised learning and, as such, might be considered a potential model of free classification behavior in humans. However, selective learning effects (e.g. Dickinson, Shanks & Evenden, 1984) suggest that human learning, at least under conditions of feedback, may be better characterized by an error-correcting system. An experiment is reported that provides preliminary evidence for the existence of a selective learning effect in free classification. Simulations indicate that Rumelhart & Zipser's algorithm does not provide an adequate account of the behavior observed, whilst an error-correcting variant of competitive learning does.

Introduction

Free classification, or free sorting as it is also called, is a procedure in which human participants are presented with a set of stimuli and are asked to group them in any way that seems sensible or reasonable to them (e.g. Bersted, Brown & Evans, 1969; Regehr & Brooks, 1995; Wills & McLaren, 1998). It may be contrasted with the more standard experimental task of category learning via trial-specific feedback that has been the dominant mode of enquiry into humans' categorization abilities for the last fifty years (e.g. Bruner, Goodnow & Austin, 1956; Medin & Schaffer, 1978; Wills, Reimers, Stewart, Suret & McLaren, 2000).

The study of categorization under conditions where each decision receives immediate feedback from a totally reliable source has allowed psychologists great control over the structure of the categories participants acquire. As a methodology, it has been successful in broadening our understanding of the category learning process. However, the level of feedback available in such tasks seems higher than that available in many real-world situations, begging the question of whether what we have learned about the categorization process will generalize to situations where the feedback is absent or scarce.

An interesting parallel may be drawn with the sort of connectionist systems that have been proposed for learning in the presence or absence of feedback. For example, Rumelhart & Zipser's (1986) competitive learning model is an unsupervised system. It extracts statistical regularities in the input to form categorical representations, and does so in the absence of feedback. In contrast, McClelland & Rumelhart's (1985) model is a supervised system. It can be taught multiple categories (cat vs. dog vs. bagel in their example) but learns to categorize because each stimulus is accompanied by an externally-provided category label.

One of the differences between these two models is the nature of the weight-update algorithms they employ. McClelland & Rumelhart (1985) employ an error-correcting algorithm, where the size of the weight change is proportional to the mismatch between an external teaching signal and internal inputs. In other words, learning only occurs when the system fails to fully predict the teaching signal. Specifically,

$$\Delta w_{ij} = \eta(e_i - i_i)a_j \qquad 1$$

where Δw_{ij} is the change in the strength of the connection from unit j to unit i, e_i is the external teaching signal to unit i, a_j is the activity of unit j, and i_i is the total internal input to unit i, this being calculated as

$$i_i = \sum_j a_j w_{ij} \qquad 2$$

In contrast, Rumelhart & Zipser's algorithm does not employ error-correction in this sense. Rumelhart & Zipser use the internal input to determine which unit is the "winner" and then change weights to the winning unit by an amount proportional to the difference between the current weight of that connection and an asymptote[1]. Specifically, the change in weight from unit j to the winning unit is

[1] Rumelhart & Zipser (1986) also discuss a variant where connections to the losing unit are also changed via Equation 3, but with a much lower learning rate. The current article concentrates on the "winner-only" version, although the conclusions drawn are valid for both variants.

$$\Delta w_j = \eta\left(\frac{a_j}{n} - w_j\right) \qquad 3$$

where n is the number of active input units, and the winning unit is the one with the highest internal input. It is assumed in the current paper that active input units have an activity of 1 and inactive input units have an activity of zero.

Error-correction is assumed by some investigators to be a fundamental aspect of human learning in the presence of feedback, as evidenced by the phenomenon of selective learning (see below). If human learning in a free classification task fails to show evidence of selective learning, concerns would arise as to the generality of an empirical research program heavily based on learning with feedback. On the other hand, if selective learning is found to occur in free classification, the sort of unsupervised system proposed by Rumelhart & Zipser may not be an appropriate model for free classification behavior.

Selective learning

Probably the best-known example of selective learning is Kamin's (1969) "blocking" effect. Kamin's study involved rats but, as will be discussed later, there is now abundant evidence that corresponding effects can also be found in human learning (with feedback).

Kamin taught hungry rats that pressing a lever would result in food. Following this, pressing the same lever whilst a noise was present resulted in a mild electric shock. Unsurprisingly, rats learned to not press the lever whilst the noise was present.

Next, the auditory tone was accompanied by a light and pressing the lever whilst this tone-light compound was present also resulted in mild shock. The rats learned not to press the lever whilst the tone-light compound was present.

Group	Stage One	Stage Two	Test
Expt.	N→Shock	LN→Shock	L
Ctrl.		LN→Shock	L

Figure 1: Kamin's (1969) blocking experiment. "N" is an auditory stimulus and "L" is a visual stimulus.

In the test phase, just the light was presented, and the rats' behavior observed. The rats in the experimental condition pressed the lever quite a lot, whilst control rats (which had participated in stage two but not stage one) pressed the lever very little indeed. The design of this experiment is summarized in Figure 1.

The rats in the experimental group appear not to have learned the relationship between light and shock even though the control rats, which received an equal amount of training with the light-noise compound, have learned the relationship. Why might this be?

A number of animal learning theorists (e.g. Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972) have essentially argued that it happens because learning is driven by surprise. For the rats in the experimental group, the shock is not a particularly surprising event because it is predicted by the noise. The rats therefore don't bother to learn about the light in stage two. Similar effects have been demonstrated with undergraduates using computer-based tasks, including simple computer games (e.g. Dickinson, Shanks & Evenden, 1984), stock market simulations (e.g. Chapman & Robbins, 1990) and simulated medical diagnosis tasks (e.g. Gluck & Bower, 1988).

As Rescorla and Wagner (1972) noted, the notion of surprise-driven learning is well-captured by an error-correcting learning rule such as the one give in Equations 1 and 2. In fact, the learning theory they proposed is basically a variant of this learning rule.

An error-correcting system reproduces the blocking effect in the following way. For simplicity, consider that there are three units - the noise unit, the light unit and the shock unit. The external input produced by the presence of a stimulus is 1, and the external input produced by its absence is zero.

Initially, shock is not expected, so the link between noise and shock, and light and shock, are small. During stage one of the experiment, the strength of the link between noise and shock increases and eventually reaches 1. In the second stage, noise, light and shock are all present. However, the internal input to the shock unit is already 1 because of the strength of the link between the noise and the shock. Therefore, no weight change can occur via Equation 1. The light does not become associated with shock, even though it clearly would in the control condition. Under the non-error-correcting algorithm given in Equation 3, the light→shock association would reach an equivalent level in both conditions.

Experiment

It is reasonably clear from previous research that humans and other animals engage in selective learning under conditions of trial-specific, informative feedback. Do humans also display selective learning in a task without such feedback? The experiment reported in this paper represents a first attempt to address this question.

In our experiment, participants had to make up their own categories, although they were constrained by the fact they were only allowed two groups. Previous research demonstrates that category learning can proceed successfully in the absence of feedback (e.g. Homa & Cultice, 1984; Wills & McLaren, 1998).

Our participants received intermittent, non-trial-specific, feedback about their overall level of performance following every 24 stimuli, in order to maintain motivation and encourage adoption of the experimenter-defined categories. We believe that such a procedure is still properly described as "free classification" as no single response can be considered correct or incorrect. Situations where all forms of feedback are entirely absent are probably almost as rare outside the laboratory as situations where feedback is always immediate and trial-specific.

Abstract, novel stimuli were employed in this experiment as we wished to study category learning with adult participants - with such participants the category learning process is probably complete or far-advanced for most realistic stimuli.

Stimulus presentation was brief and followed by a mid-gray mask. The time available for a decision was also very limited. Both of these procedures were employed to encourage participants to rely on relatively non-analytic, similarity-based categorization processes, rather than analytic, rule-based processes.

The basic design of the experiment is shown in Figure 2. The letters A to J each represent sets of features present in the stimuli shown to participants.

In the first phase of the experiment, participants were presented with examples from two different categories. Examples from category 1 were created from a base pattern that contained feature sets G and H. Examples from category 2 were created from a base pattern that contained feature sets I and J. Note that the labels "category 1" and "category 2" are essentially arbitrary in a free classification task - they could be reversed without changing anything in the design or execution of the experiment.

As Figure 2 illustrates, once the participant had mastered the GH vs. IJ categorization they were transferred to a second categorization. Participants proceeded through all five categorizations in this way, at which point the experiment was over.

The datum of central importance in this design is the category to which the first stimulus presented in phase five is allocated. The first stimulus is chosen because subsequent decisions in phase five may be contaminated by learning on previous phase five trials.

Phase	**1**	**2**	**3**	**4**	**5**
Cat. 1	GH	GE	AB	AE	CE
Cat. 2	IJ	IF	CD	CF	AF

Figure 2: Design of the experiment. Letters represent sets of features, hence category 2 in phase 3 contains feature sets C and D.

The Rumelhart & Zipser system provides the null hypothesis for this experiment because it predicts that either key is equally likely to be used. It is perhaps not immediately apparent why this should be. To elucidate, one first needs to note that in each of the first four phases, all features are equally predictive of category membership. This means that, for a system such as Rumelhart & Zipser, in each phase learning should end with two features (A and E in phase four) being equally associated to one category representation, and two features (C and F in phase four) being equally associated to the other category representation. Hence, the first stimulus presented in phase five will activate both category representations equally and so the choice of which category to place it into must be arbitrary. This conclusion is confirmed by simulation in a later section.

Why might one expect anything other than a null result with this design? One possible reason would be if people exhibited selective learning in free classification. Note that, across phases 1 to 4, E and F only occur in situations where the information they provide is partially redundant. In phase 2 the stimuli can be categorized on the basis of whether they contain G or I features, a categorization already learned in phase one. In phase four, the stimuli can be categorized on the basis of whether they contain A or C features, a categorization already learned in phase three. Hence, through analogy to selective learning effects in tasks with feedback, one might consider that E and F develop little control over responding.

Method

Due to space limitations, we are unable to report the pilot studies performed. Reports may be found in McCooe(2000) and Zwickel (2001).

Participants and apparatus

Sixteen first-year Psychology students from the University of Exeter participated to fulfil a course requirement. Participants were tested in groups in a quiet computer room. Stimulus presentation was on 17" color monitors connected to Tiny Pentium III PCs running the DMDX software package (Forster & Forster, 2000, version 2). Responses were collected via the left and right CTRL keys on standard PC keyboards. Participants sat approximately 50cm from the screen.

Stimuli

Each stimulus was made up of 12 small pictures (hereafter "elements") taken from a set of 72 that have been used in a number of previous experiments (see Jones, Wills & McLaren, 1998 for the full set). For any given stimulus, the 12 elements were randomly arranged in a square of 3 rows with 4 icons in each row, and were surrounded by a gray rectangle outline 5.5cm in height and 4cm in width. Figure 3 shows an example

stimulus. Throughout all five phases, no stimulus contained more than one copy of any given element.

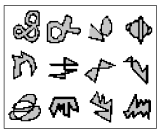


Figure 3: An example stimulus

Each of the letters A to J in Figure 2 represent a set of six elements. The assignment of elements to letters was randomly determined for each of 8 pairs of participants, with the remaining 12 elements (72 elements - 10 letter sets x 6 elements per set) being used for practice trials.

In order to control for possible effects of differential salience of the elements, one participant in each pair received the stimuli described in Figure 2 whilst the other received a design where E was transposed with A, and F was transposed with C. Hence, the putatively redundant elements were E and F for one member of the participant pair, whilst they were A and C for the other member. This means that any preference revealed in phase five cannot be due to A and C elements being more salient overall than E and F elements. To aid clarity, all participant data is reported as if E and F were the putatively redundant elements.

The stimuli actually presented to participants were generated by random distortion of the base patterns described in Figure 2. Each element in a base pattern was given a 10% chance of being replaced by a randomly selected element from the other base pattern (no element occurred more than once in any given stimulus).

An example may be helpful. To create an AF stimulus in phase five, the six A elements and the six F elements were randomly arranged in the four-by-three grid of the stimulus. Each element was then given a 10% chance of being replaced by a randomly selected element from set C or E. This method of stimulus construction produces training examples which are composed predominately of elements characteristic of a particular category but which also exhibit considerable variability.

Procedure

The five phases described in Figure 2 were preceded by some general written instructions and a brief practice phase to familiarize participants with the procedure. The experiment then proceeded in blocks of 24 trials.

On each trial, a stimulus was presented for 800ms and followed by a mid-gray mask that was presented for 1200ms. If a response was not detected within 2000ms of stimulus onset, the trial terminated with the message "You responded too slowly, please speed up!" and the participant was moved on to the next trial.

Each block comprised the sequential presentation of 24 stimuli, 12 from each of the two categories. At the end of each block a short message appeared stating the percentage of correct responses made by the participant in that block, and that they needed to score more than 80% to proceed to the next part of the experiment.

Clearly, percent correct has a slightly different interpretation in a free classification task to a task with trial-specific feedback as the relationship between categories 1 and 2 and the two response keys is arbitrary. Hence, percent correct was computed under the assumption that category 1 would receive a particular response, and the resulting number was subtracted from 100 if it was less than 50.

When a participant's score exceeded 80% they were moved on to the next phase of the experiment, after having been presented with the message "You did very well! You are now entering the next phase". If participants completed 10 blocks without ever reaching the 80% criterion they were moved on to the next phase with the message "You are entering the next phase as you have been in the last block of this phase".

Results

Consider Figure 2 again. The central null hypothesis we are attempting to reject is that, in the first trial of phase five, a participant will be no more likely to categorize AF using the response typically made to AE in phase four than the response typically made to CF in phase four. Similarly, they will be no more likely to categorize CE using their typical AE response than their typical CF response.

Of the 16 participants tested, 12 used the same response key for CE that they had typically used for CF (or the AE response key for an AF stimulus). Three participants showed the opposite response, using the CF key for an AF stimulus or the AE key for a CE stimulus. The remaining participant could not be described as having a preference for any in key in response to AE or CF as they scored exactly 50% across phase four. Treating this participant in the manner that makes it hardest to reject the null hypothesis, we can state that at least 12 participants emitted the CF→CE (or AE→AF) response, whilst no more than 4 participants emitted the opposite response. Given the null hypothesis would predict 8 responses of each type, the probability of the null hypothesis being correct is smaller than 0.05, $\chi^2(1) = 4.0$. The effect is also significant with an exact binomial test.

Participants completed a mean of 5.88 blocks in phase one, 4.13 blocks in phase two, 6.12 blocks in phase three, 5.75 blocks in phase four, and 5.56 blocks in phase five. The number of participants failing to achieve more than 80% correct in the five phases were 6, 3, 7, 7 and 7 respectively.

Discussion

The results of the current experiment appear to be problematic for those that would attempt to explain free classification behavior in terms of the competitive learning algorithm of Rumelhart & Zipser (1986). The model predicts no preference for which of the two categories developed in phase four are used to categorize the first stimulus in phase five, yet a clear preference was observed. The direction of the preference is that predicted if one assumes the presence of selective learning in free classification

One possible defense of the Rumelhart & Zipser algorithm is that its predictions were derived for a situation where learning in each phase is essentially complete before the next phase begins. Given the relatively high numbers of participants failing to reach criterion, it might reasonably be argued that asymptotic predictions are not appropriate. Does this make a difference? This is one of the questions addressed in the following section.

Modeling

We employed simulation techniques to more thoroughly investigate whether Rumelhart & Zipser's (1986) competitive learning algorithm could accurately reproduce the categorization preference observed in our experiment. To this end we set up a network with 72 input units (one for each element) and 2 output units (one for each category). Each input unit had a forward connection to each output unit, and the connection weights were initialized to small, random values.

One network simulation was performed for each participant in the experiment, with weights being initialized for each participant. The nature of the stimuli presented to a simulated participant, and the order in which they were presented, were determined by the specific stimuli presented to a corresponding human participant. After the presentation of each stimulus, the winning category node was determined in the same manner as Rumelhart & Zipser (1986). In other words, it was determined by calculating the total internal input to each unit, and selecting the unit with the larger total. The weights from each of the input units to the winning category unit were then updated in accordance with Equation 3. The weights of the losing unit remained unchanged.

The value of η (the learning rate) employed by Rumelhart & Zipser was 0.05. At this value, no preference in the categorization response to the first stimulus in phase five was found. Six simulated participants made CF→CE or AF→AE responses whilst six made the opposite response. The nature of the response made by four simulated participants could not be determined because in phase four they employed both category nodes equally for both stimulus types. Hence, unlike the human participants, the networks did not display a categorization preference in phase five.

The Rumelhart & Zipser (1986) algorithm was applied to our data with a wide range of learning rates (0.001 to 0.009 in steps of 0.001, 0.01 to 0.09 in steps of 0.01, and 0.1 to 0.9 in steps of 0.1). In no case did the algorithm display a categorization preference in phase 5.

An error-correcting competitive algorithm

We also attempted to simulate our result using an algorithm that combined the error-correcting principle of Equations 1 and 2 with the basic properties of the competitive learning algorithm of Equation 3. On any trial, the winning unit was determined in the same manner as the Rumelhart & Zipser model. The weight-update algorithm employed on each connection from an input unit *j* to the winning unit was

$$\Delta w_j = \eta \frac{(1-i)}{n} \qquad 4$$

for connections from active input units and

$$\Delta w_j = -\eta(1-i) \times \frac{n}{m} \qquad 5$$

for connections from inactive input units. In these equations, η is the learning rate, *i* is the total internal input to the winning unit, *n* is the number of active input units and *m* is the number of inactive input units. The weights from input units to the losing unit are not changed. This chimera of an algorithm is not equivalent in behavior to either of its components but does preserve some of the properties of each.

Removing the weight update algorithm of Equation 3 from our previous simulation, and replacing it with the algorithm described in Equations 4 and 5, we find a dramatic change in behavior. Now, at a learning rate of 0.05, all 16 simulated participants make CF→CE or AF→AE responses. In other words, the simulation now reproduces the behavior observed in our human participants, although the overall level of learning is slightly higher in our simulation. A reliable preference is found for a wide range of learning rates - from 0.01 to about 0.4.

Conclusion

The experiment reported in this paper provides preliminary evidence that the ubiquity of selective learning effects in tasks with immediate, trial-specific feedback extends to some categorization tasks where feedback is scarce and not trial-specific. To the extent this phenomenon is found to be general to free

classification tasks, it casts some doubt on the adequacy of certain types of competitive learning algorithms as accounts of free classification behavior. In particular, an algorithm suggested by Rumelhart & Zipser (1986) was found to have difficulty in reproducing the results found. We suggest that a competitive algorithm which includes some aspect of error-correction may be a more appropriate account. One simple algorithm of this type was described, tested, and found to be able to reproduce our results.

The two main avenues of future research suggested by the results and simulations in this paper are a) investigation of the generality of selective learning effects in free classification, b) consideration of whether other unsupervised systems (e.g. Adaptive Resonance Theory, Grossberg, 1976) are capable of accounting for the results so far found.

References

Bersted, C. T., Brown, B. R., & Evans, S. H. (1969). Free sorting with stimuli in a multidimensional attribute space. *Perception & Psychophysics, 6B*, 409-413.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A Study of Thinking*. New York: Wiley.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, 18, 537-545.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *The Quarterly Journal of Experimental Psychology, 36A*, 29-50.

Forster, K. I., & Forster, J. C. (2000). DMDX version 2. Retrived from the internet on 12/2000 from http://www.u.arizona.edu/~jforster/dmdx.htm.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal Of Experimental Psychology: General*, 117(3), 227-247.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23*, 121-134.

Homa, D. & Cultice, J. (1984). Role of Feedback, Category Size, and Stimulus Distortion on the Acquisition and Utilization of Ill-Defined Categories. *Journal of Experimental Psychology: Learning Memory, and Cognition, 10*, 83-94.

Jones, F. W., Wills, A. J., & McLaren, I. P. L. (1998). Perceptual categorization: Connectionist modelling and decision rules. *The Quarterly Journal of Experimental Psychology*, 51B(3), 33-58.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell & R. Church (Eds.) *Punishment and aversive behavior*. New York: Appleton.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal Of Experimental Psychology: General, 114*(2), 159-188.

McCooe, M. (2000). *Learning without being taught* (unpublished project report). Cambridge: Dept. of Experimental Psychology, University of Cambridge.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82, 276-298.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207-238.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87, 532-552.

Regehr, G., & Brooks, L. R. (1995). Category organisation in free classification: The organising effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory & Cognition, 21*(2), 347-363.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.*), Classical conditioning II: Current research* (pp. 64 - 99). New York: Appleton-Century-Crofts.

Rumelhart, D. E. & Zipser, D. (1986). Feature Discovery by Competitive Learning. In D. E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.) *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations.* Cambridge, London: MIT Press.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *The Quarterly Journal of Experimental Psychology, 51B*(3), 235-270.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology, 53A*(4), 983-1011.

Zwickel, J. (2001). *Categorisation without feedback* (unpublished report). Exeter: School of Psychology, University of Exeter.

Acknowledgments

This research was supported by a research fund provided by the University of Exeter. The authors would like to thank Manfred Amelang. Meg Davies, Mike McCooe and Joachim Funke for their practical assistance, and Fergal Jones, Timm Lochmann, Ian McLaren and Katrin Scharpf for their insightful comments.

Member Abstracts

A Formal Analysis of Intelligent Agents with Mathematical Tools

Zippora Arzi-Gonczarowski (zippie@actcom.co.il)
Typographics, Ltd.; 46 Hehalutz Street
Jerusalem 96222, Israel

ISAAC ('Integrated Schema for Affective Artificial Cognition') is a mathematical model of intelligent agents, which gives rise to a formal theory that could be implemented computationally.

Similar to the natural evolutionary context, the schema starts from a simple model of corresponding sensations and reactions. It then structures 'upgrades' (e.g. handle conflicting reactions, internal representation, and so on) on top of that, using generative reasoning to systematically obtain and study the properties of these upgraded structures. Among other things, this approach yields a continuous bridge from low level to high level intelligence.

A *perception* snapshot is structured as a set of *world elements* that constitutes an environment (real or imagined), a set of *connotations* that constitutes a collection of discriminations, and a set of *behaviors*. Behaviors are conjured up on the basis of a *perception predicate* that relates between world elements and their connotations in a three valued manner (*true, false, undefined*). With real environments, that basic schema approximates models of simple forms of intelligence. For higher level forms of intelligence, mind activities (cognitive, behavioral, affective) are modeled as streams of perceptions. Along these streams, all the components mentioned above could adapt dynamically: be modified, extended, merged, and so on. In mathematical terminology, perceptions are *objects*, and passages from one perception to another are modeled as *morphisms*.

In the course of a few years of ongoing research, a variety of mind processes have been modeled on the basis of these uniform, yet flexible, premises, capturing mental activities from streams of interpretations, through behavior development and integration, representation formation, imaginative design and anticipation, analogy making, to social and self perception. Each publication has been appraised as a model of the relevant mind aspect, but as a collection they also feature an additional value of an integrated whole: because they share uniform modeling premises, the various processes can be neatly composed and alternated between, modeling multifaceted intelligences.

Mathematical treatments are naturally expected to come up with equations. The equational results of ISAAC are categorical commutative diagrams that state where one formalized mind process is a 'paraphrase' of another. These diagrams are like high level 'blue prints' that provide a computational schema with basic structural guidelines for neat agent architectures.

Mathematization has, indeed, proven a powerful modeling tool for other scientific domains. In cognitive science, however, the open ended diversity of phenomena that need to be modeled has made it difficult to capture things in a uniform manner. If one were to overcome that obstacle, then this would probably be by finding a suitable level of abstraction: high enough to absorb a variety of phenomena, but not too high so that meaningful things could still be stated. Mathematical categorization is a tool that has been developed precisely for such purposes within mathematics itself. ISAAC deploys well developed category theoretical tools for cognitive science purposes, yielding a general, yet rigorous, schema.

Pre-theoretically, ISAAC is grounded by intuitions. As a theory, the treatment proceeds as if the semantic primitives were context independent: All definitions, constructions and results are tidily operated within the abstract mathematical framework. Then, they are invariably examined with regard to the grounding intuitions, pre-theoretical conceptions, and existing knowledge about the modeled processes. Results that have not been anticipated at the outset provide supporting arguments that the proposal is apparently on a promising track.

Substitution instances of the schema could be applied to describe, or to design, particular agents, filling in detailed domain-specific features to approximate particular agents. The foundational schema avoids over determinism and captures abstracted structures and mechanisms that are shared by intelligent agents. This yields a unifying theory and, hopefully, contributes to:
(1) Cognitive science becoming a science, and to
(2) Integrative artificial intelligence that does not lose the big picture by over fragmentation.

References

The publications can be found on the author's web page: `www.actcom.co.il/typographics/zippie`

Distinct Errors Arising From a Single Misconception

Ryan S. Baker (rsbaker@cmu.edu)
Albert T. Corbett (corbett@cmu.edu)
Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Data and Introduction

We present an account of the mistakes students have been observed to make when generating scatterplots, and give evidence that somewhat disparate error behaviors can be traced to the same strategic decision.

In two prior studies (Baker, Corbett, & Koedinger 2001, 2002), we observed middle-school students attempting to generate scatterplots (which have a quantitative variable on each axis) but making two conceptually similar errors. When given both categorical and quantitative variables but no advice on which to place in their graph, 15% made what we call the *choice* error, incorrectly choosing a categorical variable for the X -- 0% used the correct variables. Naming the variables to use in the question did not eliminate this error, but 77% used the correct variables. 13% of those students, however, then made what we term the *representation* error: treating the values of the quantitative X variable as if they were categorical. They wrote the variable's values along the axis in the order they appeared in the data table, rather than numerical order, e.g., placing "22 20 23 25 24 19 23" along the axis rather than "19 20 21 22 23 24 25". Labeling the axis variables for the student did not significantly reduce the representation error's frequency.

Our conjecture was that students made both errors due to a substantially greater familiarity with bar graphs, leading to the belief that graphs always have a categorical x-axis. A possible alternate conjecture for the representation error is that students don't understand the difference between categorical and quantitative variables. However, this conflicts with their comparatively greater ability to represent the Y axis quantitatively, as is done in a bar graph.

Modeling

In drawing each axis of a scatterplot, there are two key decisions for the student to make -- which variable to graph and how to represent its values. We developed an ACT-R (Lebiere & Anderson, 1998) model of these decisions and fit it to the students' behavior displayed in Table 1. In alternate fits we modeled the two correct decisions (correct variable choice and quantitative value representation) as two different productions (if-then rules) or the same production. Similarly, we modeled the two errors (categorical variable choice and categorical representation of a quantitative variable) as two productions or the same production. Modeling the two correct decisions as different productions produces a significantly better fit than modeling them as a single production ($F(1,30)=39.853$, $p<0.001$). The model where the two errors stem from the same strategic production has equal fit but superior parsimony to the model where they stem from different strategic productions. (BiC (same)=77.00, BiC(different)=79.28) The former model achieves an excellent fit to the overall pattern of data from the two experiments. ($r=0.990$, mean absolute dev =.057)

This model is generally consistent with but clarifies our early conjectures. As shown in the top row of the table, students never pick a quantitative variable for the x-axis unless one is suggested. This implies that in these studies correct selection of a quantitative variable just reflects the ability to follow directions; treating a quantitative variable as quantitative, on the other hand, is modeled as active knowledge of the difference between quantitative and categorical variables. The two errors in this account, selecting a categorical variable and treating a quantitative variable as categorical, reflect a single misconception. These students know the difference between categorical and quantitative variables, but are biased to make the X axis categorical in any way possible, consistent with their prior experience with bar graphs.

Thus, in this domain multiple error behaviors arise from a single misconception, an overgeneralization of their prior knowledge of bar graphs.

Table 1: Percent occurrence of behaviors

	No prompts	No labels	X label	Y label	Both label
Correct	0	53	59	62	61
Choice error	15	27	9	27	8
Rep error , X axis only	0	10	14	12	10
Rep error, Y axis only	0	0	0	0	0
Rep error, both axes	0	3	4	0	6
Other/ Give Up	85	7	14	0	15

References

Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

Baker R.S., Corbett A.T., & Koedinger K.R. (2001) Toward a Model of Learning Data Representations. Proceedings of the Cognitive Science Society Conference. pp. 45-50

Baker R.S., Corbett A.T., & Koedinger K.R. (2002) The Resilience of Overgeneralization of Knowledge about Data Representations. Presented at American Educational Research Association Conference. New Orleans, LA.

Belief in the Hot Hand Improves Performance: A Mathematical Model

Bruce D. Burns (burnsbr@msu.edu)
Department of Psychology, Michigan State University
East Lansing, MI 48824-1117

The widely held belief in the "hot hand" in basketball suggests that a player experiencing a streak should be given the next shot. However Gilovich, Vallone & Tversky (1985) found that streaks of hits in basketball shooting were no more likely than chance, so basketball shots are independent events. Thus it has been widely accepted that belief in the hot hand is a fallacy. Starting with the question of what are the goals of basketball players, Burns (2001) argued that the data only demonstrated that the hot hand is invalid as an *individual* cue to *when* a player will hit a shot, not that it is an invalid *allocation* cue for deciding *who* to give the next shot to. Streaks should occur more often for good shooters.

Burns (2001) used computer simulations to show that giving the next shot to players who hit their last shot improved a basketball team's scoring. However these simulations had two weaknesses. First, occasionally the simulations utilizing the hot hand did not outperform those not using it. Although less than 1% of the simulations, they raised questions about the claim that belief in the hot hand would always be expected to help. Second, in order to limit the number of free parameters and make the entire parameter space explorable, some simplifying assumptions had to be made. Although it was possible to argue that these assumptions were not critical, it would be better to demonstrate this directly.

A Markov Model

To address concerns about the simulations I constructed a Markov model of the first two shots in basketball. The first two shots were modeled because at least one shot is necessary before there can be a hot hand, and if scoring is improved in the first two shots it should be improved over any number of shots. My analysis does not however assume that the hot hand is defined by just one hit. The model could be applied to any definition of the hot hand in which it represents a temporary elevation in the probability of giving a player the next shot that is triggered by recent success.

The model has four parameters and represents Player X and Player (or Players) Y. A bias parameter b represents the probability of giving the next shot to Player X, whereas the same probability for Player Y is $1\text{-}b$. The bias parameter represents any bias to give the ball to a player (e.g., high shooting percentage, perceived ability, friendship, etc) that is independent of recent success. The model does not incorporate a parameter for belief in a "cold" hand because there is no empirical evidence for this belief.

The model has separate parameters for the shooting percentages for the two players, s_x and s_y for Players X and Y respectively. The hot hand parameter h temporarily elevates the probability of a player being given the next shot after a hit. Thus the probability of Player X being given the next shot after a hit is $b + h(1 - b)$. All parameters have a range of 0.0 to 1.0. The expected number of hits after two shots, calculated by summing the expected outcomes of all 16 possible states is:

$$E(\text{hits after two shots}) = 2(b(s_x - s_y) + s_y) + h(b - b^2)(s_x - s_y)^2$$

The $h(b - b^2)(s_x - s_y)^2$ component is *never* negative, thus belief in the hot hand can never lower the expected number of hits. Any positive value of h will raise the expected outcome so belief in the hot hand increases expected scoring, just as was shown in the simulations. However there are the same two exceptions to this: h will have no effect when $s_x = s_y$ (if there is no difference between players then it does not matter how shots are allocated), and when $b=1$ or $b=0$. This pure strategy of always giving the ball to one player is neither observed (even when it is optimal) nor would be desirable in NBA basketball. Game theory predicts that for most interesting competitive games there is a mixed strategy equilibrium.

Conclusions

This Markov model provides a mathematical proof that belief in the hot hand is beneficial if shots are independent events. In this way it expands on Burns' (2001) simulations. It also makes clear that giving the ball to good shooters and to players experiencing the hot hand are not mutually exclusive strategies. Instead allocating shots between players is a multi-cue decision making task in which both players' base-rates of success over the long term and short-term streaks are valid allocations cues. Giving weight to both of these allocation cues will improve the amount that a team will be expected to score. False beliefs that shots are dependent may be a way to maintain utilization of streaks.

References

Burns, B. D. (2001). The hot hand in basketball: Fallacy or adaptive thinking? In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Meeting of the Cognitive Science Society* (pp. 152-157). Hillsdale, NJ: Lawrence Erlbaum.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Human Reasoning:
An Analysis of the Mathematical Problem-resolution Strategies

Manoel CAETANO
LCC/CCH/UENF
Av. Alberto Lamego 2000
28015-620 Campos RJ Brazil
T: 55 22 27261589 / F: 55 22 27261589
manoelcaetano@hotmail.com

Adriana SOARES
UNIVERSIDADE GAMA FILHO
Mestrado em Psicologia
Rua Manuel Vitorino 625 Prédio CP
20748-900 Piedade Rio RJ Brasil
T: 55 21 22748407 F: 55 21 22748409
mespsi@ugf.br

In this paper we investigate the human reasoning applied to the mathematical problem-resolution process. Our approach is based on two main settings: *a.* the investigation of mental processes involved in the human reasoning applied to problem resolution; *b.* the analysis of differences in the categorization and resolution of mathematical problems by novicesrs and experts.

In *a.* we sought to contribute for the rupture of the logical-formal reasoning paradigm. In fact, we sought to contribute for the rupture of the idea that identifies the human as a completely rational entity, which invokes a thinking way that adheres the rules of an explicit form. Our results show that the human reasoning is not determined exclusively by logical-formal guidelines, but is rather determined by characteristics and pragmatics aspects of the context.

In *b.* we sought to analyze more effectively the problem-resolution process. We concentrated our discussion on the differences in the categorization and resolution of mathematical problems by novices and experts. Our results indicate for a problem categorization and subsequent resolution: experts are guided by organized logical principles, and novices are guided by superficial elements found in its enunciation.

The Role of Logical Structure and Premise Believability in Belief Revision

Dustin P. Calvillo (calvillo@psych.ucsb.edu)
Department of Psychology, University of California
Santa Barbara, CA 93106-9660 USA

Russell Revlin (revlin@psych.ucsb.edu)
Department of Psychology, University of California
Santa Barbara, CA 93106-9660 USA

Belief Revision

Belief revision occurs when one moves from one belief state to another after encountering some data that are inconsistent with one's initial belief set. Experiments in belief revision have demonstrated that the initial logical structure of an argument affects how reasoners revise their beliefs. When arguments for changing beliefs are made in a logical form, the typical finding is that the major premise is revised more frequently than the minor premise. This is evident when the modus ponens (MP) inference is contradicted (if p then q; p; therefore, q), while there is no clear preference when the modus tollens (MT) inference is contradicted (if p then q; not q; therefore, not p) (Dieussaert, Schaeken, De Neys, & d'Ydewalle, 2000; Elio & Pelletier, 1997; Politzer & Carles, 2001). Others have reported a different finding: reasoners revise belief in the major and minor premises equally often in MP problems, but prefer to disbelieve the minor premise in MT problems (Revlin & Calvillo, 2002; Revlin, Cate, & Rouss, 2001). In three experiments, we explore possible explanations for these two different patterns of results.

Three possible explanations for the inconsistent results are the types of major premises, the revision alternatives presented to participants, and the prior believability of the major premises. The major premises used by Elio and Pelletier (1997), Dieussaert et al. (2000), and Politzer and Carles (2001) were conditional (if p then q) and somewhat neutral in believability. Participants in these experiments were allowed to express uncertainty toward premises. The major premises used by Revlin et al. (2001) and Revlin and Calvillo (2002) were universal quantifiers (all p are q) and considerably more believable. Participants in these experiments were forced to decide, with certainty, to disbelieve the major or minor premise.

In Experiment 1, we assigned 80 introductory psychology students from the University of California, Santa Barbara into four groups. Logical structure (MP or MT) and type of major premise (conditional or quantifier) were between-participants variables. The major premise revision rates are presented in Table 1. The rates for both quantifiers and conditionals were similar to those found by Revlin et al. (2001). Logical structure had a significant effect, type of major premise did not, and the two variables did not interact. This ruled out the use of different major premise types as an explanation for the different previous findings.

In Experiment 2, we assigned 50 participants to two groups and presented them with MP and MT problems like in Experiment 1, but gave them the revision alternatives used by Politzer and Carles (2001). As seen in Table 1, logical structure had a reliable effect on revision rates and the major premise revision rates were similar to those of Revlin et al. (2001), ruling out the use of different revision alternatives as an explanation for the inconsistent results.

Table 1: Major premise revision rates by logical structure.

	MP	MT
Experiment 1: Conditional	0.329	0.044
Experiment 1: Quantifier	0.263	0.107
Experiment 2	0.465	0.123
Experiment 3: Low believability	0.751	0.701

In Experiments 1 and 2, the major premises used were highly believable. In Experiment 3, we gave 47 participants either MP or MT problems with major premises of low-believability. The results, as seen in Table 1, were similar to those of Elio and Pelletier (1997). There was a preference to revise belief in the major premise in both MP and MT problems and there was no effect of logical structure.

Experiments 1 and 2 ruled out the use of different types of major premises and revision alternatives explanations for the varying results in the literature. Experiment 3 showed that believability of the major premise is a likely source of the different patterns of results, demonstrating the need for models of belief revision to include initial premise believability to account for how reasoners revise beliefs.

References

Dieussaert, K., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 19,* 277-288.

Elio, R., & Pelletier, F.J. (1997). Belief change as prepositional update. *Cognitive Science, 21,* 419-460.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking and Reasoning, 7,* 217-234.

Revlin, R., & Calvillo, D.P. (2002). Stages in counterfactual reasoning. *Unpublished Manuscript.*

Revlin, R., Cate, C.L., & Rouss, T.S. (2001). Reasoning counterfactually: Combining and rending. *Memory & Cognition, 29*, 1196-1208.

Displacement Affects Duration Estimation, But Not the Other Way Around

Daniel J. Casasanto (djc@mit.edu)
Lera Boroditsky (lera@mit.edu)

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
77 Massachusetts Avenue NE20-423 Cambridge, MA 02139 USA

Introduction

How do we mentally represent and reason about domains for which little sensory information is available? One suggestion is that our understanding of abstract domains is intimately dependent on our understanding of richer, more experience-based domains (Boroditsky, 2000, 2001; Gentner et al., 2001; Gibbs, 1994; Holyoak & Thagard, 1995; Lakoff & Johnson, 1980, 1999). For example, people's understanding of time appears to be dependent on spatial knowledge (Boroditsky, 2000; Boroditsky & Ramscar, 2002). Previous evidence suggests that there is an asymmetric relationship between the domains of space and time. Whereas spatial knowledge was found to be useful for reasoning about time, temporal knowledge did not facilitate spatial reasoning (Boroditsky, 2000).

The present studies investigate whether this asymmetric relationship between space and time holds even for low-level representations of the two domains.

In a series of simple psychophysical tasks, participants viewed a moving stimulus and estimated either its displacement or its duration. Results show that temporal estimates were strongly modulated by the displacement of the moving stimulus, even when participants were encouraged to attend selectively to temporal information. In contrast, spatial estimates were not modulated by the duration of the moving stimulus when instructions encouraged selective attention to spatial information. Spatial estimates were only weakly correlated with stimulus duration when participants were required to attend to both temporal and spatial information simultaneously.

Experiment 1

Methods

Moving lines were presented on a CRT monitor. Line durations and displacements were varied parametrically. Durations ranged from 1 to 5 seconds in 0.5 second increments. Displacements ranged from 200 to 800 pixels, in 75 pixel increments. Nine durations were fully crossed with nine displacements to produce 81 distinct lines. Lines 'grew' horizontally across the screen one pixel at a time, from right to left, at rates ranging from 40 pixels/second to 800 pixels/second. Each line remained on the screen until its maximum displacement was reached.

Participants viewed 162 moving lines, one line at a time. Immediately after each line event, a prompt appeared indicating that the participant should reproduce either its duration or its displacement by clicking the mouse to indicate the beginning and end of the estimated temporal or spatial interval.

Results

Participants' temporal and spatial estimates were highly accurate. Target duration correlated positively with estimated duration (r^2=0.96), and target displacement correlated positively with estimated displacement (r^2=0.97). The effect of target displacement on estimated duration (r^2=0.72) was greater than the effect of target duration on estimated displacement (r^2=0.36).

Experiment 2

Methods

Materials and design were exactly as described in Experiment 1. The procedure was identical with the following exception: in Experiment 2, participants were notified before each trial whether they would need to reproduce the duration or displacement of the moving line.

Results

Again, participants' temporal and spatial estimates were highly accurate. Target duration correlated positively with estimated duration (r^2=0.96), and target displacement correlated positively with estimated displacement (r^2=0.99). The effect of target displacement on estimated duration (r^2=0.82) was much greater than the effect of target duration on estimated displacement (r^2=0.01).

References

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition, 75*(1), 1-28.

Boroditsky, L., & Ramscar, M. (2002). The Roles of Body and Mind in Abstract Thought. *Psychological Science, 13*(2), 185-189.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). The analogical mind: Perspectives from cognitive science.

Gibbs, R.J. (1994). The poetics of mind: Figurative thought, language, and understanding. New York, Cambridge University Press.

Holyoak, K.J. & Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MIT Press.

Lakoff, G. & Johnson, M. (1980). Metaphors we live by. University of Chicago Press, Chicago.

Evaluating Information Design for Notification Systems

C. M. Chewar (cchewar@cs.vt.edu)
D. Scott McCrickard (mccricks@cs.vt.edu)
Department of Computer Science, Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0106 USA

As computing platforms continuously grow in processing power, diminish in size, and are creatively integrated into every facet of the human experience, popular demand also increases for unfettered access to information of interest, necessitating insightful design for a variety of displays. While engaged in their daily discourse, occupied with activities such as driving, desktop computing, or interacting with others, people often want to remain notified about news items, collaborative efforts, and other changing information. Decision requirements within new settings or situations may prompt immediate interest in accessing related data. Notification systems in the form of ubiquitous computing devices, to include wearable computers, vehicle information systems, and handheld devices, are relied on support these information needs. Desktop computer users also depend on small-sized secondary display applications to provide similar notification information.

However, information conveyed through these devices and applications is often perceived with short, discrete attention shifts and glances rather than longer periods of full attention perception that has been considered typical of human-computer interaction. Certainly, this paradigm has implications for information design, rooted in cognitive processing and human attention limitations. Adding to this challenge, user goals are difficult to predict and often conflicting. For example, users may not want to be interrupted from a primary task, although they still wish to maintain awareness of information over a period of time or recognize specific information states. In other usage scenarios, users may wish to be alerted about information and attracted to some interaction. Platform capabilities may also mandate minimalist information representation, presenting an imperative for reevaluation of design guidelines for a wide array of emerging computer interfaces within these constraints.

Objectives and Related Work

Through empirical study, we seek to understand how various options for information encoding and design, presented within a dual-task situation, simultaneously affect user interruption while enabling reaction and comprehension of notifications. Although much work has been done to understand relative effectiveness and expressiveness of visual primitives within the human-computer interaction field, there are few empirically established design guidelines available for digital displays that are typically not a user's main attention focus. Cleveland and McGill's ordering of graph attributes provides guidance for primary task displays (1984), and Cleveland has extended consideration of graphical attribute effectiveness to specific information extraction tasks (1994).

However, the dual-task nature of notification systems usage requires evaluation of many other system variables for strong empirical study validity. For example, various combinations of mental and physical workload levels, cross-modal or intramodal presentation of the two tasks, and competing demand for sensory channels and short-term memory (Wickens & Hollands, 2000) will certainly have implications for fulfilling objective information design requirements. Empirical methods allowing reliable replication, measurement, and modeling of these variables are pivotal for creating notification systems guidelines.

Continuing Work

Initial findings from our work show that Cleveland and McGill's guidelines for use of visual attributes do not hold for dual-task situations where a distraction to a primary task requiring high attention and manual interaction must be minimized (Tessendorf et al., 2002). Additionally, we have seen evidence that information design for decision-support notification systems is best accomplished with cross-modal representations as a primary task's visual sensory demand level increases (tasks tested within a CAVE™ virtual environment and on a desktop computer). Continuing studies will lead to development of regression models and tables, supporting rule-based presentation adaptivity, complementary to efforts such as Horvitz's PRIORITIES system, which makes inferences about a user's attention state and calculates expected cost of an interruption to determine the most suitable presentation method (Horvitz, Jacobs & Hovel, 1999).

References

Cleveland, W. S. (1994). *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of American Statistical Association, 79*(387), 531-554.

Horvitz, E., Jacobs, A. & Hoxel, D. (1999). Attention-sensitive alerting. *15th Conf. on Uncertainty and AI (UAI '99)* (pp. 305-13). San Francisco, CA: Morgan Kaufmann.

Tessendorf, D., Chewar, C. M., Ndiwalana, A., Pryor, J., McCrickard, D. S. & North., C. (2002). An ordering of secondary display attributes. *Extended Abstracts of CHI2002* (pp. 600-1). New York: ACM Press.

Wickens, C. D. & Hollands, J. G. (2000). *Engineering Psychology and Human Performance*. 3rd edn. Upper Saddle River, NJ: Prentice Hall.

The Recognition of Overlapped Chinese Characters at Two Spatial Scales

Yu-Ju Chou (sallyc@cogsci.ed.ac.uk)
Richard Shillcock (rcs@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Introduction

The two hemispheres may have different strategies allowing various perceptions to be processed independently, such as near and far, ambient and focal, and low-level subcortical communications, contrasted to high-level cognition. Then what happen when the ambient processing of the large character, and the focal processing of small characters are merged? This study hypothesized that two repeated items should facilitate the recognition. Additionally, large stimulus captures visual attention, which should speed up the recognition of large characters, compared to small characters, and because this was a preattentive factor, the effect should not have a gender difference.

Experiment

Thirty-two Chinese characters, half high frequency and half low frequency, were presented in three versions: Large (400*400, units defined by Psyscope), Small (50*50) and Overlap version. The Overlap version was a large character overlapped by a small identical character in the centre. And the centre was emptied beforehand to make sure that the small character was not obscured by the large one. Characters and versions were arranged by Latin square design. A fixation point was presented in the centre of the monitor for 500 msec, followed by a character lasted for 150 msec, and then followed by a mask picture lasted for 500 msec. Native Chinese speakers, ten males and eleven females, participated in this study. They made a lexical decision by clicking the button on the response box with their index fingers, buttons were counterbalanced between people, and the response times were recorded by Psyscope.

Analysis and results

Frequency effect was significant by subjects and by items ($p < .01$), as High frequency characters were recognized faster than Low frequency characters. Size effect was significant only by subjects ($p < .01$) as Overlapped characters were recognized slower than Large and Small versions. Gender difference was insignificant ($p > .05$). A two-way interaction between Gender and Size was significant in the by-items analysis ($p < .05$). Post hoc tests show that females recognized Overlap characters quicker than Large and Small ones, but no significant difference for males.

Discussion

That large characters were recognized slightly quicker than small characters supports our hypothesis that the visual attention was more likely caught by larger objects, compared to small objects. And as presumed, Gender effect was not significant.

However, there were large characters in the overlapped versions, but the recognition of overlapped version was slowest. This suggests that in the overlapped version the small characters interfered with their recognition, resulting in the delay of the response latency. The degree of interference was greater in females than in males. Males did not show facilitation of recognition in the overlapped condition. It might be reasonable to interpret that in general, females performed more cautiously in doing the recognition task than males. It might be the cognitive style of the sexes that differentiate the results.

Conclusion

The result showed that Overlapped characters lengthened the response latency and seemed to interfere with the recognition process.

Acknowledgments

I am grateful to my supervisor, Dr. Richard Shillcock, for his long-term guidance and supervision.

References

Chinese knowledge information processing group (1993). *The most frequent nouns in Journal Chinese and their classification: Corpus-based research series*. Taiwan: Academia Sinica.

Chou, Y. J. (2002). Hemispheric lateralisation in Chinese character recognition. *Unpublished PhD thesis. The University of Edinburgh.*

Mack, A. & Rock, I. (1998). *Inattentional Blindness.* Cambridge, MA: MIT Press.

Mozer, M. C., P. W. Halligan, and J. C. Marshall. (1997). The end of the line for a brain-damaged model of hemispatial neglect. *Journal of Cognitive Neuroscience* 9(2): 171-190.

Nazir, T. A. (2000). Traces of print along the visual pathway. In: Kennedy, A., Radach, R., Heller, D. & Pynte, J. (Eds). *Reading as a perceptual process* .North-Holland.

Learning the Dynamics of Vowel to Vowel Phonotactics

Orlando Bisacchi Coelho (orlandoc@terra.com.br)
UMC / FEEC & IEL – UNICAMP
LAFAPE - C.P. 6045 – CEP 13084-970 – Campinas – SP - Brazil

Edson Françozo, Eleonora Albano, Laudino Roces, Pablo Arantes & Renato Basso
(edson,albano,laudino,pablo,renatomb@iel.unicamp.br)
LAFAPE – IEL – UNICAMP

Introduction

Is phonetic information encoded by distributional biases in the lexicon? Are phonotactic constraints robust enough to help a learner infer the phonic pattern of a language? Our work in progress attempts to shed light on these questions via: (i) statistical description of the distributional biases in phone sequences in a lexical database and (ii) connectionist simulation. The simulation focuses on V-to-V relations in V(C)'C(C)V phone strings since both harmony and contour constraints (the tendency for the vowels to share or avoid repetition of phonic properties, respectively) have been found in the distributional study (Albano, 2002).

Experiment and Current Results

An SRN (Elman, 1995), with a compression layer added between the input and the hidden layers, was trained to predict the next phone in the word. The 3700 penultimate stressed words in the training set were fed to the network phone by phone. The context layer activation was reset after each word. Phones were encoded as 35 orthogonal vectors; so no phonetic information was supplied to the network. The training set consisted of trisyllabic nouns only. Since the focus of the experiment was the intervocalic transitions, testing was performed by presenting the network with 12 non-words [pV'pVpV], thus controlling for the effects of consonants. The hidden unit activations produced for each phone in the test set were then analysed using PCA, averaging for the 5 harmonic and 7 contour non-words (means were tested for significance with ANOVA).

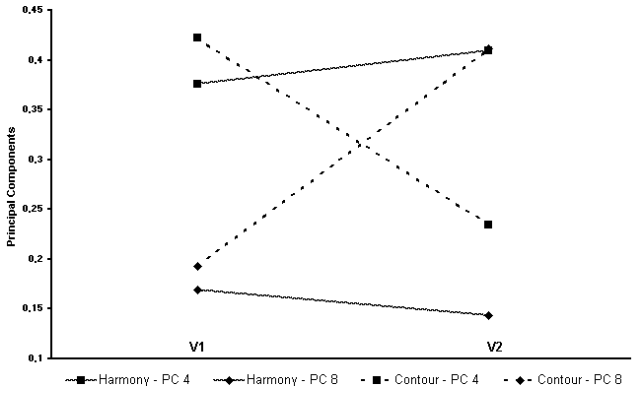


Figure 1: V1 to V2 Transitions according to PCs 4 and 8.

It is possible to identify dimensions in the principal component space which discriminate for Harmony and Contour. Figure 1 depicts the transition between vowels along the dimensions coded by the 4th and 8th principal components. It can be seen that harmonic vowel pairs are not strongly distinguished in any dimension. In contrast, contour vowel pairs differ significantly in both dimensions.

The trajectories associated to transitions in the subspace spawn by the 4th and 8th components (Figure 2) show that the dynamics for harmony is opposite to the one for contour.

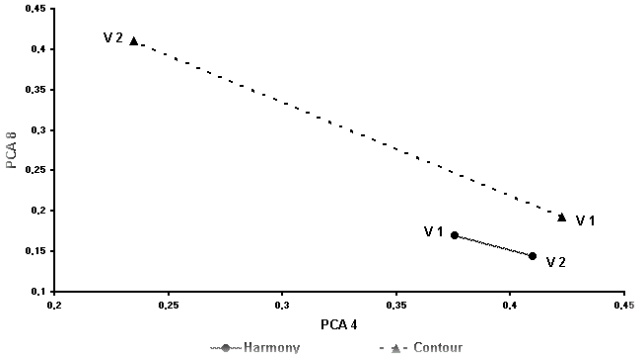


Figure 2: Harmony vs. Contour dynamics for PCs 4 and 8. The direction of the V1V2 vectors is opposed for harmonic and contour vowel pairs.

Phonotactics is best described by non-categorical, probabilistic biases, embodied in the lexicon as constraints on lexical forms which are emergent properties of the operation of dynamical systems that shape language behaviour. So, understanding phonotatics depends on adopting a dynamical point of view.

Acknowledgments

Research funded by FAPESP (01/00136-2).

References

Albano, E. (2002) V-to-V phonotactics, syllable structure, and morphological productivity. *Proceedings of the Eighth Conference on Laboratory Phonology*. http://www.ling.yale.edu:16080/labphon8/

Elman, J.L. (1995) Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion*. Cambridge, MA: The MIT Press.

The Roots of Plausibility: The Role of Coherence and Distributional Knowledge in Plausibility Judgements

Louise Connell (louise.connell@ucd.ie)
Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

Introduction

Plausibility plays a central role in human cognition, whether one is considering the alibi of a murder suspect in a crime novel, or assessing the answers of a candidate in a job interview. Other studies have mentioned plausibility judgements in the service of other phenomena (e.g. Reder, 1982), but often without being investigated in their own right. This paper presents evidence that plausibility judgements depend on inferential coherence and distributional information. In the first experiment, we show that the type of inference being made affects the plausibility of a sentence pair. The second experiment demonstrates that the distributional properties of the words in a sentence pair directly influence plausibility.

Experiments

Two experiments advance a novel paradigm in which people make plausibility judgements about sentence pairs. These sentence pairs are manipulated to invite different bridging inferences and to control their distributional scores (as determined by the Latent Semantic Analysis model LSA; Landauer & Dumais, 1997).

In Experiment 1, 40 participants were asked to judge the plausibility of sentence pairs on a scale from 0 – 10 that had been manipulated to support causal, attributal or temporal inferences, or not to invite any obvious inferences at all (i.e. unrelated pairs). The distributional information of each pair (the LSA score of the first sentence against the second) was controlled across inference types.

In Experiment 2, we manipulated distributional information across the causal and attributal sentences to look at the action of both factors together. 24 participants saw two versions of each sentence pair per page (see Table 1), one of which had a relatively high LSA score between the sentences (a strong distributional link) and the other of which had a relatively low score (a weak distributional link). Participants were asked to judge the plausibility of each pair as before, but to make certain that any perceived difference in plausibility between the two versions of each sentence pair was reflected in the scores.

Results & Discussion

Experiment 1's results demonstrate that different inference types differentially affect the perceived plausibility of a discourse. The causal pairs were rated the highest in plausibility (M=7.8), followed as predicted by attributal (M=5.5), temporal (M=4.2) and unrelated (M=2.0). An analysis of variance yielded a significant effect of inference type on plausibility scores, $F(3, 472) = 93.683$, $p < 0.0001$.

Table 1: Sample Experiment 2 sentence pair variants.

Sentence 1	Sentence 2	Inference X Distribution
The pack saw the fox.	The hounds growled.	Causal Strong
	The hounds snarled.	Causal Weak
	The hounds were fierce.	Attributal Strong
	The hounds were vicious.	Attributal Weak

Experiment 2's results show that the distributional information of a sentence pair affects how plausible it is perceived to be. We examined the proportion of times a participant judged either the strong or weak version of a sentence pair to be more plausible. This analysis shows that in both the causal pairs [M=59.4%, $t(10)=4.893$, $p<0.001$] and in the attributal pairs [M=60.3%, $t(11)=3.753$, $p<0.005$], the weak sentence pair was proportionally rated more plausible than the strong pair.

This gives rise to a very interesting explanation of the joint effects of coherence and distributional strength. We suggest that when there is a strong distributional link, there is an expectation that a coherent inference will be found, and this expectation suggests an initial level of plausibility. When the expectation is borne out – by finding a bridging inference for a strong link, or by not finding one for a weak link – then the level of plausibility suggested by the expectation remains unchanged. On the other hand, when the expectation is contradicted – by unexpectedly finding a bridging inference for a weak link, or failing to find one for a strong link – then the level of plausibility rises or falls accordingly. While distributional information plays an essential role in the judgement process, the degree of coherence is what ultimately validates the plausibility level.

References

Landauer, T. K. & Dumais, S. T., (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Reder, L. M. (1982). Plausibility judgments vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review, 89*(3), 250-280.

Measures of Real Time Assessment to use in Adaptive Augmentation

Martha E. Crosby (crosby@hawaii.edu)
Curtis Ikehara (cikehara@hawaii.edu)
David N. Chin (chin@ hawaii.edu)
Department of Information and Computer Sciences, University of Hawaii
1680 East West Rd. POST 317, Honolulu HI 96822

Cognitive capabilities vary considerably due to people's different levels of expertise and aptitude. Also the cognitive capabilities of any one person will vary greatly over time because of stress, fatigue, injuries, attention lapses, and distractions. In order to augment the cognition of any individual, one strategy is to first assess the real-time cognitive capabilities of that individual and then tailor any augmentation to the current cognitive capabilities of the individual.

The first major problem in real-time adaptive augmented cognition is assessing cognitive capabilities of the person instantaneously. A number of sensors have detected a student's physiological states and actions. Riseberg et al. (1998) frustrated highly motivated users of a game interface by degrading mouse performance at irregular intervals. Galvanic skin conductivity, blood volume pressure, and muscle tension were recorded and correlated to frustration and non-frustration states. Ark, Dryer, and Lu (1999) established the feasibility of correlating physiological measurements to emotions. Crosby, Chin and Iding (2001) demonstrated how eye fixations can be used to analyze users' strategies. Other research projects by Crosby, et al. (2001) suggest that a person's pressure on a mouse is also a good candidate for real-time biometrics to predict cognitive load. Identifying emotions is a complex process and it may not always be necessary to know the precise emotion. If other information such as performance data and focus of attention are available, it may be sufficient to only sense changes in the physiological data.

The data collected by our projects include temperature, heart waveform, galvanic skin response and physical pressures applied to a mouse. Sensors are connected to an electrically isolated micro-controller, which converts the analog sensor signals to digital data for processing, by a master computer. The first sensor is on the finger tip to monitor peripheral temperature, the second sensor is attached to the wrist as a reference to body temperature and the third sensor monitors the ambient temperature at a distance of about 5 cm away from the wrist. The sensors are not attached to a substantial thermal mass that would reduce the response rate, and the minimum detectable change of temperature is calculated to be 0.005 C.

The heart waveform is monitored using an infrared reflective light sensor that is mounted on the same elastic band as the fingertip temperature sensor. The change in blood flow is measured by changes in reflected light at the skin surface where blood flow is apparent. The heart beat rate is extracted by measuring the time between peaks and the relative blood volume change can be determined by integrating the heart waveform over the desired time. When the sensor is pressed against the fingertip, too much force will obstruct blood flow and too little force will produce unreliable data. The sensor also detects the mean reflected light, which is used to determine the optimum amount of pressure to be applied to the sensor.

The experiments performed in our projects show that simple physiological measurements of GSR, heart rate, temperature, and pupil diameter provide information on changes in user's emotional and subjective states while engaged in cognitive tasks. A subject's physiological state and actions can be indicative of their cognitive performance. Physiological data, collected from people as they use computers to form their tasks, might not consistently predict their emotional state. However, results from our experiments suggest that it is possible to provide the computer with information about the users' cognitive state in real time.

Acknowledgments

This research was supported in part by Office of Naval Research grant no. N00014970578 and DARPA grant NBCH1020004. The authors thank to IBM's project Blue Eyes for supplying two prototype emotion mice.

References

Ark, W., Dryer, D. and Lu, D. (1999). *The Emotion Mouse*. In Human-Computer Interaction: Ergonomics and User Interfaces, Bullinger. H. J. and J. Ziegler (Eds.), Lawrence Erlbaum Assoc., 818-823.

Crosby, M., Auernheimer, B., Aschwanden, C., and Ikehara, C. (2001). *Physiological Data Feedback for Application in Distance Education*. Workshop on Perceptive User Interfaces, Florida.

Crosby, M., and Chin, D. and Iding, M. (2001) *Using Search Behavior in Complex Interfaces for the Design of Adaptive Systems* Proceedings of 2001 User Modeling Conference, Germany.

Riseberg, J., Klein, J., Fernandez, R., and Picard, R.W. (1998) *Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state*. *CHI'98 Late-Breaking Results*, 227-228.

Semantic Memory Retrieval During Conditional Reasoning: Every Counterexample Counts

Wim De Neys (Wim.Deneys@psy.kuleuven.ac.be)

Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)

Géry d'Ydewalle (Géry.dYdewalle@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Introduction

Reasoning with conditionals involving causal content is known to be affected by retrieval of counterexamples from semantic memory. This study focuses on the characteristics of this search process.

In Markovits' (2000) recent specification of the memory search process, the number of stored counterexamples is important because it determines the probability that at least one can be retrieved. This specification does not address the impact of additional counterexample retrieval. Indeed, the search process is assumed to stop after the successful retrieval of a single counterexample.

The present study tests an alternative specification of the search process. We examine the assumption that the search process does not terminate after the retrieval of a single counterexample and that every retrieved counterexample has an additional impact on the reasoning process. Here, the number of stored counterexamples will be important because it determines the number of counterexamples that can be retrieved and this number would determine the degree to which inferences will be accepted.

Experiment

A generation pretest measured the number of counterexamples (alternative causes or disabling conditions) participants could retrieve for a set of causal conditionals. One month after the pretest, participants were presented a reasoning task with the same conditionals. We looked at participants inference acceptance ratings for each conditional in function of the number of counterexamples they could retrieve for that conditional.

Results showed that every alternative or disabler that can be retrieved has an impact on the inference acceptance. Acceptance of Modus Ponens and Modus Tollens linearly decreased with every additionally retrieved disabler. Likewise, Affirmation of the Consequent and Denial of the Antecedent acceptance linearly decreased in function of the number of retrieved alternatives.

These graded effects of up to four different numbers of available counterexamples can not be explained if the semantic search process during conditional reasoning would stop after successful retrieval of a single counterexample. This makes it clear that Markovits' (2000) conditional reasoning model needs to be revised.

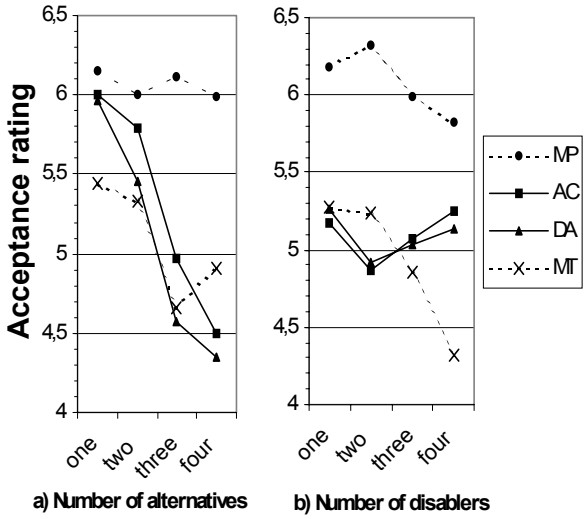


Figure 1. Inference acceptance (7-point scale) in function of the number of alternatives (2a) or disablers (2b) participants could retrieve for a conditional.

References

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Thinking and Reasoning, 6*, 335-347.

Categorization of Emergent Processes by Students at Different Levels of Expertise

Randi A. Engle (RAEngle@pitt.edu)
Michelene T. H. Chi (chi@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260 USA

Introduction

Chi and Roscoe (2002) proposed that one reason for the persistence of scientific misconceptions is that students classify scientific phenomena into incorrect ontological categories. In particular, Chi (submitted) has hypothesized that many commonly misunderstood science concepts (like evolution and electric current) are emergent processes in which a macro-level phenomena emerges from complex interactions between entities at a micro-level. Rather than correctly categorizing them this way, students are thought to miscategorize them as non-emergent processes in which there is a more direct relationship between the two levels.

To test this hypothesis, we are conducting a study comparing how participants at different levels of expertise categorize science problems across domains, using a card sorting task modeled on Chi, Feltovich & Glaser (1981). The prediction of Chi's theory is that participants with more expertise will be more likely to distinguish emergent from non-emergent problems while those with less expertise will often conflate them. Experts will also be more likely to refer to the 11 features of the emergent schema (e.g., disjointness, parallelism; see Chi, submitted) in defining their categories.

Method

Participants

Participants consisted of 10 undergraduate and 9 doctoral students from the biology, chemistry, or physics departments at a local university. Undergraduates had all completed 1st year courses in biology, chemistry, and physics. Eight were single majors in one of these disciplines and two were double majors in biology and chemistry. Doctoral students were in their third year or above.

Materials

Participants sorted 24 science problems, 8 drawn from each discipline's 1st year course. Within each discipline, half were emergent and half were non-emergent processes.

Procedure

Participants were asked to sort the problems into piles with similar mechanisms for linking the macro and micro levels. In the 1st sort, they were allowed to make as many piles as they wished. In the 2nd sort, they were asked to divide the cards into just two piles. In the 3rd sort, the experimenter sorted the cards into emergent vs. non-emergent processes, and participants were asked to infer the distinction. In all cases, participants were asked to explain the explanatory mechanism each pile represented and why each problem fit.

Results

As a preliminary measure of the degree to which participants distinguished emergent from non-emergent processes, we calculated a weighted average (by pile size) of the absolute difference between the number of emergent versus non-emergent processes in each pile by the number of cards in the pile. For example, 3 piles—one with 5 emergent & 5 non-emergent problems, a 2nd with 1 emergent & 7 non-emergent problems, and a 3rd with 6 emergent & 0 non-emergent problems—would get a score of $[10(0/10) + 8(6/8) + 6(6/6)] / 24 = 12/24 = .50$. A value of 1 on this score represents perfect separation of emergent versus non-emergent processes while 0 represents perfect 50/50 mixing. Mean separation scores were .58 for doctoral students, .53 for single majors, and .73 for double majors, although none of these differences are statistically reliable.

A complicating factor is that participants sometimes put, for example, only emergent problems in a given pile for reasons unrelated to their emergence. Thus, we analyzed participants' definitions of their categories to determine whether they referred to any of the 11 features of the emergent schema. Doctoral students' categories referred to more emergent features (1.78) than single majors (0.13; $t(9) = 3.08$, $p < .05$). Double majors referred to even more of them (3.50), but more data is needed to see if this is reliable.

Discussion

Doctoral students and double majors used more features of the emergent schema in sorting science problems than single majors, although they were no more likely than single majors to create piles that distinguished emergent and non-emergent processes. In future work, we will include in our sample more double majors as well as professors to further investigate the nature of expertise in emergent processes.

References

Chi, M. T. H. (submitted). "Causal" and "Emergent" Explanatory Mechanisms: Potential Schemas for Overcoming Misunderstandings in Science.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Chi, M. T. H., & Roscoe, R. (2002). The processes and challenges of conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice.* Dordrecht, NL: Kluwer.

The Autonoetic Hypothesis On Creativity: Memory and Cognition in Pollock s Abstract Art.

Carlos H. Espinel (espinelmd@aol.com)
The Blood Pressure Center, Washington, and Georgetown University Medical Center, Dept. of Medicine
1715 North George Mason Dr., Suite 401, Arlington, VA 22205 USA

Art and Cognitive Science.

Art presents a challenge and an opportunity for study in cognitive science (Espinel, 1994; 1996; 1998). This research concerns the creation of abstract art. One that is supposed to impact the viewer without recognizable, without memory-evoking images (Schapiro, 1980). This ideal was achieved by Jackson Pollock, who with it pioneered modern American art. His paintings, tangles of smears of paint on canvas, have been explained as unconscious expressions (Schapiro, 1980 & O Connor FV, Thaw EV, (Eds.), 1978). An Autonoetic Hypothesis on creativity would suggest otherwise. Here is a study of a painting representative of Pollock s abstracts, Lavender Mist, at the National Gallery of Art (Figure 1).

Method.

An investigative approach was devised previously to study works of art (Espinel, 1994; 1996; 1998; 2002). This method is applied to analyze concerning Lavender Mist : 1. Reports, Pollock s own and art experts , about his creation; 2. In the painting, by direct observation, photography, and digital techniques devised by the author (Espinel, 2002): a) physical and aesthetic characteristics, and b) systematically, cues of space and form; 3. Other works by Pollock, exhibited and published.

Results.

1. Reports did not disclose memories, thoughts or experiences that might have originated Pollock s creation. 2. In Lavender Mist : a) Direct observation and photography showed no recognizable space or form cues (Figure 1); b) Digital analysis by color revealed 9 layers of smears, called here, digital paintings (Figure 2); c) Of these, 8 showed no recognizable space or form; but d) In the ninth, The Pink, the image of a horned animal was discovered (Figure 3). 3. In a Pollock drawing from his symbolic stage, 10 years before, an image was found that is reminiscent of the one hidden in Lavender Mist (Figure 4).

Conclusions.

1. The discovery of the horned animal image under tangles of smears suggests Pollock s awareness of it, and a conscious, deliberate concealment. 2. That the image can be traced to a past event, to a contextually specific stage of his artistic development, suggests Pollock s autonoetic recall. 3. The image, perhaps as a form of episodic memory (Tulving, 2001), might have been germinal to the creation of Lavender Mist.

The Autonoetic Hypothesis .

Autonoesis entails the conscious re-experience of past events in the present. A form of episodic memory might be crucial to the creation of abstract art.

References.

Espinel CH. (1994). Caravaggio's "Il Amore": A sleeping cupid with Juvenile Rheumatoid Arthritis. *The Lancet,* 344,1750-1752.

Espinel CH. (1996). de Kooning's late colours and forms: dementia, creativity, and the healing power of art. *The Lancet,* 347,1096-8.

Espinel CH. (1998). Art and neuroscience: how the brain sees Vermeer's "Woman Holding a Balance". *The Lancet,* 352,2007-9.

Espinel CH. (2002). ArtMedicine Software: Neuroart-analysis Applications. Patent pending.

Schapiro M. (1980). *Modern art:19th and 20th centuries.* New York: George Braziller.

Jackson Pollock: a catalogue raisonne of paintings, drawings, and other works. Volumes I-IV. (1978). O Connor FV, Thaw EV, (Eds.) New Haven: Yale University Press.

Tulving, E. (2001). The origin of autonoesis in episodic memory. In Nilsson, LG and Markowitsch HJ (Eds.), *The Nature of Remembering*. Washington DC:Am. Psychol. Assoc.

Epistemic Belief and Semantic Categorization

Zachary Estes (estes@uga.edu)
University of Georgia
Department of Psychology
Athens, GA 30602 U.S.A.

Introduction

People tend to believe that membership in an artifact category (e.g., FURNITURE) is a subjectively decided matter of opinion, while membership in a natural category (e.g., FRUIT) is an objectively determined matter of fact (Malt, 1990). I argue that these different beliefs across domains affect categorization in important and predictable ways. If membership in a natural category is an objective, right-or-wrong matter, then categorization should be an absolute, all-or-none decision. But if membership in an artifact category is a subjective matter of opinion, then categorization need not be absolute, but rather may be a matter of degree. Epistemic belief may also affect the confidence with which category membership is judged. If membership in a natural category is objective, then it is possible for the category judgment to be incorrect, and therefore people may sometimes lack confidence in their category judgments. But if membership in an artifact category is subjective, then individuals are entitled their own opinions of the matter. Because opinions are not open to verification or rejection, people may have confidence in their category judgments. Thus, people may be more confident in their judgments of artifacts than of natural kinds.

Experiment 1

In Experiment 1, participants judged the category membership of artifacts and natural kinds, and also rated their confidence in those category judgments. Results indicated that artifact categories were more graded than natural categories. Artifact categories were also judged with more certainty than natural categories. This pattern of results is precisely what one would predict, given the prior evidence that people consider membership in artifact categories to be subjectively decided, while membership in natural categories is believed to be objectively determined. Thus, one may infer that belief affects categorization.

Experiment 2

Experiment 2 attempted to predict an individual's tendency to give graded membership ratings, on the basis of his or her epistemic beliefs. Epistemic beliefs were measured by Schommer's (1998) "certainty of knowledge" questionnaire, which consisted of statements intended to measure one's belief that truth is objective and certain (e.g., "truth is unchanging,"). Participants rated the extent to which they agreed with these statements. They then completed the same categorization task used in Experiment 1. The correlation between "certainty of knowledge" scores and the proportion of graded responses to artifact categories did not approach significance, $r = -.01$. Critically, however, the correlation between participants' "certainty of knowledge" scores and their proportions of graded membership responses to natural categories was significant, $r = -.37$, $p = .02$. The more a participant believed that knowledge is certain or objective, the less likely she was to provide graded judgments for natural categories. Thus, the belief that knowledge is certain reliably predicted categorization behavior.

Discussion

People's categorization behavior was consistent with their epistemic beliefs (Experiment 1), and moreover, one's epistemic beliefs predicted his own categorization behavior (Experiment 2). Thus, epistemic belief may determine semantic categorization. The claim that lay philosophical beliefs affect categorization is not without precedent. Psychological Essentialism (see e.g., Medin & Atran, 1999) posits that people hold essentialist beliefs, and that these beliefs affect cognition. The present argument is similar. The belief that membership in a natural category is an objectively determined matter of fact leads people to provide absolute judgments of natural kinds, despite the fact that people have relatively low confidence in this objective knowledge. And the belief that membership in an artifactual category is a subjectively decided matter of opinion leads people to provide graded judgments of artifacts, and people have high confidence in these subjective opinions. The present experiments, by showing a correlation between epistemic belief and categorization behavior, provide the first direct demonstration of the relation between epistemic belief and semantic categorization.

Acknowledgments

I am especially grateful to Marlene Schommer for the use of her Epistemological Questionnaire. I also thank Elizabeth Barrett, Chris Evans, Roger Gaudreau, and Katie Goodrum for their diligent help with data collection.

References

Malt, B.C. (1990). Features and beliefs in the mental representation of categories. Journal of Memory and Language, 29, 289-315.

Medin, D.L. & Atran, S. (1999). Folkbiology. Cambridge, MA: MIT Press.

Schommer, M. (1998). The influence of age and schooling on epistemological beliefs. The British Journal of Educational Psychology, 68, 551-562.

Learning from Transformational and Derivational Worked-out Examples

Peter Gerjets (p.gerjets@iwm-kmrc.de) & Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Knowledge Media Research Center & Department of Psychology, University of Tuebingen
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Stefan Kleinbeck (Stefan.Kleinbeck@Psychologie.Uni-Freiburg.de)
Department of Psychology, University of Freiburg
Niemensstrasse 10, 79085 Freiburg im Breisgau, Germany

Ute Schmid (schmid@informatik.uni-osnabrueck.de)
Department of Computer Science, University of Osnabrueck
Albrechtstrasse 28, 49069 Osnabrück, Germany

In this research two different solution formats for instructional worked-out examples were compared experimentally with regard to several measures of learning outcomes. A typical example problem from the domain of probability theory used for experimentation is the following:

Problem statement: At the Olympics, 7 sprinters participate in the 100m-sprint. What is the probability of correctly guessing the winner of the gold, silver, and the bronze medal?

The worked-out solution designed for this type of problem according to the *transformational example format* was inspired by a "structure mapping view" of analogical transfer (Gentner, 1983). According to this view, transformations of complex problem representations into another (in terms of structure mapping/ schema induction) are pivotal processes for learning and problem solving. Instructional worked-out examples designed from this perspective may aim at conveying structural problem features necessary to recognize problem categories. Problem categories comprise classes of isomorphic problems that can be transformed into another (using analogy) and that can be represented in a more abstract way (using problem-type schemas). A problem-type schema consists of information about the defining structural features and the appropriate solution procedures for a class of problems. In our experiments the transformational example solutions had the following structure:

Problem features: *Selection* of 3 sprinters out of seven sprinters; *order* of selection is important; each sprinter can only be selected *once* (*without replacement*)
Formula: A = n! / (n-k)!
Inserting: $n = 7$, $k = 3 \Rightarrow$ A = 7! / (7-3)! = 210
Result: 1/210 ≈ 0.48%

Learners are known to have difficulties in acquiring structural problem features and problem categories as well as in adapting solution procedures to novel problems that differ from the problem categories conveyed. Thus, we compared the transformational approach to a different instructional approach that was inspired by AI models of derivational analogy (Carbonell, 1984). The main idea of the derivational approach is to convey knowledge on *how to derive solutions* for problems regardless of their problem category and to abandon the mapping/ categorization of problems as well as the application of category-specific solution procedures. In our experiments the basic rationale for the *derivational example format* is to decompose a complex event into a sequence of individual events. The overall probability is calculated by multiplying the individual-event probabilities. Contrary to the transformational approach, the solution strategy conveyed in the derivational approach is not seen as a solution schema that is applied to the problem as a whole. Instead it is presented as a sequence of solution steps where each step can be justified (and modified) by concrete features of the problem at hand. Therefore, the derivational example format is characterized by a high modularity. The derivational example format had the following structure:

Rationale: Calculate probability of correctly guessing the winner of each medal; each medal can be taken into account *separately.*
Step 1: *7* possible choices (sprinters), *1* acceptable (winner) ⇒ 1/7
Step 2: *6* possible choices (winner can't win two medals), *1* acceptable choice ⇒ 1/6
Additional steps: Analogous procedure
Result: *Overall probability*: 1/7 * 1/6 * 1/5 = 1/210 ≈ .48%

In order to compare both example formats two hypertext-based experiments using different dependent measures (e.g., learning time, example processing strategies, problem-solving performance, problem classification, problem comparison) were conducted. The results show a clear superiority of the derivational example format with regard to learning time and problem-solving performance.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft . We thank R. Catrambone for helpful comments and D. Bruns, C. Krämer and V. Länge for conducting the experiments as well as S. Albers for programming.

References

Carbonell, J. G. (1984). Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell & T. M. Mitchell (Eds.), *Machine learning: An intelligence approach* (pp. 137-161). Berlin: Springer.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

The Role of Cognitive Modeling in Enhancing Dynamic Decisions

Cleotilde Gonzalez (coty@cmu.edu)
Social and Decision Sciences Department
Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3890 USA

Dynamic Decision Making is characterized by multiple and interdependent decisions, autonomous environments, and real-time evaluation and action (Brehmer, 1990). There seems to be increasing agreement on how we make decisions in this type of situations. Decision makers in dynamic environments recognize typical situations and typical responses and use their knowledge to adapt their strategies "on the fly" (Payne, Bettman, and Johnson, 1993; Klein, 1998).

Many cognitive models have been developed in the past decade to investigate different aspects of cognition in these environments: attention and multitasking (Lebiere, Anderson and Bothell, 2001; Altmann and Gray, 2000); judgment and choice in decision making (Gonzalez, Lerch and Lebiere, submitted); and skill acquisition in dynamic situations (Wallach and Lebiere, 2000; Schoppek, Holt, Diez and Boehm-Davis, 2001). These efforts are commendable for focusing on complex, real world tasks. However, the role of cognitive modeling in enhancing decision-making in these situations has been very static and limited.

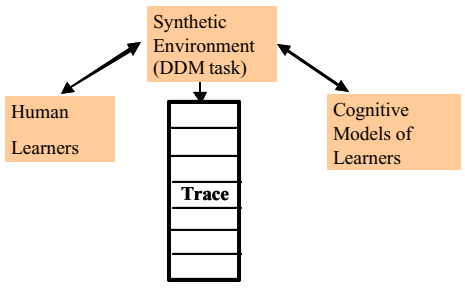
Traditionally, cognitive models are developed within the context of a task. A synthetic version of the task exists in the form of a computer program, and a cognitive model is developed to interact with such a tool. On the other hand, the synthetic task environment supports the collection of behavioral data and the development of the cognitive model. Theories expand and conclusions are reached about human cognition based on the comparisons between human and model data. I believe that cognitive modeling may and should take a more dynamic role in enhancing decision-making. Creating cognitive models to perceive a situation, combine goals and beliefs, choose a course of action, and react over the environment, is a challenge that we should overcome.

In my current research, I intend to enhance existent cognitive models to take an active, real-time role in decision making and learning. An example of this effort is the cognitive model of instance-based decision-making, developed in the Water purification Plant (WPP) task (Gonzalez, Lerch and Lebiere, submitted). WPP is a resource allocation, scheduling task, isomorph of a real world task in an organization with large-scale logistical operations (the United States Postal Service: USPS) (Lerch, Ballou, and Harter, 1997). WPP is highly dynamic because of exogenous events occurring without the decision maker's control and it is complex because of the number of options and alternatives to consider at any given time. All these elements interact with the decision maker's actions in a real-time environment, and events occur as the simulation clock is running. Based on an Instance-Based Decision Making framework, I have developed a cognitive model (built on ACT-R) to interact with the WPP task. At present, the model accurately simulates the variability of choice in this dynamic situation and allows strategy adaptation and flexibility. But, the model works off-line, not directly incorporated in monitoring the status of a real decision maker in real-time.

I intend to use this cognitive model to help track and adapt the conditions of WPP in real-time, according to the cognitive necessities of a human decision maker interacting with the task. By tracking each of the actions taken by a human decision maker and monitoring the status of the task environment, the cognitive model would be able to predict the possible contents of working memory and the state of knowledge in the form of instances (See figure). Based on the predicted cognitive status of human decision makers, the cognitive model may give advice to the decision maker in the form of:

1. attracting attention to the parts of the screen that have not been acted on and require awareness according to the environmental status
2. Presenting information in multiple modalities to spread the load to multiple channels
3. Perform urgent actions that the decision maker would not have time or resources to react to

Although simple and limited, I believe that, this effort may demonstrate an active and dynamic role of cognitive modeling in complex situations.



References

Due to space limitations, I have not included the references. Please request this list to: coty@cmu.edu

Developing a Framework for Understanding Scientific and Technological Thinking: Notes from a Workshop

Michael E. Gorman (meg3c@virginia.edu)
Division of TCC, SEAS, University of Virginia
Charlottesville, VA 22904-4744 USA

Alexandra Kincannon (kincannon@virginia.edu)
Department of Psychology, University of Virginia
P.O. Box 400400, Charlottesville, VA 22904-4400 USA

On March 24, 2001, Alexandra Kincannon, Ryan Tweney and Michael Gorman convened a workshop on cognitive studies of science and technology at the University of Virginia (Gorman, Kincannon, & Mehalik, 2001). We assembled a multi-disciplinary group of practitioners to discuss the latest research and methodologies, identify the stumbling blocks to advancement in this area, and think about directions for the future. The workshop was dedicated to Herb Simon, who was slated to participate.

Two questions became central themes. First, how can we combine in vitro experimentsÕwith in vivo case studies of actual practice? Results obtained in the laboratory may have low ecological validity. Fine-grained case studies are often domain-specific and hard to generalize.

Second, how can we deal with academicsÕattachments to their own hypotheses and methods? Researchers tend to overgeneralize hypotheses developed under specific in vitro or in vivo conditions.

One way of avoiding this kind of overgeneralization is to combine in vitro and in vivo methods. With help from workshop participant David Klahr and others, we developed a preliminary framework based on the idea of searches in multiple problem spaces, and identified which had been investigated in vitro and which in vivo. For example, hypothesis and experiment spaces have been investigated both in vitro and in vivo, but function and design spaces have only been studied in vivo. This approach helps identify areas for future research.

The workshop illustrated that frameworks can be shared and that in vitro and in vivo studies have to complement one another. Theories need to deal rigorously with the distributed character of scientific and technological problem solving. We hope this workshop will suggest directions for future applications as well as research.

See http://repo-nt.tcc.virginia.edu/cogwkshop/index.html for more information about the workshop and the participants.

Acknowledgments

This workshop was made possible by the generous support of the National Science Foundation, the Strategic Institute of the Boston Consulting Group, and the National Collegiate Inventors and Innovators Alliance.

References

Gorman, M. E., Kincannon, A., & Mehalik, M. M. (2001). Spherical horses and shared toothbrushes: Lessons learned from a workshop on scientific and technological thinking. In K. P. Jantke & A. Shinohara (Eds.) *Discovery Science: 4th International Conference Proceedings* (pp. 74-86). Berlin: Springer-Verlag.

Automated Detection of Strategies in Free Text Responses

Anthony Harrison (anh23@pitt.edu)
Lelyn Saner (les53@pitt.edu)
Celestine Cookson (clcst70@pitt.edu)
Darcie Kunder (dakst67@pitt.edu)
Christian D. Schunn (schunn@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260 USA

When solving problems, people often use a wide array of different strategies. Effective teaching often requires isolating what strategies students are using (or not using) in order to more effectively structure the instructional intervention. Nowhere is this truer than in the realm of intelligent adaptive tutors. The classification of strategy use in complex domains presents an interesting challenge to intelligent tutors. This is made even greater if the strategies are to be extracted from free text responses given by the students.

To this end, we have been using Latent Semantic Analysis (LSA) as an automatic strategy classification tool. LSA is a computational tool that extracts the co-occurrence of words in a corpus. Through high-dimensional matrix decomposition, LSA is able to produce a "semantic-space" allowing all experienced words, phrases, and sentences to be represented as vectors within that space. The more similar the vectors are to each other, the more similar their meanings. As LSA has matured, some have suggested that it may be a psychologically plausible theory of semantic learning. We remain noncommittal in this regard, choosing instead to rely upon LSA in its original capacity as a fast and efficient text-processing tool.

Strategy Classification

Our current endeavor is to use LSA to intelligently classify strategy use in day-to-day military operations. The hope is that by accurately classifying young officers' strategy uses, we can develop tutoring systems to broaden their range of strategies as well as train them to more appropriately apply the strategies.

The strategy classification system relies upon a series of key steps. First the LSA semantic space was generated based on a set of military handbooks, training documents, and pedagogical examples. Free text responses to military scenarios were collected from officers in training as well as experienced military officers. These were then human coded into different strategy categories. The responses were then fed into LSA to generate their vector representations in the semantic space. These two sources yielded two databases of semantically coded (vectors in semantic space) strategies. The novice database (officers in training) is used as a descriptive reference, while the expert database (experienced officers) provides the normative references. The final steps are to take the free text responses of other novices to the same vignettes, transform them into vectors in semantic space, and then each of these vectors is compared against those in the databases. Since they are vectors, the cosine between the two serves as a simple similarity score. As similarity increases, the cosine value will approach one. This process yields a ranking of similarities to the descriptive database, where the classified strategy is merely the most similar. Additionally, since we have a sample of strategy exemplars, we can also look at the distribution of similarity scores across strategies. This yields a simple measure of *confidence*: the greater the number of high similarity matches within a given strategy gets, the more confident we can be that it is representative of that strategy.

At this early stage in the development of the system, we were pleased to see that LSA was classifying strategies about as well as our human coders, with almost equivalent inter-rater reliabilities. This is a significant accomplishment given how limited our semantic space is currently (only 100,000+ words, in comparison to the millions of most other LSA corpora), and the limited scale of our descriptive database (10 strategies, approx. 16 exemplars each).

Future Directions

Aside from increasing the scale of both the semantic space and the reference databases, we hope to begin working on the tutoring system proper. This will mean developing a training system that adapts to the strategy use of the individual to provide sufficient scaffolding to enable them to explore alternative strategies, as well as to learn how to appropriately apply them. Then, as the student progresses through the tutor, the normative database (provided by experienced military officers) will come into greater play.

References

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2).* 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

ACT-R/S: A Computational and Neurologically Inspired Model of Spatial Reasoning

Anthony M. Harrison (anh23@pitt.edu)
Christian D. Schunn (schunn@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara St, Pittsburgh, PA 15260 USA

The field of cognitive modeling has seen a recent push in two major areas: embodied cognition, and neurological realism. No longer is it sufficient to show that a model of cognition can produce a specific behavior without actually interacting with its environment in someway, be it a real environment or simulated. Nor can psychologists ignore the fact that for every system, representation, rule, and computation proposed there must be some underlying neurological reality behind it. With both these constraints in mind, we set out to develop an extension to ACT-R (Anderson & Lebiere, 1998) allowing it to enter into a three-dimensional world in a neurologically plausible manner.

ACT-R/S (spatial) relies specifically upon three processing modules, only two of which are new to the architecture. Each of these modules has been shown to be both behaviorally and neurologically separate. The representations and computations of each of the systems are similarly distinct.

Three Visiospatial Systems

Visual System The primary function of a visual system is to identify a set of visual features as an object. The visual system needs to be able to take fine-grained detail and through special processing, recognize an object. A feature of this system is that it is able to perform its task based off of basic two-dimensional retinotopic information. An object's depth or spatial extent is not typically necessary for its accurate identification. This functionality is currently available in Mike Byrne's ACT-R/PM (perceptual & motor extension).

Neurologically, the visual system is seated in the primary visual areas as well as the ventral visual processing stream which limits processing to fine detail, color perception, local form perception, visual scanning and visual feature analysis (see Previc, 1997 for review).

Manipulative System When it comes to grasping and manipulating objects, we need to be able to represent them in a manner that will enable us to effectively prepare the motor system for the task ahead. The manipulative system is concerned entirely with a metric, geon-based (Biederman, 1987), three-dimensional representation of objects. These representations are then typically fed to the motor system permitting the development of complex motor programs. The manipulative system can represent almost any three-dimensional object, but its primary purpose is to support actual manual manipulation.

The manipulative system relies upon the dorso-lateral visual stream as well as the parietal cortex. The involvement of the parietal cortex is not surprising given that these tasks often involve actual manipulation. However, when subjects are asked to imagine object rotations, the parietal cortex is still often activated (see Previc, 1997 for review).

Configural System The configural system is concerned with representing objects in space to facilitate navigation. It represents the world around us as spatial blobs that need to be navigated around, above, or below. Its representations are nowhere near as precise as those found in the manipulative system. It encodes the locations of objects in terms of egocentric vectors that are continuously updated through path-integration. The utilization of multiple landmarks allows the system to uniquely position itself in space and return to locations at later points in time.

The discovery of "place-cells" in the rat hippocampus has been viewed as the definitive location of cognitive-maps in the brain (O'Keefe & Nadel, 1978). Recent research has shown that the parahippocampal regions are more important in primate navigation but they still represent some form of a map of the environment. Our own meta-analysis brings the "egocentric" assumption of "place-cells" into question, hence our usage of egocentric vectors in the configural representations.

Summary

With the proposal of two additional processing systems that specialize specifically in three-dimensional processing, it is hoped that we will be able to expand the range of phenomenon that computational cognitive models can represent. We present this not only for the ACT-R architecture, but also so that other architectures might get a foothold in three-dimensional embodiment.

References

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Mahwah, NJ: Erlbaum.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115-117.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon.

Previc, F. H. (1997). The neuropsychology of 3-D space. *Psychological bulletin, 124,* 123-164.

Belief Revision and Reasoning

Uri Hasson (uhasson@princeton.edu)
Psychology Department, Princeton University
Princeton, NJ 08544 USA

Philip N. Johnson-Laird (phil@princeton.edu)
Psychology Department, Princeton University
Princeton, NJ 08544 USA

Which beliefs do you abandon when you discover that their consequences conflict with the facts? Suppose, for example, you have the following beliefs:

Belief 1: If the drink is cold, then it is caffeinated.

Belief 2: The drink is cold.

You make an inference of the form known as *modus ponens* (MP): The drink is caffeinated. But then you discover:

Fact: The drink is not caffeinated.

You are likely to reject any belief that is dubious or from a dubious source. Perhaps you will be more likely to reject a simple categorical assertion, such as belief 2, than a generalization, such as belief 1 (Revlis, 1974). Conversely, you will be more likely to retain believable generalizations (Dieussaert et al., 2000, Politzer & Carles, 2001). But, if your beliefs are equally plausible, then a pertinent factor is whether there is an apparent conflict between the facts and your beliefs. According to the theory of mental models, such conflicts can occur (Girotto et al., 2000). The model theory postulates that belief 1 calls for two mental models:

Cold Caffeinated
. . .

The first model represents the possibility in which the antecedent is true; the second model has no explicit content but represents the possibilities in which the antecedent is false. The model of the fact:

Not caffeinated

conflicts with the explicit model above, and so you should reject the conditional. In fact, the conflict is apparent, not real.

Consider a contrasting case in which you believe:

Belief 1: If the drink is cold, then it is caffeinated.

Belief 2: The drink is not caffeinated.

You can make an inference of the form known as *modus tollens* (MT), but its conclusion is contradicted by the fact:

Fact: The drink is cold.

The fact matches the explicit model of the conditional, but is not represented in the model of the categorical premise (belief 2). Hence, the theory predicts that you will be more likely to reject the categorical premise.

We carried out experiments to test these predictions. Participants were presented with scenarios such as the one above, and were asked to decide which of the beliefs they found more credible. Half of the scenarios were in the form of conflicts with MP inferences, and half of them were in the form of conflicts with MT inferences. In addition, the consequent of the conditional statement was either affirmative or negative. And in half the scenarios, the conditional statement was the first belief in the set, and in the other half it was the second belief.

The participants were more likely to reject the conditionals in the MP scenarios (60% rejected) than in the MT scenarios (47% rejected; see also Dieussaert et al., 2000; Elio & Pelletier, 1997), but the difference was reduced when the conditionals had negative consequents. There was also a bias to believe whichever statement was presented first: for MP scenarios, the conditional was more believable when it came first in the set, but less believable when it came second. But, for MT scenarios, the effect of order was diminished.

When individuals notice the MP inconsistency, then they can readily reject the conditional, especially when it is the most recent statement in the set. MT inferences are harder, but individuals can also notice the inconsistency by, in effect, converting the scenarios into MP inconsistencies. They use the fact and the conditional to draw a conclusion, and then they notice that the conclusion conflicts with the categorical belief. And so they reject this belief.

We conclude that belief revision depends on how individuals represent their beliefs, and on how they reason about them. They may reject a belief because it merely appears to be inconsistent with the facts.

References

Dieussaert, K., Schaeken, W., Neys, W. D., & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Current Psychology of Cognition, 19*(3), 277-288.

Elio, R., F.J. Pelletier. (1997). Belief change as propositional update. *Cognitive Science*, 4, 419-460.

Girotto, V., Johnson-Laird, P.N., Legrenzi, P., and Sonino, M. (2000) Reasoning to consistency: How people resolve logical inconsistencies. In Garcia-Madruga, J., Carriedo, M, and Gonzalez-Labra, M. J. (Eds.) Mental Models in Reasoning. Madrid: UNED. Pp. 83-97.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking and Reasoning, 7*(3), 217-234.

Revlis, R. (1974). Prevarication: Reasoning from false assumptions. *Memory & Cognition, 2*(1A), 87-95.

Recency Effects in Category Learning are Dynamic and Adaptive

matt jones (mattj@umich.edu)
Department of Psychology, The University of Michigan
525 E. University Ave, Ann Arbor, MI 48109-1109 USA

Winston R. Sieck (sieck.3@osu.edu)
Department of Psychology, The Ohio State University
1827 Neil Ave, Columbus, OH 43210 USA

Recency Effects (REs) have been well established in both memory and probability learning paradigms. However, these effects have received relatively little attention, and virtually no empirical investigation, in research on multiple-cue category learning.

Some category learning models do make specific a priori predictions regarding REs. In particular, simple, static REs arise naturally in models incorporating the δ-rule, an error-driven learning mechanism (Gluck & Bower, 1988; Kruschke, 1992). Specifically, the δ-rule implies that responses are based on a weighted average of past cases, with recent cases receiving more weight. Weights drop at an exponential rate with temporal distance. Exemplar models make no a priori predictions regarding RE, but can mimic δ-rule predictions by explicitly incorporating exponential trace decay (Nosofsky, Kruschke, & McKinley, 1992). Both of these approaches predict RE to be unresponsive to characteristics of the task environment.

The present research contrasts this assumption with the alternative possibility that people explicitly test hypotheses concerning the predictive validity of recent outcomes. The latter idea implies that the magnitude of recency effects will adapt to the level of autocorrelation in the category sequence.

Method

Participants (n=100) engaged in a simulated medical diagnosis task in which they assigned each of a series of 150 hypothetical patients to one of two fictitious diseases based on the presence or absence of 3 symptoms, with outcome feedback given after each trial.

The experiment consisted of three conditions based on the manner of autocorrelation present in the sequence of disease outcomes. In the *positive* condition, each patient's disease matched that of the previous patient with 70% probability; in the *negative* condition that probability was 30%. Trials in the *control* condition were independently sampled (i.e., a 50% transition probability). For each subject, the diseases were first randomly generated according to the condition, and then the symptoms were randomly generated with probabilities dependent on disease outcomes.

Results

A logistic regression model was fitted to the data from each subject to determine the degree of influence of both present and past information on that subject's responses. The predictors in this model were the three symptoms, the disease of the previous patient ($Disease_{n-1}$), the similarity of the previous patient's symptom profile with that of the present patient (C^n_{n-1}, given as the number of matching minus the number of mismatching symptom dimensions), and the interaction $Disease_{n-1} \times C^n_{n-1}$.

The regression coefficient for $Disease_{n-1}$ was significantly higher for subjects in the Positive condition as compared to Control, indicating a heightened recency effect (Fig. 1). The reverse effect was not found for the Negative condition, despite its logical symmetry with the Positive condition.

In addition, the coefficient for $Disease_{n-1} \times C^n_{n-1}$ was significantly positive in all 3 conditions, showing that cue commonality moderates RE.

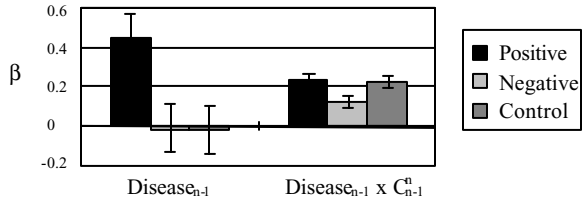


Figure 1: Subjects' use of past information, by condition

Discussion

Reliance on recent outcomes increased dramatically when those outcomes were positively related to the current case. No corresponding effect was found for negatively related recent outcomes. In fact, no overall RE was found in the negative and control conditions. These effects are not anticipated by the standard proposal that impact of outcomes simply decays exponentially with time. People appear to assess the predictive validity of recent events, but are nevertheless biased towards expecting a positive relationship.

References

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory & Language,* **27**, 166-195.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review,* **99**, 22-44.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* **18**, 211-233.

This paper

withdrawn

by the author

The Roles of Context and Working Memory in Probability Matching

Alexandra P. Kincannon (kincannon@virginia.edu)
Department of Psychology, University of Virginia
P.O. Box 400400, Charlottesville, VA 22904-4400 USA

Introduction

Probability matching, where a participant's choice frequency matches the probability of an alternative, is the modal response strategy in many probabilistic choice tasks. According to traditional norms, the probability matching strategy results in a sub-optimal payoff, compared to the utility maximizing strategy of always choosing the most probable alternative. Some researchers, however, have argued that probability matching is evolutionarily adaptive in certain environments (e.g., Gigerenzer, 1996) and recent evidence suggests that use of the matching strategy is sensitive to different kinds of feedback and incentives (Gallistel, 1990; Wolford, Newman, Cutler, & Miller, 2001).

Others have applied a dual-systems approach to explain the strategies that participants use in probabilistic choice and other tasks (Stanovich & West, 2000), finding that those who use a utility maximizing strategy have higher cognitive ability on average than those who use a probability matching strategy. This evidence supports the theory that the two strategies are products of two different reasoning processes; one that is rule-based and analytic and one that is based on evolutionarily derived heuristics.

The present experiment explores the roles of working memory and task context in probability matching. These factors are proposed to differentiate between the two reasoning processes. Analytic processing requires working memory resources; thus taxing these resources with a secondary task should reduce the use of the maximizing strategy in the probabilistic choice task. Heuristic processing requires a meaningful, socially relevant context; thus an enriched context should increase the use of the matching strategy in this task.

Method

A 2x2 between-participants design was used with two levels for each of the independent variables. Participants were assigned at random either to do the probabilistic choice task by itself or in parallel with a random number generation task (single vs. dual task condition). Half the participants saw a contextually sparse version of the choice task in which there were two blank squares on a computer screen and they had to guess which square would not change color. The other half saw a contextually enriched version in which there were two parking lots and they had to guess which lot would not be ticketed. Participants received feedback after each trial. The experiment was programmed to make the target event occur 75% of the time at the location on the left side of the screen.

Results and Discussion

The 150 trials were divided into five blocks of 30 and a difference from matching score was calculated for each block by subtracting the number of times the left side was chosen from the number of times the target event occurred on the left. There was a significant main effect for both the task condition and the context condition (see Figure 1.). In the first block, participants in all conditions chose either side with equal frequency. In the last block, participants in the single task/rich context condition were more likely to use a maximizing strategy, whereas participants in the other conditions were more likely to use a matching strategy.

These findings suggest that analytic processing resources are needed to use the maximizing strategy, as was predicted. Surprisingly, the enriched context facilitated a maximizing, rather than a matching strategy.

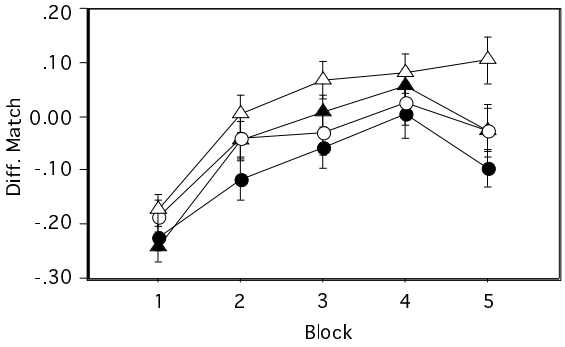


Figure 1: Single task condition and rich task context facilitate a utility maximizing strategy. Filled shapes = Sparse context, open shapes = Rich context; circles = Dual task, triangles = Single task.

References

Gallistel, C. R. (1990). *The Organization of Learning.* Cambridge, MA: The MIT Press.

Gigerenzer, G. (1996). Rationality: Why social context matters. In P. B. Baltes & U. M. Staudinger (Eds.), *Interactive minds: Life-span perspectives on the social foundation of cognition*. New York: Cambridge University Press.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645-726.

Wolford, G., Newman, S., Cutler, J. & Miller, M. (2001). *The effect of gain and loss on probability guessing*. Paper presented at the 42nd Annual Meeting of the Psychonomic Society, Orlando, FL.

The Structure of Linguistic Spatial Representation
A test for psychometric structure using Japanese spatial terms

Takatsugu KOJIMA (kojima@cpsy.mbox.media.kyoto-u.ac.jp)
Takashi KUSUMI (n50609@sakura.kudpc.kyoto-u.ac.jp)
Faculty of Education, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan

Introduction

Although it is not clear whether our system of spatial representation as a whole (integrated spatial representation) is structured initially by perception or by language, it is certain that forms of spatial representation generally are ultimately based largely on perception, especially on vision, and on language. Furthermore, when we encode or categorize forms of spatial representation, we do so following systematic rules that are founded on the structures of spatial representation formed by vision, language and so on.

The question remains, then, whether these spatial structures resemble each other? If so, are they grouped in any way? In addition, do they connect with each other, and, if they do, what is the nature of the relation between the specific spatial representational structures formed by vision and by language? According to Hayward and Tarr (1995), the spatial structure encoded by language (e.g., above, below) seems to be based on, and to correspond to, the spatial representational structure based on perception. Crawford, Regier and Huttenlocher (2000), however, have insisted that these structures do not correspond.

Both Hayward and Tarr (1995) and Crawford et al (2000) examined this issue from the viewpoint of categorical prototype and boundary. In linguistic spatial categorization, prototypes and boundaries are dependent on what spatial terms are used. Therefore, their method is not appropriate for comparing spatial structures.

In this study, instead of prototypes and boundaries, we examined fit patterns for four Japanese spatial terms (*ue, shita, hidari, migi*) using Thurstone's law of comparative judgment (case V). From this fit distribution and the prototypical spatial structure of visual representation (Huttenlocher , Hedges, and Duncan, 1991), we investigated whether the spatial structure of perceptual representation corresponds to that of linguistic representation.

Method

Ten Japanese graduate and undergraduate students participated. Stimuli were generated by an IBM/PC compatible computer and presented on a CRT at a viewing distance of approximately 115cm. For each trial, an instruction word would first appear in the center of the screen for 1000ms. Then, a black square (11°×11° side) was centered as a reference object, and two black dots (0.12°×0.12° diameter) were randomly presented as target objects, occupying 21 fixed locations that were based on a former experiment (Kojima & Kusumi, 2002). Four Chinese characters ( ), each of which expresses spatial locations (e.g.," (*ue*)" in Japanese is nearly equal to "above" in English, and, in the same way, " (*shita*)" to "below", " (*hidari*)" to "left" and " (*migi*)" to right) were used as the instruction words in a square (11°×11° side). The participants were required to compare the locations of the two dots in relation to the reference object, and to choose the dot that best suited the location expressed by the prior instruction word.

Results and Discussion

The paired comparison data were processed and scaled by Thurstone's law of comparative judgment (case V). The fit patterns of the four Japanese spatial terms are shown in Fig1, based on scaled value. Here, the width and depth are equivalent to the horizontal and vertical frame lines on the CRT. The height expresses the scaled value (from –3.00 to 3.00). This fit pattern differs from that of the prototypical structure; It shows a less simple gradient pattern.

From this pattern and former studies' result (Huttenlocher et al, 1991), we supported the suggestion of Crawford et.al (2000). That is, we also found that the structure of visual spatial representation does not correspond to the structure of linguistic spatial representation.

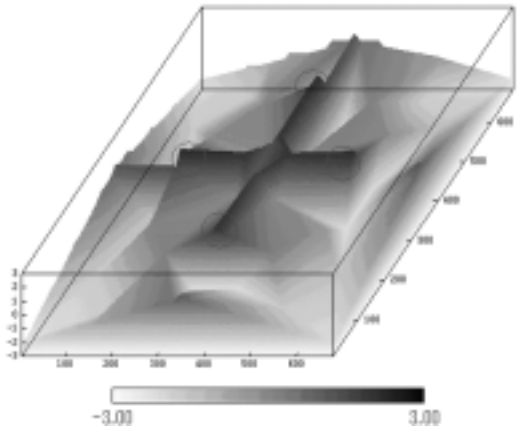


Figure 1:The fit pattern of four Japanese spatial terms

References

Crawford, L., Regier, T., & Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. *Cognition*,73(5),209-235.

Hayward, W., & Tarr, M. (1995). Spatial language and spatial representation. *Cognition*,55,39-84.

Huttenlocher, J., Hedges, L., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review*,98(3),352-376.

The Effect of Attentional Distraction in the Tempo-Naming Task

Laura Leach (lleach@gmu.edu) and **Christopher Kello (ckello@gmu.edu)**
Department of Psychology, George Mason University
4400 University Drive, Fairfax, VA 22030 USA

Errors in word naming can reveal dysfunctions in the component processes of word reading. An abundance of regularization errors (e.g., naming PINT to rhyme with MINT) suggests an over-emphasis on sub-lexical spelling-sound correspondences relative to lexical knowledge. An abundance of lexicalization errors (e.g., naming WIFE as WHITE) suggests a malfunction in the influence of lexical knowledge. An abundance of positional errors (e.g., naming BROAD as BOARD) suggests a malfunction in the positioning of letters and/or sounds.

Kello and Plaut (2000) introduced the tempo-naming task as a method for inducing naming errors under extreme pressure for speeded responding. The results of three experiments showed that lexicalization errors predominated compared with regularization errors. They interpreted this pattern as indicating that pressure for speed caused an increase in the emphasis on lexical knowledge in the process of converting print to sound. Emphasis on lexical knowledge can also be found under manipulations of strategic control (Herdman, 1992), and in the reading errors that define phonological dyslexia (Coltheart, 1996).

In the current study, we examined the effect of distraction of visual attention on performance in the tempo-naming task. Attentional distraction was hypothesized to interfere with the process of identifying and positioning the letters of a word stimulus. Errors were used as window into the effect of attentional distraction on processes of word reading.

Method

Participants

Twenty undergraduates participated in the experiment for course credit. All participants reported English as their native language, and all had normal or corrected vision.

Stimuli

The experiment consisted of 600 monosyllabic words, sampled from a full corpus of English words to preserve the distributional characteristics of the full corpus. Words varied in frequency and regularity. For each participant, each word was randomly assigned to one of six blocks such that there were 100 words per block.

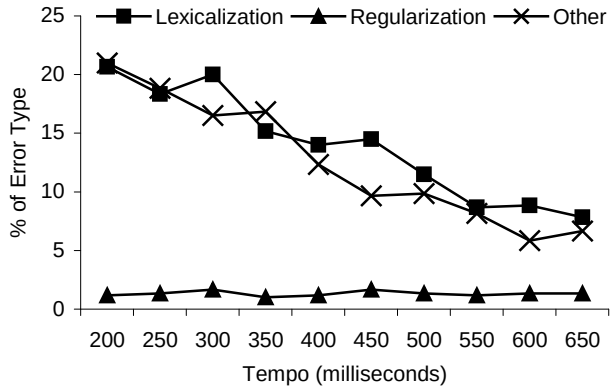
Procedure

On each trial, five beats of a simple audiovisual tempo were presented, and a printed word was displayed on the fifth beat. The tempo was varied between 200 ms and 650 ms (in steps of 50 ms), and each word was assigned to each tempo twice across participants. Participants were instructed to name the word in time with the sixth beat, and feedback on timing was given on every trial. Visual distraction was created by flashing white disks of varying size in random positions on the screen during stimulus presentation.

Results and Discussion

As tempo decreased (i.e., more pressure for speed), naming latencies and durations decreased, and overall error rates increased. These results replicated Kello and Plaut (2000).

The percentage of different types of errors is graphed as a function of tempo in the figure below. Lexicalizations were much more frequent compared with regularizations, and lexicalizations increased with faster tempos, but regularizations did not. The rate of lexicalizations was greater than that found in a comparable experiment (not reported here) in which there was no attentional distraction. Other error types included nonwords, stutters, and garbled pronunciations.



Results indicated that attentional distraction caused an increase in the emphasis on lexical knowledge. Future analyses are planned to examine the effect of attentional distraction on the rate of positional errors.

References

Coltheart, M. (Ed.). (1996). Special issue on Phonological Dyslexia. *Cognitive Neuropsychology, 13*, 749-934.

Herdman, C. M. (1992). Attentional resource demands of visual word recognition in naming and lexical decisions. *Journal of Experimental Psychlogy: Human Perception and Performance, 18*, 460-470.

Kello, C.T., & Plaut, D.C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 719-750.

Why Animated but not Static? The Spatial-Temporal

Terence C. P. Lee (faypui@graduate.hku.hk)
Albert W. L. Chau (awlchau@hku.hk)
Benise S. K. Mak (benise@hkusua.hku.hk)
Department of Psychology, The University of Hong Kong
Pokfulam Road, Hong Kong

Background

Past research has shown that pictures in general enhance learning with text. Glenberg & Langston (1992) found that pictures compatible with text facilitate the learning of a procedural concept with concurrent steps. Glenberg & Kruley (1992) evinced that pictures assist anaphor resolution, thereby improving comprehension.

One of the major theoretical constructs adopted to account for such facilitatory effect is the dual-code theory by Paivio (1971), which suggests that information could be stored either verbally or visually, and that these codes together lead to better retention than either one alone.

Owing to the advancement in information technology, multimedia instructional materials, such as computer-generated animation, have become popular. Compared with static pictures, animation facilitates learning only under more specific conditions. Rieber (1991) showed that students with animated presentation outperformed those with static presentation, but only when frames were presented in chunks. Schnotz, *et al.* (1999) showed that animation better assists learning than static pictures for individual learning, but not cooperative one, which leads to cognitive overload. Mayer and Sims (1994) demonstrated that high-spatial ability students benefit more from contiguous animation than their counterparts. Large, *et al.*(1996) found that animation improves the comprehension of a procedural text more than a descriptive one.

Hypotheses

The aforementioned findings converge to suggest that animation learning demands additional cognitive processing that consumes extra cognitive resources, compared with static-picture learning. In the light of 1) the advantageous position of high-spatial ability students, 2) the facilitatory effect upon learning sequential concepts, and 3) the essential disparity between animated and static pictures, it is postulated that the animation-over-static-picture advantage, when it occurs, is attributable to better spatial-temporal coding in the former condition.

Methodology and Findings

The design was a modified version from the study by Moreno and Mayer (1999). The participants were forty-four undergraduates enrolling in Introduction to Psychology at the University of Hong Kong. Participants in the animation-narration condition viewed an animation of 190 seconds on lightning formation, while participants in the static-picture-narration condition viewed 11 static pictures that were snapshots representing critical steps of lightning formation extracted from the animation. After the learning section, participants were tested with three tasks: 1) matching verbal labels with to-be-circled objects, 2) verbal recall of narration and 3) sorting the sequence of 11 pictures, which were the pictures shown in the static condition. The last measure, which had not been adopted in previous studies, was developed to assess spatial-temporal coding.

One-tailed independent-samples t tests showed that the animation-narration group outperformed static-picture-narration group, on the matching task ($t(42) = 2.630$, $p = 0.006$), the verbal-retention task ($t(42) = 3.077$, $p = 0.02$) and the visual-spatial-retention task ($t(42) = 1.895$, $p = 0.0325$).

Discussion

Results showed that animation plus narration is a better combination than static pictures plus narration in facilitating learning. The former led to stronger verbal and visual representational connections, and closer referential connections between the two modules, all of which are beneficial to the learning of a sequential concept. In addition, animation was found to be superior to static graphics in assisting spatial-temporal coding. These findings are consistent with a modified version of Paivio's dual-code theory.

References

Glenberg, A. M., & Kruley, P. (1992). Pictures and anaphora: Evidence for independent processes. *Memory & Cognition, 20*(5), 461-471.

Glenberg, A., & Langston, W. E. (1992). Comprehension of illustrated test: Pictures help to build mental models. *Journal of Memory and Language, 31*, 129-151.

Large, A., Beheshti, J., Breuleux, A. & Renaud, A. (1996). The effect of animation in enhancing descriptive and procedural texts in a multimedia learning environment. *Journal of American Society for Information Science, 47(6),* 437-448.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual coding theory of multimedia learning. *Journal of Educational Psychology, 86*, 389-401.

Moreno, R., & Mayer, R.E. (1999). Cognitive principle of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 358-368.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Reinhart & Winston.

Rieber, L. P. (1991). Effects of visual grouping strategies of computer-animated presentations on selective attention in science. *Educational Technology Research and Development, 39(4),* 5-15.

Schnotz, W., Böckheler, J., & Grzondziel, H. (1999). Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education, XIV*, 245-265.

Domain Knowledge and False Memory

Yuh-shiow Lee (psyysl@ccu.edu.tw)
Han-yu Lin (hanyu@ns1.mit.edu.tw)
Department of Psychology, National Chung-Cheng University
Chiayi, 621, Taiwan, ROC

Introduction

Many studies have demonstrated the power of schema or knowledge structure in organizing incoming information, which led to various kinds of memory errors. What a person already knew determined whether and how information would be remembered. Thus, it is clear that there is a close link between cognitive processes and structures and the types of memory errors committed. Based on this logic, the present study examined how participants' prior knowledge affects memory errors. This study used the DRM paradigm to examine false memory produced by a group of industrial design experts as compared to the control group.

In the DRM paradigm, participants study list of semantic related words (bed, awake, rest…) that are all related to a critical word that is not presented (sleep). High levels of false memory for lures (e.g. sleep) have been demonstrated in tests such as free recall and recognition (e.g., Roediger & McDermott, 1995). Studying expert behaviors offers a unique window into human cognition. While experts' behaviors on chess were the most researched area, a wide range of domains has been examined and various methods have been used in this area. The clearest finding from these studies was that memory performance on meaningful stimuli has found to be correlated with domain expertise. However, very few studies in this area have focused on the pattern of errors as affected by domain expertise.

McEvoy, Nelson, and Komatsu (1999) looked at the influence of preexisting knowledge on the production of false memory. They found that the probability of producing false memories in free recall varied with the strength of connections from the list words to the critical word and the density of the interconnections among the list words. In addition, false recognition was more likely when the list words were more densely interconnected. McEvoy et al. (1999) determined the association strength based on the word association norm and examined how the strength relates to false memory. Since the association strength between words develops though experiences, it is reasonable to assume that for the words that come from specific domain knowledge, experts of that domain and novices would have different types and strength of association. This would lead to different types and amount of false memory. This study investigated whether domain specific knowledge would induce or reduce false memories. A group of experts and novices were tested on words either related or unrelated to their knowledge of expertise.

Results and Discussion

Two types of semantically related list items were used in this study. One type of items were words selected from technical terms used in the domain of industrial design whereas the other type were common words. Four groups of participants were recruited: senior and junior non-industrial design students and industrial design students. Results showed that for the design-related lists, senior industrial design students not only performed better on both recall and recognition, but also had a higher rate of false recognition. In addition, age had an effect on the false recognition of common words, while the rate of false recognition of design-related terms was mainly determined by participants' domain knowledge. These results support the view that domain knowledge plays an important role in creating false memory.

Table. Mean Percentages of Correct Recall, Recognition and False Recognition as a Function of Age and Domain Knowledge

	correct recall		*correct recognition*		*false recognition*	
	common words	design-related words	common words	design-related words	common words	design-related words
Senior						
non-experts	.71	.49	.95	.86	.50	.29
experts	.73	.69	.94	.92	.57	.44
Junior						
non-experts	.72	.52	94	.86	.35	.28
experts	.72	.52	.93	.89	.38	.28

Acknowledgments

This research was supported by the National Science Council of R.O.C. Grant NSC-90-2413-H-194-015.

References

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What is the connection between true and false memories? The differential roles of interitem associations in recall and recognition. *Journal of experimental psychology: Learning, memory and cognition, 25*, 1177-1194.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803-814.

Acquisition of Landmark Knowledge from Static and Dynamic Presentation of Route Maps

Paul U. Lee (pauly@psych.stanford.edu)
Department of Psychology, Bldg. 420
Stanford, CA 94305 USA

Heike Tappe & Alexander Klippel ([tappe, klippel]@informatik.uni-hamburg.de)
Department for Informatics, University of Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany

Introduction

Route maps have recently gained much attention as effective tools to convey route information. Benefits of maps are attributed to their ubiquitous existence in culture and their analogous properties representing spatial knowledge. Route maps have become widely available through the internet and within on-board navigation systems. Despite their prevalence, optimal design criteria are still missing.

For example, route maps integrated in on-board navigation systems present routes dynamically with a moving dot that traverses a map. In contrast, internet maps present information statically with lines representing the route. At first glance, dynamic, animated presentation seems to be more effective than static one. (e.g. Nathan, Kintsch, & Young, 1992). However, advantages of animation may be due to other factors, such as interactivity or inclusion of information not present in static conditions. Furthermore, other studies fail to demonstrate superiority of animations at all (e.g. Morrison, 2000).

Extending research on effects of static vs. dynamic route presentation on conceptualization and memory (Klippel, Tappe, Habel, submitted), we examined the influence of presentation mode on memory for landmarks.

Dynamic vs. Static Presentation of Maps

Material and Procedure

Participants learned a route from a map of a fictitious town. The route was presented to them either as a solid line (i.e. static), a moving dot (dynamic), or a dot superimposed on a line (mixed).

The participants viewed the map three times, each for 1.5 minutes. Afterwards, they were given a blank map with only the streets and were asked to recall the landmarks.

Recall Memory of Landmarks

In the dynamic condition, landmarks at turning and non-turning intersections were recalled equally well (49.4% vs. 48.8%), but in the static condition landmarks at turning intersections were remembered more often (52.9%) than at non-turning intersections (43.8%) (see Table 1). Since landmarks at turns are more critical to route directions, we conclude that static displays of route information is preferable over dynamic displays.

	Turns	Non-turns	Total
Dynamic	**49.4**	**48.8**	49.1
Static	**52.9**	**43.8**	48.4
Mixed	**57.7**	**41.5**	49.6

Table 1: Proportion of recalled landmarks (in %)

Surprisingly, in the mixed condition participants recalled even more landmarks at turns (57.7%) than at non-turns (41.5%). The combination of different presentation modes and the resulting memory improvement for vital route information support findings on the benefits of redundant information displays (Hirtle, 1999).

Acknowledgments

This research was supported by DAAD PKZ A-01-49336 to the first author and by the Deutsche Forschungsgemeinschaft (DFG) HA 1237-10, (Conceptualization processes in language production), and FR, 806-8 (Aspect maps) to the second and the third author.

References

Hirtle, S. C. (1999). The use of maps, images, and "gestures" for navigation (pp. 31-40). In C. Freksa, W. Brauer, C. Habel, K.F. Wender (eds.). Spatial cognition II, integrating abstract theories, empirical studies, formal methods, and practical applications. Springer: Berlin.

Klippel, A., Tappe, H., & Habel, C. (submitted). Pictorial representation of routes. Chunking route segments during comprehension.

Morrison, J. B. (2000). Does animation facilitate learning? An evaluation of the congruence and equivalence hypothesis. Ph.D. Thesis, Stanford University.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. Cognition and Instruction, 9, 329-389.

Bongard Problems and Symbolic Approaches: A Skeptical Look

Alexandre Linhares (linhares@fgv.br)
EBAPE/FGV, Praia de Botafogo 190
Rio de Janeiro 22257-970, Brazil

Introduction to Bongard Problems

Three decades ago the intelligence theorist Mikhail Bongard (1970) posed an outstanding challenge to artificial intelligence, bringing a remarkable set of 100 visual pattern understanding problems where two classes of figures are presented and the pattern recognizer (either a human or a machine) is asked to identify the conceptual distinction between them. Sometimes the classes are opposite in terms of this conceptual distinction, such as large figures versus small figures, and other times there may be properties or relations holding between boxes in one class, but not in the other, such that there is always some aspect to distinguish the classes.

Figure 1 displays two very simple Bongard problems. One of the most important characteristics of such problems is that, although humans can generally solve them intuitively, their automation is simply daunting: there is always much relevant information to be perceived and much irrelevant information to be discarded.

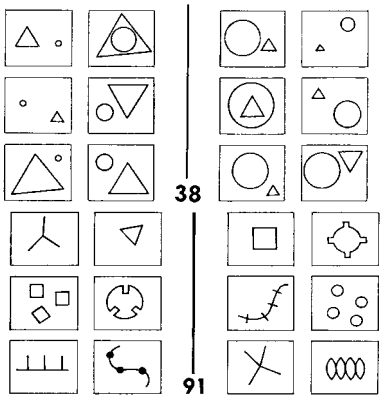


Figure 1: Bongard problems BP#91 and BP#38. What abstract aspect distinguishes the boxes on the right (class 1) from the boxes on the left (class 2)? [From M.M. Bongard (1970) *Pattern Recognition*, Spartan books]

A philosophical problem

Let us focus on the following issue: suppose that a specific Bongard problem includes the following box shown in figure 2. Would it be appropriate (as done in Saito and Nakano 1994; Saito and Nakano 1995) to discard the raw geometrical information in favor of a simple symbolic description, such as that presented?

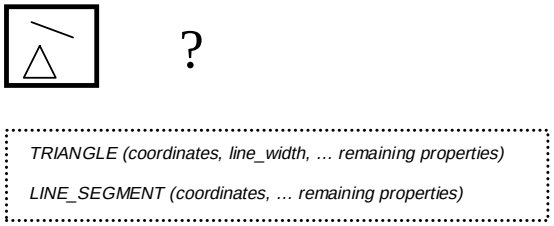


Figure 2: Raw geometrical information versus symbolic descriptions.

This is the core question of our investigation. We argue that such approach is unattainable, as it leads to the inadequate philosophical grounds of *metaphysical realism*. Purely symbolic representations are not capable of containing all forms of concepts and categories expressed (and expressible) in Bongard problems. Furthermore, they lead to inadequate architectural models, which can easily be seen to falter (Linhares, 2000; see also Hofstadter 1979, Hofstadter 1995a, Hofstadter 1995b). Finally, we propose that the philosophical grounds sketched in Smith (1984) are sound alternatives to current theory.

References

Bongard, M.M. (1970) *Pattern Recognition*, Spartan Books, New York.

Hofstadter, D.(1979) Gödel, Escher, Bach: an Eternal Golden Braid, Basic Books, New York.

Hofstadter, D. (1995a) *Fluid Concepts and Creative Analogies*. New York: Basic Books.

Hofstadter, D. (1995b) On seeing A's and seeing As, *Stanford Humanities Review*, 4, 109-121.

Linhares, A. (2000). A glimpse at the metaphysics of Bongard Problems, *Artificial Intelligence*,121, 251-270.

Saito, K. and R. Nakano (1995), A concept learning algorithm with adaptive search, in: K. Furukawa, D. Michie, and S. Muggleton, *Machine Intelligence 14 – Applied Machine Intelligence*, pp. 347-363, Oxford: Oxford University Press.

Saito, K., and R. Nakano (1994) Adaptive concept learning algorithm, *IFIP Transactions A – Computer Science and Technology*, 51, 294-299.

Smith, B.C. (1996) *On the origin of objects*, MIT Press, London.

Language-Like Representation in Embodied and Situated Cognition: A Case Study of a Situated Robot's Planning

Hsi-wen Liu (hwliu@pu.edu.tw)
Division of Humanities, Providence University
200 Chung-Chi Rd, Shalu, Taichung County 433, Taiwan

Interests in embodiedness and situatedness have increased in all disciplines of cognitive science (Clark, 1999). Such interests generally concern the complex interplay between robotic (or neural) systems and their local environments. A number of theorists and philosophers with such interests cast strong doubt on the need of a rule-based search and even the need of overall *internal* representations (Brooks, 1991; Beer, 1995; Keijzer, 1998).

Also with those interests, Andy Clark and some co-authors raise significant reasons for reconsideration regarding the above doubt (Clark, 1997ab, 1999; Clark and Grush, 1999; and others). The gist in general is that in the above systems internal representations can exist, by playing a two-fold role. On the one hand, the internal representations in the systems constitute descriptions of the world, and on the other, these representations serve as *internal* representations/codes to *control* those systems' tight coupling agent-environment interactions. With such a two-fold role, those internal representations are named *action-oriented representations* (Clark, 1997ab, 1999). This two-fold role has been exemplified by a number of robotic architectures with on-line control of situated activities (e.g., see Mataric's TOTO discussed in Clark (1997ab, 1999)). A question with the above theory of action-oriented representations is whether some of those representations can be characterised in language-like codes. If the answer is yes, it would constitute a link between the situated-and-embodied approach and traditional cognitive science.

Conforming to the above theory of action-oriented representations, this work contends that the answer is indeed yes. That is, language-like codes can play the above two-fold role in the control of tight coupling agent-environment interactions. This claim is exemplified by the robot arm designed by Maes (1990).

The robot activities, as Maes (1990) sees, lead (though not in terms of abstract thoughts) to the *planning* of various actions in support of a goal—catching an appropriate tool to paste sand on a board. The architecture of that robot arm is hybrid. It consists of several *action modules*, which perform certain motor actions. The architecture of each action module includes *both* symbolic codes (i.e. language-like codes) which describe environmental conditions (or arm states) *and* the control over the (numerical) energy flowing across those action modules. When the robot system observes, or initiates an action, those modules change their inherent energy and their condition lists of the environment (or the arm states). An action module initiates its action, when the energy flowing in the module goes beyond a certain threshold. Yet, the energy flowing in a certain module may go below its threshold and hence need a fine-tuning of the energy flowing among certain relevant modules, and even a fine-tuning of the threshold itself.

The present work argues that the language-like codes in the architecture of Maes' (1990) robot arm play the aforementioned two-fold role of action-oriented representations. The steps of argument are as follows.

1. The architecture of the robot arm clearly adopts certain language-like codes, which describe certain conditions.
2. There is a limited degree of searching among those language-like codes, which serve as initiating conditions of the action to be carried out by an action module.
3. The activities of the robot arm are situated, because of the tight coupling agent (arm)-environment interactions.
4. The computation for the control of the above interactions is embodied, because the energy spreading among the relevant action modules (and even a threshold itself) is sensitively fine-tuned in response to the recurrent re-try of initiating a single module's action.

As the above argument shows, the language-like codes in the architecture of Maes' (1990) robot arm play the two-fold role of action-oriented representations, in support of the robot's situated and embodied activities. This constitutes a link between situated-and-embodied approach and traditional cognitive science.

Acknowledgments

This work is supported by the National Science Council, Taiwan (90-2411-H-126-017).

References

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72*, 173-215.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139-160.

Clark, A. (1997a). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A. (1997b) The dynamical challenge. *Cognitive Science*, *21*, 461-481.

Clark, A. (1999) An embodied cognitive science? *Trends in Cognitive Science*, *3*(9), 345-351.

Clark, A., and Grush, R. (1999) Toward a cognitive robotics. *Adaptive Behavior*, *7*, 5-16.

Keijzer, F. (1998). Doing without representations which specify what to do. *Philosophical Psychology*, *11*, 269-302.

Maes, P. (1990). Situated agents can have goals. In Maes, P. (Ed.), *Designing autonomous agents*. MIT Press.

The Comprehension of Novel Noun-Noun Compounds: The Influence of Out-of-Context Interpretations on In-Context Understanding

Dermot Lynott (dermot.lynott@ucd.ie) and Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland.

Introduction

A key question for language research, and in particular conceptual combination, is the dependence of in-context understanding on out-of-context meanings. Gerrig & Bortfeld (1999) contrast two views of conceptual combination comprehension in context, the interdependence and independence views. The *interdependence view* states that out-of-context meanings influence the in-context comprehension of novel combinations while the *independence view* (adopted by Gerrig and Bortfeld) maintains that context is the prevailing factor and prevents the activation of interpretations that might normally be available out of context. We test this hypothesis by generating a set of compounds whose interpretations differ in their frequency of production out of context and then varying the contexts in which the high-frequency or low-frequency interpretations are embedded. We can then establish whether these out-of-context interpretations have a bearing on in-context processing.

Experiment 1

We collected participants' out-of-context interpretations for novel noun-noun compounds and categorised them by their frequency of production. High-frequency (HF) and low frequency (LF) interpretations for each compound were selected from these frequency-scored sets. For example, for the compound *rhinoceros horse*, the HF interpretation was "a horse that has a horn" while the LF interpretation was "a horse that has tough skin". To confirm a difference between the HF and LF interpretations a response time experiment was run. The difference between high and low frequency interpretations was reliable, $F1(1, 20) = 5.845$, $p = 0.0253$.

Experiment 2

We define 3 context types - neutral, supportive and alternative. Supportive and alternative contexts make explicit reference to the relation between the head and modifier, while the neutral context makes no mention of the relation. The Supportive Context is defined as the condition where the paraphrase judgement question at the end of the story supports the interpretation suggested by the story. By contrast, the Alternative Context is the condition where the paraphrase judgement question supports an alternative question to the story i.e. if the story supports a HF interpretation then the question that follows will refer to the LF interpretation.

If out-of-context interpretations do not effect in-context processing then we would expect no difference between the HF and LF conditions. If there is an influence then a difference in response time should be evident. This should be clearest in the Alternative condition where people move from one interpretation to another. If the independence view holds then it should take the same amount of time to go from HF to LF as it does to go from LF to HF, since their frequency of production out of context should not impact on processing time. This, however was not the case. We found a reliable difference between the high and low frequency interpretations, $F1(1, 40) = 12.933$, $p < 0.001$, and also a reliable trend (using Page's L) showing that the supportive context was responded to most quickly followed by the neutral and then the alternative $L(12) = 158.5$, $p < 0.005$**.** The differences between the contexts is shown in Figure 1.

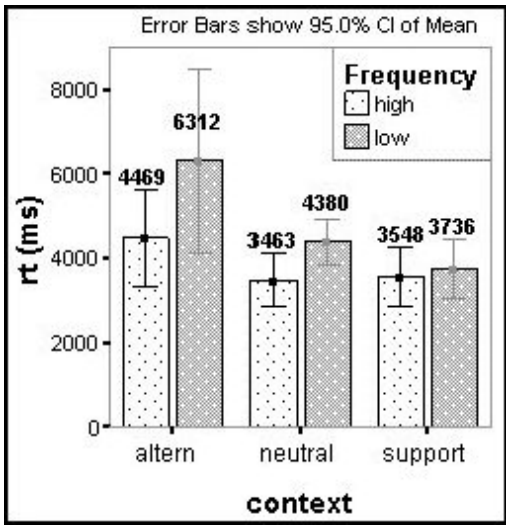


Figure 1: Mean RTs for HF / LF interpretations per context.

Discussion

We have shown here that in certain context types (alternative and neutral) out-of-context interpretations to have an effect on in-context processing. This shows that one aspect of out-of-context interpretations, namely their frequency of production, has an impact on the ease with which interpretations are comprehended in-context, which violates the basic assumption of the independence view.

References

Gerrig, R. J. & Bortfeld, H. (1999). Sense creation in and out of discourse contexts. *Journal of memory and Language*, 41, 457-468.

Allocation of Attention in Neural Network Models of Categorization

Toshihiko Matsuka (tm249@columbia.edu)

James E. Corter (jec34@columbia.edu)

Department of Human Development, Teachers College, Columbia University, 525 W. 120th St., New York, NY 10027 USA

Arthur B. Markman (markman@psy.utexas.edu)

Department of Psychology, University of Texas, Austin, TX 78712 USA

We compared ALCOVE (Kruschke, 1992), RASHNL (Kruschke & Johansen, 1999), SUSTAIN (Love & Medin, 1998), and the Cortico-Hippocampal Model (CHM) (Gluck & Myers, 1993) to see how they account for selective attention in category learning. Such comparisons may usefully augment comparisons of the models' classification accuracy.

Method

We simulated the results of studies of classification learning by Medin and Schaffer (1978) and Medin, Altom, Edelson & Freko (1982). The parameter values used for each model were adjusted to minimize the SSE in reproducing the training classification responses by human subjects.

Attention allocation predictions for the models were derived as follows. ALCOVE and RASHNL have explicit attention weight parameters, which are reported below. For SUSTAIN, the dimension-specific tuning parameters, λ, are reported. In the CHM there are no explicitly defined dimension attention parameters. We defined implicit measures of a dimension's attentional salience, by summing the absolute values of weights from all input nodes associated with a given dimension to the hidden node layer in the hippocampal net component of the CHM. To enhance comparability among the models, we computed and report relative attention weights for all the models.

Summary of Results

For Experiment 2 of Medin and Schaffer (1978), all the models fit the training set classification probabilities roughly equally well, but RASHNL and SUSTAIN were somewhat more accurate in predicting classification responses for the transfer stimuli. In this stimulus structure Dimensions 1 and 3 are highly predictive of the binary classification task, and Dimension 4 is moderately predictive. Somewhat surprisingly, ALCOVE, RASHNL, and the CHM gave as much or more attention weight to Dimension 4 as to the more diagnostic dimensions.

For Experiment 4 of Medin, Altom, Edelson & Freko (1982), RASHNL and the CHM fit the training set classification probabilities best, but RASHNL was the best and the CHM worst in predicting the transfer classifications. In this stimulus structure, Dimensions 1 and 2 are diagnostic in the sense that each is highly correlated with the criterion classification response, but Dimensions 3 and 4 have a simple XOR pattern in regards to the criterion classification. ALCOVE, RASHNL, and SUSTAIN all learn to allocate more attention to Dimensions 3 and 4, that together define the classification in terms of a simple XOR relationship. In contrast, the CHM pays more attention to the individually, but merely probabilistically, diagnostic Dimensions 1 and 2.

Conclusions

The four models give different predictions about attention weights for some stimulus structures. Examining and comparing these predictions may shed light on how the models learn. A promising line for future research is to gather direct data on how humans allocate attention in category learning (Matsuka, 2002).

References

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491-516.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 1083-1119.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceeding of the Fifteenth National Conference on AI (AAAI-98),* 671-676.

Matsuka, T. (2002). Attention processes in category learning. Unpublished doctoral dissertation (draft), Teachers College, Columbia University.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8,* 37-50.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238

How Goals Affect Evaluations of Animation Effectiveness

Julie Bauer Morrison (morrison@bryant.edu)
Department of Applied Psychology / Bryant College
1150 Douglas Pike / Smithfield, RI 02917-1284

Introduction

Animation commands our attention, it entertains us, but does it instruct us? Recent studies have shown animation is no more effective in instruction than comparable text or static graphics (for a review, see Tversky, Morrison, & Betrancourt, in press). This is particularly surprising because these studies often focus on teaching the kind of information that ought to be animation's strength, namely, change in time.

Despite the research to the contrary, the perception of animation is that it is an effective means of presenting information, specifically, information regarding movement. Why do these perceptions of animation differ from what research tells us of its effectiveness? The present research shows that the goals one has when evaluating different instructional media affect those evaluations, such that when we must learn from animation we judge effectiveness by our perception of what we have learned, whereas when we are simply evaluating animation we judge on aesthetics.

Media Comparisons

Method

Participants reviewed three learning interfaces, text, text plus static graphics, and text plus animated graphics, each displaying rules of movement through an environment. Thirty-one participants were under instructions to imagine they would be subsequently tested on the information (No Learning group), while 55 were to be tested (Learning group). Following the entire review process, participants rated each interface on three criteria using a 1-7 scale: how *effective* they thought the interface would be/was in helping them learn the material, how *confident* they would be/were about subsequent tests of the material, and how *enjoyable* it would be/was learning from the interface.

Results

For each rating criteria, effectiveness, confidence, and enjoyment, those in the No Learning group rated the interface with animated graphics the highest, followed by the ratings for the static graphics interface and the text interface (Effectiveness: $F(2,70)=16.1$, $p<.01$, Confidence: $F(2,70)=15.5$, $p<.01$, Enjoyment: $F(2,70)= 21.6$, $p<.01$). All paired-sample t-tests showing the differences between the three interface types were significant at the $p<.001$ level (see Figure 1).

Participants in the Learning group showed a different pattern of results in which the ratings in the graphics conditions were indistinguishable. Despite there being overall differences for each rating across the three media

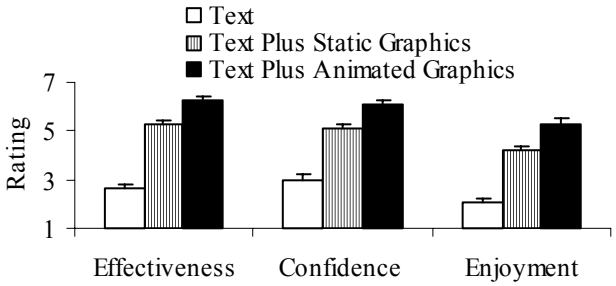


Figure 1: No Learning group ratings for effectiveness, confidence, and enjoyment.

(Effectiveness: $F(3,51)=15.1$, $p<.01$, Confidence: $F(3,51)= 4.0$, $p<.05$, Enjoyment: $F(3,51)= 3.2$, $p<.05$), the only t-test to reach significance was for the effectiveness rating comparing text and animated graphics (see Figure 2).

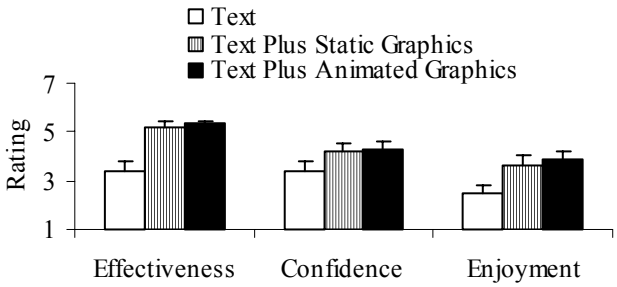


Figure 2: Learning group ratings for effectiveness, confidence, and enjoyment.

Subsequent analyses showed that the pattern of ratings for the Learning group mirrored actual performance. In other words, participants based their ratings on accurate judgments of learning, rather than on aesthetic elements of the interface.

Discussion

Judgments of different instructional media, in terms of their effectiveness, ability to inspire confidence, and enjoyability, differed based on the learner's goals. The attractiveness of animation, and, secondly, static graphics, influenced judgments of those who were not required to learn the information. Those who were required to learn made judgments that superceded the superficial aspects of the interface and focused instead on accurate perceptions of what had been learned.

References

Tversky, B., Morrison, J.B., & Betrancourt, M. (in press). Animation: Does it facilitate? *International Journal of Human-Computer Studies*.

Cognitive Principles in a Computational Engineering Design Methodology

Jarrod Moss (jarrodm@cmu.edu)
Kenneth Kotovsky (kotovsky@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213 USA

Jonathan Cagan (cagan@cmu.edu)
Department of Mechanical Engineering, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213 USA

A-Design

A-Design is a multi-agent computational system that automates the conceptual design process and is currently capable of solving several electromechanical design problems (Campbell, Cagan, & Kotovsky, 1999). A-Design takes input and output constraints for a desired electromechanical device and produces an array of conceptual designs that satisfy the problem constraints.

In order to solve a design problem, the system initially generates a population of candidate designs. These candidates are then evaluated along multiple dimensions, and all but the best candidates are eliminated before another iteration of design generation begins. In each iteration of the design process, the best designs from the previous iteration are modified and a number of new designs are also generated. This iterative design process is similar to the iterative design process employed by engineers solving design problems (Smith & Tjandra, 1998).

While A-Design's development was guided by some aspects of human cognition, it is not intended as a cognitive model of the design process. However, it does provide a point of departure for an investigation of the cognitive processes occurring in the field of engineering design. As an initial step in this investigation, modifications were made to A-Design in order to allow it to learn design knowledge during problem solving. A-Design was then tested to see if this learned knowledge could transfer to new design problems.

Learning from Design Experience

A-Design already had the ability to examine a group of designs and extract common subsets of electromechanical components that appear in every design of the group (Campbell, 2001). A set of interconnected components that appears in multiple designs will be referred to as a common subsystem of those designs. In this research, A-Design was said to have completed a design problem after it had run for a specified number of design iterations on that problem. Useful design knowledge was extracted after a problem had been solved by examining the best six designs produced in the final design iteration. Common subsystems were extracted from these designs, and these subsystems were added into a permanent memory store. Subsystems were indexed in this memory by the input and output constraints of the subsystem. A-Design could then add these subsystems to designs that it generated while solving new design problems.

The declarative memory component of ACT-R (Anderson & Lebiere, 1998) was utilized to store these subsystems. The ACT-R model of memory provides the capability to retrieve a chunk based on information in any of its slots, which gives A-Design the capability of retrieving a subsystem based on only a subset of the information contained in the input and output constraints of the subsystem. The ACT-R system may also provide the basis for an expanded version of A-Design which will attempt to capture some of the cognitive processes underlying design.

Results

A-Design was tested on a number of design problems to see if knowledge learned in one problem could be transferred successfully both within and across problems. Results indicate that A-Design applies learned knowledge very successfully in the same design problem where the knowledge was learned, however there is very little successful knowledge transfer across problems. This lack of transfer highlights some aspects of representation and knowledge transfer that A-Design does not have but which human designers obviously do.

Acknowledgments

This work was supported by a National Defense Science and Engineering Graduate Fellowship.

References

Anderson, J. R., & Lebierre, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Campbell, M., Cagan, J., & Kotovsky, K. (1999). A-design: An agent-based approach to conceptual design in a dynamic environment. *Research in Engineering Design, 11*, 172-192.

Campbell, M., Cagan, J. & Kotovsky, K. (2001). Learning From Design Experience: Todo/Taboo Guidance. *Proceedings of the 2001 ASME Design Engineering Technical Conferences and Computers in Engineering Conference: Design Theory and Methodology Conference*, DETC01/DTM-21687, Pittsburgh, PA.

Smith, R., Tjandra, P. (1998). Experimental observation of iteration in engineering design. *Research in Engineering Design, 10*, 107-117.

The Role of Exploration and Forward Checking in Human Scheduling

Stefani Nellen (Stefani.Nellen@urz.uni-heidelberg.de)
Joachim Funke (Joachim.Funke@psychologie.uni-heidelberg.de)
Department of Psychology, University of Heidelberg
Hauptstr. 47-51, 69117 Heidelberg, Germany

Research Objective: Connecting two strategies

We investigated how human participants work with the interactive Plan-A-Day (PAD) task (Funke & Krüger, 1993), which implements the task of scheduling numerous appointments during a fictitious day. We postulate that two strategies work together to enhance scheduling performance. The first strategy, remindful of instance based learning (cf. Logan, 1988) is to *explore* the feasibility of specific partial schedules by entering them into the PAD Interface. The second strategy, remindful of *forward checking* for Constraint Satisfaction Search (e.g. Russell & Norvig, 1995), checks in advance whether meeting an appointment would render another appointment impossible. In order to "verify" the results of forward checking, a certain amount of exploration is necessary, and in order to restrain exploration, forward checking is necessary. We conducted a study to determine which patterns of exploration are present in human scheduling and to validate our assumption that forward checking increases between two different PAD tasks.

Empirical results

The results reported in this section were obtained by presenting 43 student participants with two different PAD tasks (PAD 4 and PAD 5; a more detailed account of the analytic procedure can be found in Nellen, 2002).

Patterns of Exploration

The number of times participants modify their schedules during a PAD session is positively correlated with the number of complete restarts ("R"; abandoning a schedule completely and placing another appointment at the start), the number of different appointments placed at the start of a schedule ("Dif".), and negatively with the mean length ("ML") of the tried schedules. This pattern is consistent with exploration aimed at collecting a wide variety of experiences.

Table 1: Correlations between the number of schedule modifications and other process measures (explained in the text). Asterisks indicate significance at the level of $p<.01$ according to Fisher's Z test for correlations.

(N=43)	# of modifications PAD 4	# of Modifications PAD 5
R	-.57***	-.461***
Dif	.74***	.665***
ML	.61***	.556***

Increase in forward checking

The amount of forward checking in the data was assessed by computing the percentage of "deliberate" modifications that are performed *before* participants are too late at an appointment, relative to the total number of modifications. Table 2 shows the considerable increase of forward checking between the two PAD tasks.

Table 2: Increase of the percentage of deliberate modifications (forward checking) between the two PAD tasks.

	Deliberate modifications	
	PAD 4	PAD 5
average	41.7 %	59.6 %
median	44.0 %	58.0 %
mode	0.0 %	100 %

Conclusion

Participants consistently explore the feasibility of partial schedules. However, they also acquire the skill of forward checking between two PAD tasks, resulting in an enhanced quality of the exploration, which now yields fewer dead ends. The quick and considerable increase of forward checking suggests a mechanism of skill acquisition as production composition as defined by Anderson (1987); while the continuous presence of exploration implies that the importance of specific experiences throughout the scheduling process.

References

Anderson (1987). Skill acquisition: Compilation of weak method problem solutions. *Psychological Review, 94,* 192-210.

Funke, J. & Krüger, T. (1993). *"Plan A Day" (PAD): Ein Diagnostikum zur Erfassung von Planungskompetenz. Manual zum Programm (unveröffentl. Manuskript).* Bonn: Psychologisches Institut der Universität Bonn.

Logan, G.D. (1988). Towards an instance theory of automatization. *Psychological Review, 22*, 1-35.

Nellen, S. (2002). *How humans solve scheduling Problems.* Heidelberg: University Of Heidelberg (Diploma thesis)

Russell, S. & Norvig, P. (1995). *Artificial Intelligence: A modern approach*. NJ: Prentice Hall.

Cognitive Functional Processing System: Reasoning about Quantitative Relationships

Kent L. Norman (kent_norman@lap.umd.edu)
Department of Psychology, University of Maryland
College Park, MD 20742-4411 USA

Introduction

How people acquire or generate rules for combining information for decision-making has long been of interest (Norman, 1974). A cognitive functional processing system is proposed which assesses the relationship between any two variables in the environment by taking the partial derivative of an assumed or known response surface for one variable (A) relative to another (B). The response surface is a general concept that can include stochastic and determinist functions whether correlational or causal. The system is used to reason about the composition and decomposition of functions for combining information to make a single judgment. Brunswik's (1955) Lens Model, Norman Anderson's (1981), Information Integration Theory, and Kenneth Hammond's (1975) Social Judgment Theory are instances of such functions. In this system, decision makers store gradient functions in a matrix as shown in Table 1.

Table 1: Gradient Functions.

	A	**B**	**C**
A	--	f(A,B)	f(A,C)
B	f(B,A)	--	f(C,A)
C	f(C,A)	f(C,B)	--

f(A,B) = + when increases in B lead to increases in A.
f(A,B) = 0 when changes in B do not affect changes in A.
f(A,B) = - when increases in B lead to decreases in A.

When called upon to make subsequent judgments, decision makers generate composition rules from these gradients.

An experiment in which participants used a dynamic map query system to access information about geographic regions was used to investigate relationships learned and inferred between variables.

Experiment

Thirty undergraduates participated in an experiment in which they were asked to find states that satisfied one, two, and three variable range queries using a dynamic map query system (Dang, North, & Shneiderman, 2001). They adjusted sliders which continuously showed the dynamic set of states that met the search criteria.

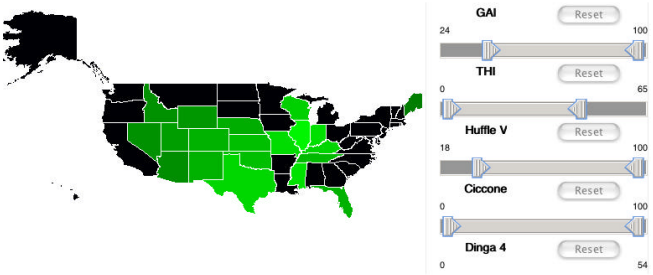


Figure 1: Dynamic Query Interface.

The results showed that participants were able to code the existence and direction of the gradients between the criterion variables; that is, they were aware of the inter-relationships of the search variables. A follow-up experiment was used to demonstrate the ability of the participants to generate new composition rules given the decomposed gradients they had stored from the queries.

Conclusion

The cognitive function processing systems accounts for how people are able to generate information integration rules in novel situations. They re-use previously stored gradients from other decision or search functions.

Acknowledgments

This work as funded in part from a grant from the U.S. Census Bureau, Statistical Research Division, Grant 50YABC166008.

References

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193-217.

Dang, G., North, C., Shneiderman, B. (2001). Dynamic queries and brushing on choropleth maps. *Proc. International Conference on Information Visualization 2001*, 757-764. IEEE Press.

Hammond, K. R., Stewart, T. R., Brehmer, B., Steinmann, D. O. (1975) Social judgment theory. In Kaplan, M. F. & Schwartz, S. (eds.) *Human judgment and decision processes* (pp. 271-312). New York: Academic Press.

Norman, K. L. (1974). Rule learning in a stimulus integration task. *Journal of Experimental Psychology, 103,* 941-947.

Strategies and eye movement of an expert in a video-game

Hidemi Ogasawara (hidemi@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Sciences, Chukyo University; 101 Tokodate, Kaidu-cho
Toyota, Aichi 470-0393 Japan

Takehiko Ohno (takehiko@brl.ntt.co.jp)
Communication Science Laboratories, NTT; 3-1 Morinosato, Wakamiya
Atsugi, Kanagawa 243-0198 Japan

Introduction

The purpose of this study is to explore human cognition in dynamic environment. Using a video-game, "Pac-man" (Fig.1), we have found the process of acquiring the expertise, which included the play strategy shift from a safer defensive strategy to a risky offensive one (Ogasawara & Ohno, 1999). These play strategies require different information acquisition strategies. To explore the relations between the play and information acquisition strategies, we examine again the player's eye movement data in Ohno & Ogasawara (1999).

Case

"Pac-man" is a game that a player controls Pac-man to eat dots while escaping from ghosts. If Pac-man eats one of four extra large dots (PPs), it becomes "strong" and can attack the ghosts for a limited time. The killed ghosts are transferred to the center of the screen and re-join the game. The game ends if the player cleared all dots on the screen or all Pac-men were lost. The game was implemented on Sun Sparc 10 with an eye mark recorder, NAC EMR-NC. One undergraduate student participated in this study. One session of the task usually included five games. The participant performed 24 sessions with one session in a day.

Result and Discussion

The sessions were divided into four periods (6 sessions/period) in the following analysis. The subject showed the similar play strategy shift as the previous subject (Ogasawar & Ohno, 1999) in the early periods of the 80 sessions. For the eye mark data, we examined the distance among Pac-man, the ghosts, and the eye mark. The result showed a tendency that the subject looked less around the Pac-man as he played more games.

Next we examined the details of the strategies and eye movement. One of the offensive strategies observed in the previous study was to move Pac-man to the center area after the consumption of PP. This strategy gives chances to kill ghosts twice: Pac-man can kill ghosts nearby, and after moving closer to the center it can kill once again the reappearing ghosts. But it is risky because Pac-man's "strong" time might be expired on the way to the center. This strategy was also observed in this study. For example, for PP (LR) in the lower right corner this type of strategy was observed more frequently in the later session (from the 1st to the 4th period, 0%, 7.1%, 42.3%, 25.0% respectively). The four squares in Fig.1 show distributions of the player's eye marks in Pac-man's "strong" time after the consumption of the PP, from the 1st to the 4th period respectively. The eye marks are more concentrated around the PP and the center area as the periods progressed. The same tendency was observed for the other PPs.

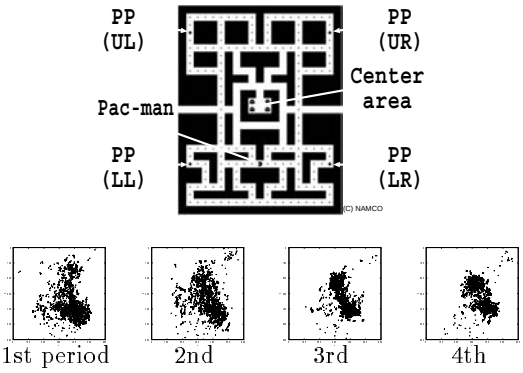


Figure 1: Display for "Pac-man" and distribution of eye marks after consumption of PP(LR)

These results indicate that the play expertise in this case involves the acquisition of the play strategy of paying less attention on Pac-man and shifting it to the score-gaining but risky moves.

References

Ogasawara, H.,& Ohno, T. (1999). Expertise process of human video-game players. *Proceedings of ICCS/JCSS99* (pp. 388–393).

Ohno, T., & Ogasawara, H. (1999). Information acquisition model of the highly interactive tasks. *Proceedings of ICCS/JCSS99* (pp. 288–293).

Not so Fast! (And not so Frugal): Rethinking the Recognition Heuristic

Daniel M. Oppenheimer (bigopp@psych.stanford.edu)
Department of Psychology, Stanford University
Building 420 – Jordan Hall, Stanford, CA 94305 USA

People face a lot of decisions and it stands to reason that we would want to expend as little cognitive effort as possible while still remaining accurate. Gerd Gigerenzer and his colleagues (1996; 1999) have contended that individuals have limited cognitive capacity, and are unable or unwilling to utilize complex statistical methods in decision making. Thus, individuals use heuristics in order to approximate "optimal" strategies more quickly, and at a much lower cognitive cost; hence the term "fast and frugal".

The simplest of these heuristics is the recognition heuristic (RH) (Goldstein & Gigerenzer, 1999). Simply stated, RH claims that when making a judgment about two items, an individual who only recognizes one of the items will consider the known item to have a higher value. This is an important heuristic not only for its elegant simplicity, but also because it is the first step in a variety of other fast and frugal heuristics (Gigerenzer & Todd, 1999)

To test whether individuals actually use RH, Goldstein and Gigerenzer (1999) asked Americans to make population comparisons among pairs of cities taken from the 30 largest cities in Germany. Participants were also quizzed as to which cities they recognized. The researchers found that when a participant recognized only one city in a pair, he/she judged that city as larger about 90% of the time.

Goldstein and Gigerenzer (1999) clearly assert that the level of recognition is not important in using RH, "the distinction relevant for the recognition heuristic is that between unrecognized objects and everything else". They discuss the "inconsequentiality of further knowledge" as an essential feature to maintain the frugality of the heuristic.

Accordingly, an individual using RH should judge a recognized city as larger than an unknown one *even if the recognized city is known to be small.* To test this, 50 participants were asked to judge populations of local cities that were known to be small, as compared to made-up cities (which, by virtue of being fictional, were unrecognizable).

Across all cities, only 37% of responses were consistent with RH. Thus, participants were significantly more likely to be inconsistent with RH than chance ($\div 2 = 4.25$, $df = 1$, $p < .05$). Results by city are summarized in table 1.

Table 1: Results of Experiment 1.

City	% using RH	City	% using RH
Cupertino	.30	Milpitas	.33
Sausilito	.20	Berkeley	.35
Foster City	.46	Freemont	.53
Total	**.37**		

One explanation for the discrepancy between these results and those of Gigerenzer & Goldstein (1999), might be found in attributions of mental states. Individuals may recognize a city, and attempt to determine *why it is that they do so*. One reason for recognition might be size (large cities are more likely to be well known). However, when there is an alternate reason for recognition – in this case proximity – individuals may attribute their mental state to the alternative. That is, when there are reasons other than size that one might recognize a city, an individual may be less likely to use recognition as a cue that the city is large.

To test this, 172 participants were asked to make population estimates on cities which were famous for virtues other than their sizes (e.g. nuclear accident, featured in literature, etc.) as compared with made-up cities.

Slightly over 40% of the trials were consistent with RH. Subjects were significantly more likely to be inconsistent with RH th
are summarized in table 2:

Table 2: Results of Experiment 2.

City	% using RH	City	% using RH
Los Alamos	.38	New Haven	.52
Cherynoble	.29	Timbuktu	.40
Nantucket	.36	**Total**	**.40**

This data suggests that although individuals do use recognition as a cue for size estimations, they do so in a more complicated manner than conjectured by Goldstein & Gigerenzer (1999). Individuals appear to make attributions about their mental state of recognition, and perform some kind of Bayesian discounting based upon that attribution. While it is beyond the scope of this abstract to discuss the mechanism thoroughly, it is clear that RH may not be as fast or frugal as it was originally postulated.

Acknowledgments

This material is based upon work supported under a National Science Foundation Research Fellowship. The author would like to thank Josh Tenenbaum and the Tenenbaum lab for advice and support.

References

Gigerenzer, G. & Todd, P.M., (1999) Simple Heuristics that make us smart. New York: Oxford University Presss.

Goldstein, D.G. & Gigerenzer, G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review,103(4)

Goldstein, D.G. & Gigerenzer, G. (1999) The recognition heuristic: How ignorance makes us smart.. In G. Gigerenzer & P. Todd (Eds.), *Simple Heuristics that make us smart.* New York: Oxford University Press.

Neural Correlates of Perceptual/Semantic Encoding and Implicit/Explicit Retrieval: An fMRI Study

T. Park (tpark@chonnam.ac.kr)
Department of Psychology, Chonnam National University
Kwangju 500-757, Republic of Korea

The important distinction between implicit and explicit memory tests is based on intentional effort and conscious recollection experience during retrieval process (Schacter, 1987), which have been suggested to be associated with increased activity in prefrontal and medial temporal regions, respectively (Schacter & Buckner, 1998). The present study investigated brain areas activated during tasks involving implicit and explicit memory retrieval, and examined neural correlates of conscious recollection and intentional effort during memory retrieval.

Methods

Whole-brain functional MRI was used to examine 8 subjects during retrieval in a block-designed fMRI experiment (Fig. 1). Two incidental study conditions were manipulated: Semantic and perceptual word encoding conditions. This manipulation of level of processing (LoP) was expected to yield two retrieval conditions that differed with regard to intentional retrieval effort and successful conscious recollection: Semantic encoding yielding low level of retrieval effort with high level of retrieval success and perceptual encoding yielding high level of effort with low level of retrieval success. After studying, word fragment completion (WFC) task was presented and then cued recall (CR) task was presented with word fragment cues.

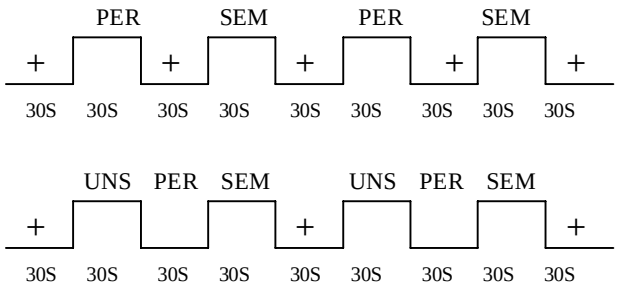


Fig. 1. A schematic illustration of study (above) and retrieval (below) paradigm. Critical study blocks (Semantic Encoding and Perceptual Encoding) were 30-s long separated by 30-s blocks of fixation (+). Critical retrieval blocks (Unstudied, Perceptually studied, Semantically studied) were also 30-s long separated by 30-s blocks of fixation in each of WFC and CR tasks.

Results

During explicit retrieval (CR) of the semantically encoded words, right inferior frontal regions (Brodman Areas 45, 47) were activated but right anterior frontal regions (BA 10) were deactivated. These results suggest different roles of different prefrontal regions during explicit (episodic) memory retrieval: BA 45/47 involved in conscious recollection and BA 10 in intentional effort (McIntosh et al., 1997). Also, parahippocampal gyrus was activated during explicit retrieval of the semantically encoded words, and this result supports the idea that medial temporal lobe is a neural correlate of conscious retrieval success.

During implicit retrieval (WFC), occipital lobe (BA 17, 18 including fusiform gyrus) showed reduced activation when word fragments were primed, and this result supports the view that posterior areas are neural correlates of perceptual priming. Unexpectedly, right parahippocampal gyrus showed increased activation during implicit retrieval of the semantically encoded words. This result suggests that LoP effects which were often observed in studies of implicit memory retrieval could be the result of involuntary recollection (explicit contamination).

Acknowledgements

This research was supported as a Brain Neuroinformatics Research Program sponsored by Korean Ministry of Science and Technology.

References

McIntosh, A. R., Nyberg, L., Bookstein, F. L., & Tulving, E. (1997). Differential functional connectivity of prefrontal and medial temporal cortices during episodic memory retrieval. *Human Brain Mapping, 5*, 323-327.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501-518.

Schacter, D. L., & Buckner, R. L. (1998). On the relations among priming, conscious recollection, and intentional retrieval: Evidence from neuroimaging research. Neurobiology of Learning and Memory, 70, 284-303.

Mental Rotation Transfer

Philip Pavlik (ppavlik@andrew.cmu.edu) and John Anderson (ja@cmu.edu)
Carnegie Mellon University
Department of Psychology, 5000 Forbes Ave.
Pittsburgh, PA 15213 USA

The mechanisms underlying the effects of practice on mental rotation are still incompletely understood. Explanations that may apply include improved rotation processes (Wallace & Hofelich, 1992) and the learning of figure exemplars (Tarr & Pinker, 1989). One comparison of these two theories supported an exemplar explanation: Heil, Roesler, Link and Bajric (1998) looked for transfer of mental rotation performance in three conditions where 3-D block figure stimuli pairs were either identical to practiced pairs, the same objects around mixed old and new axes, or new objects around mixed axes. They found that mental rotation skill only transferred to the identical condition.

In contrast to this, in an experiment that included less practice and a transfer condition in which only the figure views were different, we found transfer of rotation performance.

There were two parts to the experiment. Participants trained for five blocks of 32 trials each, following which they tested in one mixed block for 128 trials. The four stimuli figures were identical to Shepard and Metzler (1971) and presented in pairs at four angular disparities: 0°, 40°, 80°, or 120°. Presentation was randomized and one-half of the trials were unanalyzed mirror-image foils.

Of the 128 test trials, 32 were identical to training, 32 involved the same axis of rotation but substantially new views on the same stimuli (at least a 90 degree oblique rotation), 32 involved the same view on the stimuli but an orthogonal axis of rotation (in this case the stimuli from the same/ same condition were simply presented rotated 90° around the Z), and 32 involved both an orthogonal axis of rotation and new views on the stimuli. This design allowed for a two by two (same vs different figures, same versus different axes) within-subjects comparison. Four between-subjects conditions of the experiment were run to counterbalance for possible effects of the stimuli set or the trained axis of rotation. Results for these counterbalanced conditions were not significantly different and were aggregated. 47 undergraduates participated of which 9 were discarded for failing to meet criterion performance.

Average millisecond per degree rotation speeds for each of the training blocks were calculated based on the assumption of a linear relationship of rotation latency and angular disparity. There were significant indications of learning indicating that participant RS's improved over training. In transfer participants showed a significant advantage to rotation around the same axis as training that did not depend on whether the figure views were the same or new. Mean times to judge 0° rotation pairs (intercept times) were significantly faster in the same-same condition than in any of the other conditions. See Table 1.

Table 1: Data Summary

	Train Blk. 1	Same/ Same	Same/ Orth.	New/ Same	New/ Orth.
Mean RS (ms/°)	18.8	15.2	16.3	13.3	17.2
0° (ms)	1830.5	1134.9	1345.3	1350.4	1327.7
40° (ms)	2800.5	1845.6	2242.7	2255.9	2345.4
80° (ms)	3314.3	2381.1	2949.4	2861.3	2972.4
120° (ms)	4164.7	2984.9	3280.8	2922.8	3410.6

Unlike the Heil, et al. (1998) study, our results suggest that transfer of rotation skill involving rotation around a particular axis can occur to new stimuli views. While we do not claim to refute Heil et al. (1998), it does seem that non-stimulus-specific transfer can occur. On the other hand the special advantage for 0° rotation pairs that were repeated from training does suggest that there is also a stimulus-specific component to the learning. The fact that the new view/same axis transfer condition showed transfer of RS, yet showed no benefit in the 0° disparity (identity recognition) trials suggests that our new view stimuli condition was significantly unique and transfer of rotation skill to these new views depended on general rotation learning independent of any exemplar view strengthening. As revealed by the data and a subsequently formulated ACT-R model, there is a complex variety of learning that is taking place in the first 200 trials of a mental rotation experiment.

Acknowledgments

Preparation of this paper was supported by grant F49620-99-1-0086 from the Air Force Office of Scientific Research.

References

Heil, M., Roesler, F., Link, M., & Bajric, J. (1998). What is improved if a mental rotation task is repeated--the efficiency of memory access, or the speed of a transformation routine? *Psychological Research, 61,* 99-106.

Shepard, R. N. & Metzler, J. (1971). *Mental rotation of three-dimensional objects. Science, Vol. 171,* 701-703.

Tarr, M. J. & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233-282.

Wallace, B. & Hofelich, B. G. (1992). Process generalization and the prediction of performance on mental imagery tasks. *Memory & Cognition, 20,* 695-704.

What do you understand for X?

Célia Lúcia GOMES PESSANHA
LCC/CCH/UENF
Av. Alberto Lamego 2000
28015-620 Campos RJ Brazil
T: 55 22 27261589 / F: 55 22 27261589
Celiapessanha@aol.com

Adriana SOARES
UNIVERSIDADE GAMA FILHO
Mestrado em Psicologia
Rua Manuel Vitorino 625 Prédio CP
20748-900 Piedade Rio RJ Brasil
T: 55 21 22748407 F: 55 21 22748409
mespsi@ugf.br

This study approached the language and the thought under the cognitive perspective. We investigated the interaction, dependence, or independence of these cognitive processes, but above all J. Fodor's inatist presupposition of a "language of the thought." Such presupposition transcends the specific subject of the language and of the thought. Fodor (1975) affirms that it would exist in the human brain a structure that would take us to develop a formal system of linguistic registrations that would incorporate all of the universal properties of the language. This "language" would be the communication between the mental states and the structures of the thought. Therefore, it would allow us to do hypotheses regarding the knowledge that we want to acquire, and, still, to classify, and to classify it. This way, we elaborated an experimental situation in which we showed that children from 6 to 7 years (in the first school stage), are not capable to acquire real new concepts not belonging to the immediate universe of knowledge. In general terms, our results point that after the contextualization is notable the index of understanding of the new concept for the children. Although this concept is distant of their cognitive universe, it is significant for them, because they are part of real events of their lives. Partly, we corroborated the theory of Fodor that we acquired new concepts starting from the formulation of hypotheses. However, unlike Fodor that affirms that the innate mental structures are responsible for the acquisition of new concepts, Jean Piaget sustains that such concepts are acquired through the individual's interaction with the middle in that he is always not having any innate determinant. In this case, the categorization of new concepts involved a mental representation purely abstract, as foresaw Fodor, however, sustaining bonds cultural, historical and social in the understanding and apprehension of the new concepts, as announced Piaget. We believed that the results demonstrate that the contextualization of the new concept in the children's cognitive universe is the first step for a significant learning.

Mental representation in mathematical problem resolution

Maridelma POURBAIX
LCC/CCH/UENF
Av. Alberto Lamego 2000
28015-620 Campos RJ Brazil
T: 55 22 27261589 / F: 55 22 27261589
manoelcaetano@hotmail.com

Adriana SOARES
UNIVERSIDADE GAMA FILHO
Mestrado em Psicologia
Rua Manuel Vitorino 625 Prédio CP
20748-900 Piedade Rio RJ Brasil
T: 55 21 25997139 F: 55 21 25997139
mespsi@ugf.br

In this paper we investigate mental-representation strategies applied to solve contextual problems that involve mathematical calculus. We are interested in formal procedures, algorithms, and strategies that could be used by three groups of peoples: with specific mathematical knowledge (*the expert group*); without this knowledge (*the control group*); and without this specific knowledge but acquainted with the designed problem context (*the familiar group*).

Our investigation demonstrates that the reasoning developments of these three groups are quite different from one another: in the expert group the reasoning is only based on algorithmic operations; in the control group the reasoning combines algorithm with other mathematical strategies; and the familiar group uses a more intuitive reasoning, probably influenced by their familiarity with the problem context and by their special mental strategies.

In this research we aim to come through with Cognitive Science studies on the nature of human knowledge. Mathematical problem resolutions in general and arithmetic calculus in particular are very interesting tools to recognize the difference between expert and novice knowledge and reasoning.

Browsing Multiple Texts under Time Pressure

William R. Reader (readerwr@cardiff.ac.uk)

Stephen J. Payne (paynes@cardiff.ac.uk)

School of Psychology, Cardiff University
Cardiff CF10 3YG, UK

The Problem of Multiple Texts

With the expansion of the World Wide Web and other electronic information sources, it is becoming increasingly important for learners to actively allocate their time among texts in order to maximize their learning. Finding relevant texts is no longer the main problem; rather the problem is one of adaptive time allocation among multiple relevant texts. What constitutes a good text is dependent on, among other things, the individual's background knowledge, since comprehension requires textual information to be integrated with this knowledge so as to construct a situation model. If there is too much overlap between the text and the reader's background knowledge, then the text affords little opportunity for learning, but if there is too little overlap then the text would be incomprehensible. Good texts for learning therefore fall in the middle ground that Wolfe, Schreiner, Rehder, Laham, Folz, Kintsch & Landauer (1998) call the zone of learnability.

Thus, at least one task facing the self-directed learner is to allocate his or her time selectively to texts that fall within this zone of learnability and ignore the rest. Experimental studies of metacognition (e.g. Son & Metcalfe, 2000) have explored the way in which prior judgments of text difficulty influence study-time allocation, but have been silent about our main question, which concerns the strategies by which difficulty judgements are integrated with browsing to produce preferential study. To understand these browsing strategies we follow Pirolli & Card (see 1999) in drawing on optimal foraging theory (see Stephens & Krebs, 1986).

Foraging Theory and Browsing

One of the findings in the optimal foraging literature that is particularly relevant to the issue of selective browsing is that animals will *sample* unfamiliar food patches in order to decide which to exploit. Krebs, Kacelnik & Taylor (1978) observed that great tits initially switched rapidly between two food patches before settling down to exploit the higher-value patch.

Do readers use a similar sampling strategy when deciding how to allocate their time among multiple texts? Sampling strategies have as an objective to choose the best source (of food or information). An alternative to such the sampling strategy is a satisficing strategy, in which readers continue to read any text that is good enough (which we take to mean that the text would still fall within the zone of proximal learning).

Experiments and Findings

We have conducted a number of experiments on reading multiple texts under time pressure, investigating the prevalence of sampling and satisficing strategies and the effectiveness of these strategies for the preferential allocation of time among texts.

The results of Experiment 1 suggested that readers were adaptive in that more expert readers allocated more time to more difficult texts, and that satisficing was a much more common strategy than sampling. The results of Experiment 2 suggested that the provision of outline overviews led to participants being more selective in the documents that they read, and encouraged sampling to the extent that it became the modal strategy. Other experiments have confirmed the general adaptive character of browsing and shown how document preference is influenced by the nature of the learning task (e.g. studying to answer factual questions leads to a preference for more difficult texts than does studying to write a general essay).

Acknowledgments

This research was funded by the Engineering and Physical Sciences Research Council of the UK under grant GR/M43302/01 awarded to S.J. Payne.

References

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature, 275*, 27-31.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 105*(1), 58-82.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning Memory and Cognition, 26*(1), 204-221.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes, 25*, 309-336.

Color Palettes for Displays: Optimization by Genetic Algorithm

John Rehling (rehling@cs.cmu.com)
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Introduction

The use of color can make computer displays and interfaces more readable. Ideally, the colors in a display will be maximally different from one another. For small values of *n*, it is not difficult to choose instinctively colors that are clearly distinct from one another. For larger *n*, however, it becomes more difficult to assure that the colors are immediately distinguishable from one another. A genetic algorithm can be used to pick large sets of colors such that all pairs in the set are, according to one metric, nearly maximally distinguishable. Even allowing for imperfections in the scheme, these sets are in practical terms very easily distinguishable, and could be of use to designers.

The approach

The metric used for the differentiability between two colors is the CIE 1976 L*U*V* color space CITE. A genetic algorithm was used to find sets of colors that optimized a fitness function based on the CIE metric. With *n* as the number of colors being explored in a given run, an individual consisted of 3*n* values, represented as real numbers from 0 to 255, which is a twist from most GAs, which use strings of Boolean bits to represent individuals. The mutation operator, traditionally the toggling of a bit when bitstrings are the basis of the representation, is instead the alteration of a 0-255 value by adding to it a random number with a distribution with tails that ranged from –100 and 100 (but clustered tightly near the mode of zero).

In each generation, the population would consist of the ten individuals rated highest by the fitness function, plus 45 copies of those individuals pass through the mutation operator, plus five new, random individuals. There was no breeding of individuals, since a set of colors is satisfactory only as a function of the whole set; except when *n* is very small, the cleaving together of two partial solutions is exceptionally unlikely to yield a good solution.

After some trial and error, a fitness function was derived that consisted of the CIE distance between two two closest (i.e, least-easily distinguishable) pair of colors in the set plus a small constant (0.00001) times the sum of CIE distances between all pairs of colors in the set. This makes the first priority that no two colors in the set are very much alike; as a secondary priority, it also tries to maximize other separations between color pairs given that the worst pair has been dealt with.

The genetic algorithm was run to find sets of colors for values of *n* from 2 to 12. In one set of runs, the colors were constrained only by what could appear on a typical computer monitor. In other sets of runs, it was required that black, white, or both black and white be included in the set of colors. Runs lasted until they appeared to converge upon a best value. At least three runs were conducted for each situation, with the best of all runs reported below.

Results

The printed form of the Proceedings is not the ideal medium for the presentation of the color palettes that the GA generated. The palettes are described here in terms of the R3 colors that most closely match the RGBs produced by the output. In many cases, the program produced values that differed slightly from any named color, although the differences from those listed below are almost always imperceptible.

Unconstrained: n=2 {lime, magenta}; n=3 {red, lime, blue}; n=4 {red, yellow, aqua, magenta}; n=5 {deepPink, orange, lime, aqua, deepViolet}; n=6 {black, red, deepPink, orange, lime, blue}

White plus: n=1 {red}; n=2 {red, lime}; n=3 {red, lime, blue}; n=4 {red, lime, teal, magenta}

Black plus: n=1 {red}; n=2 {red, blue}; n=3 {red, lime, magenta}; n=4 {darkOrange, limeGreen, blue, magenta}

Black, white plus: n=1 {red}; n=2 {red, lime}; n=3 {red, lime, blue}; n=4 {red, lime, blue, magenta}

References

Boff, K. and Lincoln, J., eds. (1987). *Engineering Data Compendium: Human Perception and Performance.* Dayton, Ohio: Armstrong Aerospace Medical Research Library.

Letter Spirit: An Architecture for Creativity

John Rehling (rehling@cs.cmu.com)
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Introduction

The Letter Spirit program is a model of human creativity in the domain of typeface design. The task of Letter Spirit is to take as input a few gridletters, letters rendered on a medium-resolution grid, that are intended to represent the same style. Using those seeds as a beginning, Letter Spirit creates versions of the remaining lowercase letters of the roman alphabet until it has completed an entire *gridfont* of 26 stylistically consistent gridletters.

Approach

Letter Spirit consists of three modules, each being a relatively complex program solving a vital subtask of gridfont design. The three modules are all based upon an architecture common to the cognitive models implemented by the Fluid Analogies Research Group. Most of those projects, including Copycat (Mitchell, 1993), Tabletop (French, 1992), and Metcat (Marshall, 1999) aimed at implementing increasingly refined models of analogy. One of the Letter Spirit modules (McGraw, 1995) pursues a similar approach, although its task is gridletter categorization rather than analogy.

The Letter Spirit program has a top-level loop that coordinates the three modules into a single strategy of design called *review-and-revision*. The program has one module, the Examiner, that detects letter category, while the Adjudicator evaluates style, and the Drafter, which creates new gridletters that aim to represent a goal letter category and that style.

To get around the brittleness that afflicts many AI models, Letter Spirit hands the Drafter's output to the other modules and they rate each gridletter for how well it fits its intended letter category and the intended style, respectively.

The design phase of a Letter Spirit run thus amounts to the execution of a loop, in which a letter category is selected and then the Drafter renders a gridletter that, ideally, incorporates the goal style as well as that letter. The Drafter's attempt is run past the Examiner and the Adjudicator, and if the attempt is the best version thus far for that category, as determined by the scores that the Examiner and the Adjudicator generate, then it is kept as the current version of that category in the gridfont. This loop runs many times, and as it does so, the quality of the gridfont should incrementally increase.

Results

Figure 1 shows five gridfonts that resulted from runs that each began with the program receiving five gridletters (in each case, 'b', 'c', 'e', 'f', and 'g') as input.

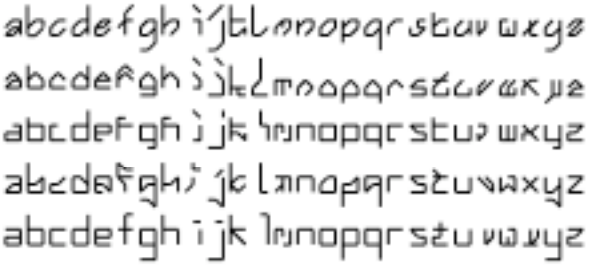


Figure 1: Five Letter Spirit-designed gridfonts.

While the Letter Spirit program captures human creativity, at best, partially, it is our belief that the model captures many essential aspects of human creativity. Perhaps more important, we argue, is the opportunity that the output provides to identify shortcomings in the current approach so that future models of creativity can converge more closely upon human creativity (Rehling,. 2000). More detail is at www.cogsci.indiana.edu/farg/rehling/lspirit/thesis

References

French, R. (1992). *Tabletop: An emergent stochastic computer model of analogy-making*. PhD thesis, University of Michigan, Ann Arbor, Michigan.

McGraw, G. (1995). *Letter Spirit (part one): Emergent high-level perception of letters using fluid concepts*. PhD thesis, Indiana University, Bloomington, Indiana.

Marshall, J. (1999). *Metacar: A sel-watching cognitive architecture for analogy-making and high-level perception*. PhD thesis, Indiana University, Bloomington, Indiana.

Mitchell, M. (1993). *Analogy-making as Perception.* Cambridge, Massachusetts: MIT Press/Bradford Books.

Rehling, J. (2000). *Letter Spirit (part two): Modeling creativity in a visual domain*. PhD thesis, Indiana University, Bloomington, Indiana.

How to Make a Computer Conscious

Alexei V. Samsonovich (asamsono@gmu.edu)
Krasnow Institute for Advanced Study, George Mason University
4400 University Dr. MS 2A1, Fairfax, VA 22030-444 USA

Introduction

Why do computers lack common sense initiative? What is so special about human brain and consciousness that computers still cannot reproduce? Much brain information processing deals with abstract representations of agents, such as instances of the self and others. Therefore, implementing the right sense of agency in a computer could provide a solution. Here I outline a general approach to the problem of a computer-based implementation of the mind.

Basic Formalism

This approach is based on the formalism of schemas. The notion of a schema used here is more general than any notion of a schema used, e.g., in psychology, or in computer science, or in social sciences. Examples of schemas range from very abstract notions (a thing, an entity, an event, a property, a relation, an agent, an act of learning a schema), to very concrete things (an apple, minus one, red, grasping a pen with the right hand, etc.). In this framework schemas can represent any cognizable thing, including qualia, thoughts, intentions, feelings, and so on. A predicate, a variable, a quantifier, an instruction or even an analog signal can be represented in terms of schemas as well. Elements and components of schemas, as well as complexes made of schemas, can be viewed as schemas on their own. Generally, a schema can be characterized as an abstract model, a template or a prototype that can be bound to a particular content. Definitions of particular schemas must specify semantics, syntax, pragmatics and dynamics. Schemas can be represented in a computer symbolically in a standardized format with the structure of a nested list.

When an instance of a schema is bound to a particular content, it forms a state. Schemas and states are dynamical objects in this framework. They evolve in time according to rules defined by schemas. E.g., states can be initiated, executed and terminated; dynamics of a state may affect other states and schemas. A chart is a special state that represents a mental perspective of an agent (e.g., I-Now, I-Yesterday, He-Imagined). When a state is bound to a chart, it represents a mental state. Perspectives change with time according to the flow of the subjective time: I-Now becomes I-Previous, and so on. In addition to perspectives, charts have internal dimensions that determine attitudes of the associated states: e.g., I-Now may contain a belief about a past or a future. The general idea of this framework based on schemas is not new; however, its known analogs (e.g., OPS, SOAR, ACT-R) lack generality and do not implement a proper concept of the self possessing free will.

Implementing Free Will and The Self

Charts labeled "I-…" represent instances of the self of the virtual individual, including cognitive and metacognitive perspectives. Other charts may represent mental simulations of "third persons": in other words, a "theory of mind". Only the content of I-Now represents the current content of consciousness in this framework. Mental states in I-Now have special privileges: e.g., they can directly determine scheduling of voluntary acts and may access all other charts.

A distinguishing feature of this framework is that it is based on the fundamental properties of the human self (*error fundamentalis*: Nadel & Samsonovich, 2002), including its uniqueness, its localization in a particular context, its indivisibility, its self-consistency over time, and finally, its apparent free will and self-awareness that are present in the contents of consciousness at any instance.

These properties are introduced via the design of the system or via dynamical constraints. They guarantee certain degree of coherence in system's behavior. Their products are the emergence and the maintenance of a unique working scenario (i.e., a consistent sequence of charts leading from I-Now to I-Goal) and the generation of voluntary actions. An action is considered voluntary, iff it results from an intention of "I" (represented by a mental state in I-Now), is consistent with the working scenario and is not biased by any additional factor. Intentions are selected among the available ideas based on their fitness into the working scenario, and ideas (i.e., mental states in I-Now that represent feasible actions) result from the activity of states.

Previously active charts may disappear with time together with their contents – or be remembered, thus contributing to the episodic memory and/or to the system of values, i.e., the memory of dreams and goals. In contrast, the semantic memory is represented by the set of schemas. Creation of new schemas is controlled by the learning schemas and may be supervised or not. A complete system should be able to work interactively with a human instructor, using a specially designed language. The range of interaction paradigms may include listening to and completing stories, learning how to play or playing games, learning languages or scientific disciplines, operating a robot in a virtual environment, simulating virtual characters, designing or training other systems of the same kind, etc. Primary applications of this framework should relate to natural language understanding.

References

Nadel, L., & Samsonovich, A. (2002). The conscious self. To appear in: S. Jess (Ed.). *Brain, Mind & Consciousness*. CA: University Press.

The Role of Prior Beliefs in Processing Analogical Arguments

Lelyn Saner (les53@pitt.edu) Christian D. Schunn (schunn@pitt.edu)
Department of Psychology, University of Pittsburgh
Learning Research and Development Center, 3939 O'Hara St.
Pittsburgh, PA 15260

Introduction

How is the background opinion of the person who is processing an analogy related to that person's evaluation of the analogy, both in terms of how the analogy is mapped to form a conclusion and in terms of the conclusion itself? Assuming that people have pre-existing perceptions of the analogs referred to, will they be more or less likely to reflect on and evaluate the analogy?

Several researchers have examined the degree to which the process of mapping between source and target analogs is influenced by the context in which it is done (Blanchette & Dunbar, 1997; Dunbar & Baker, 1994; Holyoak, 1985), including the type of problems, situations, or issues that are being compared, the environment in which the comparison is being made, and the goals or reasons that a person has for engaging in the mapping process. We explored the possibility that the attitudes that people have associated with the analogs prior to the presentation of the analogy might influence the way people respond to the analogy.

Methods

We elected to conduct this study over the Internet in order to get a broader, more representative sample of perspectives on the analogs, and participants were recruited primarily through Usenet newsgroups, where discussions and debates on various topics occur on a continuous basis. Two controversial socio-political issues, Abortion and Gun Control, were selected as target analogs for this study. Two novel source domains were selected for each of the two target domains such that they could be mapped to argue either for or against the key issue in the target domains.

Participants read a general introduction to the experiment on the first page of the website and were then asked for some demographic information. After completing a 14-item background opinion questionnaire to measure their opinions on the target issues, each participant performed two analogy-processing tasks, which instructed them to read a scenario that related one of the issues to one of its two analogs, and then to write a response to the scenario.

Each scenario argued for either a Pro- stance on the issue (Pro-Abortion or Pro-Gun Control) or an Anti- stance (Anti-Abortion or Anti-Gun Control). With the participants' actual background opinions on the issues and the record of which scenario that participant received for each issue, we determined if, for each issue, they received the statement supporting the same stance on that issue as their own. Participants were also asked, immediately after reading the scenario, to rate their level of agreement with it.

Responses were classified according to whether or not the analogy was processed in the response. Those that either presented an argument for a position that did not include any analogy at all or described their opinion on the issue only, without employing any argument, were coded as not processing the analogy. Three indicators were used to code a response as processing the analogy; (1) explicitly discounting the validity of the analogy presented in the scenario, (2) building upon the presented analogy in the response, or (3) choosing new analogies for the issue in favor of the one presented. Between two raters, using Cohen's Kappa as the measure of reliability, the agreement in classification of the responses was relatively high (κ=0.86, n=50).

Results

We observed that the likelihood of processing the analogy in a response was related to the average level of overall agreement with the scenario that was read. As the level of agreement with a particular scenario increased, so did the proportion of people who processed the analogy. At the same time, contrary to our expectations, there was no significant relationship, for either issue, between participants' background opinions on the issues and the likelihood of processing. On the basis of the first result, however, we concluded that there is evidence that peoples' opinions influence how they process analogies. We suspect that background opinion is one component of agreement with the scenario and that further exploration may reveal how it contributes to the way people process analogies.

References

Blanchette, I. & Dunbar, K. (1997). Constraints Underlying Analogy Use in a Real-World Context: Politics. Poster presented at the *19th Annual Meeting of the Cognitive Science Society*, Stanford, CA.

Dunbar, K., & Baker, L. M. (1994). Goals, analogy, and the social constraints of scientific discovery. *Behavioral and Brain Sciences, 17*, 538-539.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19) (pp. 59-87). New York: Academic Press.

A Pyramid Model of the Perception of Partially Visible Figures

Michael R. Scheessele (mscheess@iusb.edu)
Department of Computer & Information Sciences, Indiana University - South Bend
1700 Mishawaka Ave., South Bend, IN 46634 USA

Zygmunt Pizlo (pizlo@psych.purdue.edu)
Department of Psychological Sciences, Purdue University
1364 Psychological Sciences Bldg., West Lafayette, IN 47907 USA

Introduction

Frequently, figures in our visual field are only partially visible. One figure may partially **occlude** another, for example, or a particular figure may appear **fragmented** due either to camouflage or to low contrast between it and the background. Despite such challenges, the human visual system routinely perceives figures that may only be partially visible.

One prior theory of the perception of partially occluded figures (Nakayama, Shimojo, & Silverman, 1989) states that contours "intrinsic" to a figure of interest must be distinguished from those "extrinsic" to it and that this classification requires depth cues. Our theory proposes that the human visual system can use a *variety of cues, local or global,* to perform this classification and that this classification serves as the basis for perception of both partially occluded *and* fragmented figures. Further, we propose that an exponential pyramid, from the machine vision literature, provides a good model of how the human visual system implements this classification.

Exponential Pyramid Model Description

The Exponential Pyramid has been proposed as an adequate model of the human visual system (Rosenfeld, 1990; Pizlo, Salach-Golyska, & Rosenfeld, 1997). Our model uses a "non-overlapped quad-pyramid". Assume that the bottom layer of the pyramid has n processing nodes. The next layer has $n/4$ nodes, the one above that $n/16$ nodes, and so on. The top layer has only one node. Each node in a layer connects with four distinct 'child' nodes in the immediately lower layer and one 'parent' node in the immediately higher layer. Such a pyramid has $(\log_4 n) + 1$ layers. Each node in the pyramid has limited memory and processing capability. An image is input to the bottom layer (Jolion & Rosenfeld, 1994). The image may also be represented at each higher layer (with increasing spatial scale or 'receptive field size').

Our model features a bottom-up processing stage followed by a top-down stage. In the bottom-up stage, local variance of various contour features (e.g., orientation, length) is computed. When the variance of a contour feature abruptly changes between successively higher layers, the presence and position of a figure in the image is indicated (i.e., the figure 'comes into view'). In the top-down stage, the statistical information computed in the bottom-up stage is used to classify image contours as either intrinsic or extrinsic to the target figure. The model has only one free parameter: the standard deviation of decisional noise. Model and human performance were compared across 11 experimental conditions.

Method

In each trial of the **human psychophysical experiments**, a polygonal figure was partially occluded by simple shapes – diamonds (Exp. 1, two conditions) and squares (Exp. 2, nine conditions). A subject's task was to respond whether the figure was presented in its upright or rotated (180^o) position. Contours of occluders differed from those of the figure in terms of orientation (Exp. 1) and length (Exp. 2). Depth cues from occluders were minimal. **Model simulations** were run for all 11 conditions using the same sets of stimuli as those used by the human subjects.

Results

Subjects used orientation (Exp. 1) and length (Exp. 2) differences between the contours of a target figure and those of occluders, in detecting the figure. Model simulations accounted well for human performance in the 11 conditions of Experiments 1 and 2.

Conclusions

The human visual system can detect and use a variety of cues, local or global, to classify contours as either intrinsic or extrinsic to a partially visible figure. Our exponential pyramid-based computer model provides a good account of how the human visual system implements this process.

References

Jolion, J. M., & Rosenfeld, A. (1994). A pyramidal framework for early vision. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. Perception, 18, 55-68.

Pizlo, Z., Salach-Golyska, M., & Rosenfeld, A. (1997). Curve detection in a noisy image. Vision Research, 37, 1217-1241.

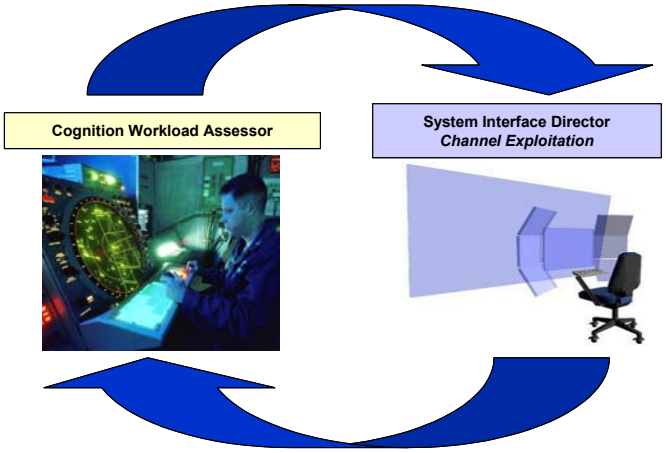
Rosenfeld, A. (1990). Pyramid algorithms for efficient vision. In C. Blakemore (Ed.), Vision: Coding and efficiency. Cambridge, Great Britain: Cambridge University Press.

Tomorrow's Human Computer Interaction from Vision to Reality: Building Cognitively Aware Computational Systems

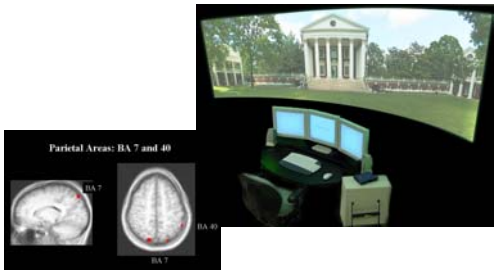
LCDR Dylan Schmorrow (dschmorrow@darpa.mil)
DARPA IPTO
3701 N. Fairfax Dr. Arlington, VA 22203 USA

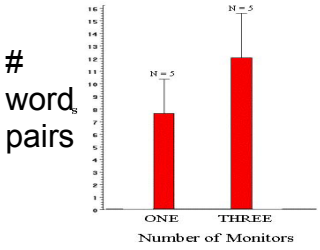
Amy A. Kruse (akruse@snap.org)
Strategic Analysis / DARPA
3601 Wilson Blvd., Suite 500 Arlington VA 22201 USA

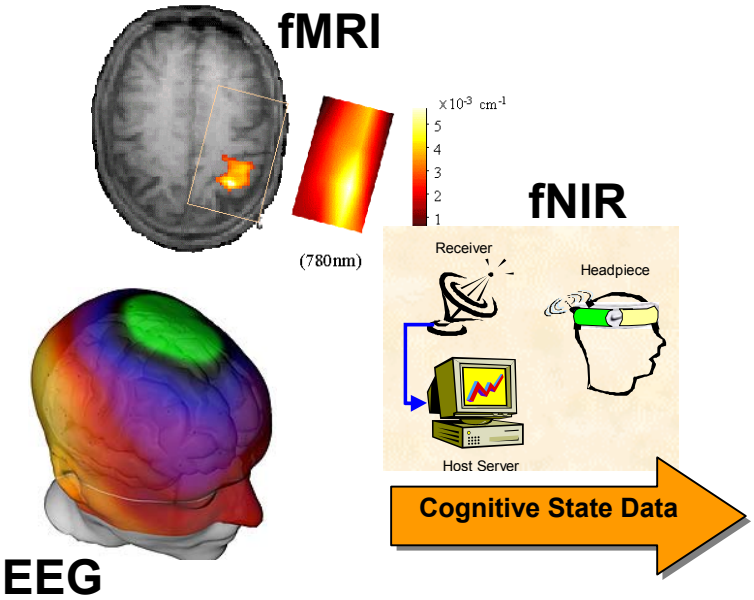
The goal of the DARPA Augmented Cognition program is to extend, by an order-of-magnitude or more, the information management capacity of the "human-computer" interaction by developing and demonstrating enhancements to human cognitive ability in diverse and stressful operational environments. Specifically, in its current phase, the program is focused on development of the technologies needed to measure and track a subject's cognitive state in real-time - these include functional near infrared imaging (fNIR) devices, as well as single site electroencephalographic (EEG) recordings. Through measurement and understanding of cognitive state in real-time, the ultimate goal of cognitively aware computational systems is within reach. Military operators are often placed in complex human-machine interactive environments that have been shown to fail when a stressful situation is encountered. The technologies under development in the Augmented Cognition program have the potential to enhance operational capability, support reduction in the numbers of persons required to perform current functions, and improve human performance in cognitively challenging environments.



InfoCockpit experiments explored ways to make it easier for people to encode, store, and retrieve information were conducted…







Learning by Collaborating Revisited: Individualistic vs. Convergent Understanding

Hajime Shirouzu & Naomi Miyake (shirouzu, nmiyake@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, 470-0393 Japan

Introduction

Roschelle (1992) characterized the process of "learning by collaborating" as a search of convergence among members. This paper makes a contrasting claim that each member is individualistic in how s/he interprets the learning task, how s/he solves it, and what kind of understanding s/he gains from collaboration. The others serve as a "monitor" (Miyake, 1986) to observe what the member is doing from a slightly broader perspective and to check its validity, which triggers the member's re-interpretation. This leads a learning pair to an iterative chain of re-interpretations, not to a search of the common ground among the two, which are often exchanged by verbal expressions in collaborative situations. In this paper we re-analyzed protocol data of collaborative learning processes in Roschelle (1992) and Shirouzu, Miyake & Masukawa (2002) to show that members' verbalizations reflecting their interpretations or re-interpretations are individualistic through the processes.

Re-analyses

Roschelle (1992)

Two students, Carol and Dana, elaborated their conceptions of velocity and acceleration using a computer simulation of a Newtonian micro-world. As Roschelle pointed out, they gradually revised and refined their verbal "metaphors" to mean these notions. A closer look at their protocol reveals, however, that the two did not seem to converge on the usage of particular metaphors to mean particular things. Dana started with a geometric "lengthen (addition)" metaphor to indicate the velocity vectors, the acceleration ones or their relations. Carol heavily used a "pull" metaphor to represent the dynamic relations between these factors. Carol finally verbalized an expression of "travel along" in the last episode, Episode 5, to explain the composition of velocity and acceleration, which Roschelle interpreted as an integration of the two metaphors, "lengthen" and "pull" one. Dana, however, did not use the verb "pull" other than in Episode 1, while Carol superimposed her metaphor to paraphrase Dana's insight into additive nature of vectors in Episode 3 (she said "right that's what I'm saying" to Dana without any specification of "what"). Besides, Dana did not share the expression "travel along" during training sessions and used "move along" in their post-training interview. They thus independently revised their verbal expressions. There might be two independent shifting processes of understanding.

Shirouzu, Miyake & Masukawa (2002)

When paired subjects were asked to indicate 2/3 of 3/4 of the area of a square sheet of paper, they shifted their strategies from the non-mathematical one in the first trials to the mathematical one in the second trials than solo subjects. Seven out of nine shifting pairs gradually generated the variations of solutions, from the most externally oriented two-step solution (making 2/3 out of the 3/4) to the external-internal mixed one (reinterpreting the externalized answer as one-half) to the most abstract one (2/3x3/4=1/2). Shirouzu et al. found that, though the members shared the algorithmic view of the task at the end of the first trial, the member who verbalized such a view during the first trial tended to propose the abstract solution in the second trial. Shirouzu (2001) also used a similar task in a small-case learning experiment with six 6^{th} grades only to find individual differences in the quality of their reports of six months later depending on their verbalization during the experiment.

Discussion

Under seemingly "one voice" in the collaborative situations, there were different courses or levels of understandings of the members, which were reflected by their particular language use (methodologically, the transfer task or the post interview--especially individually conducted one--reveals differences well). Roschelle thinks much of convergence because it warrants not only shared understanding between members but also their integration of scientific concepts. However, we can assume another, more real course of knowledge integration, in which each member gradually revises their interpretation of the task or the solution processes using the other's monitoring as stepping stones. Shirouzu et al. showed that the shift between solution variations coincided with members' role shifts between task doing and monitoring. Carol and Dana often proposed "what if" cases to each other to monitor their understanding. This series of re-interpretations enables the interactive and gradual integration of the solution variations of different abstraction levels. Careful analyses on the language use in collaboration makes us possible to feed a better folk-model of collaborative learning back to everyday learners (also see Miyake & Shirouzu, 2002 in this conference).

References

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science, 10,* 151-177.

Roschelle, J. (1992). Learning by collaborating: convergent conceptual change. *The Journal of the Learning Sciences, 2,* 235-276.

Shirouzu, H. (2001). Children's algorithmic sense-making through verbalization. *Paper presented at the 23rd Annual Conference of the Cognitive Science Society*, Edinburgh, The UK.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*

Retrieval Effects on Confidence in General Knowledge

Winston R. Sieck (sieck.3@osu.edu)
Department of Psychology, The Ohio State University
1827 Neil Ave, Columbus, OH 43210 USA

J. Frank Yates (jfyates@umich.edu)
Department of Psychology, The University of Michigan
525 E University Ave, Ann Arbor, MI 48109 USA

At a recent panel discussion sponsored by the Ohio Board of Education, credential-laden fellows of the Discovery Institute argued with absolute conviction that the Darwinian theory of evolution is flawed, and that we must have been created by intelligent design.

What underlies the confidence we have in our beliefs and knowledge? A typical assumption underlying many models of confidence in general world knowledge is that assessments of arguments that favor or oppose chosen answers will primarily, if not exclusively, determine confidence in choice (e.g. Griffin & Tversky, 1992; Koriat, Lichtenstein, & Fischhoff, 1980; also see Allwood & Granhag, 1996). Differences between models largely reflect distinct proposals for how evidence assessment is accomplished, and the extent and manner of bias that is postulated to exist in the process. For example, Koriat et al. (1980) suggested that memory search produces reasons for and against presented alternatives, and that assessment is biased in that choice-consistent reasons are relied upon more heavily than choice-inconsistent reasons.

The current study tests the hypothesis that confidence also depends in part on successful retrieval of topical information that is not directly relevant towards arriving at a choice. Successful retrieval of facts about the general topic is used as evidence that one is knowledgeable about the subject area, and thus likely to be reasoning correctly.

Method

Participants (n=159) were presented with a series of questions following the form: "Which species has a longer gestation period: (a) chimpanzees, or (b) humans?" Participants first chose one of the two alternatives as more likely to be correct, and then specified a probability between 50% and 100% that their choice was, in fact, correct. Subjects in a *reasons* condition were asked to write all possible reasons for and against each alternative answer, prior to choosing. Participants in a *recall* condition were asked to recall all of the facts they could about the topic of each question before choosing. *Control* condition participants simply answered the questions with standard, non-directive instructions.

Results

The principal results are shown in Table 1. Mean confidence, proportion correct, and overconfidence were virtually equivalent across conditions. Accuracy discrimination (mean confidence given correct – mean confidence given incorrect) was larger for recall than the other two conditions. Also, the correlation between confidence level and choice accuracy was larger for recall than the other conditions.

Table 1: Confidence/Correct Indices by Condition.

Indices	Control	Reasons	Recall
Mean Conf.	.73	.72	.72
Prop. Correct	.66	.66	.66
Accuracy	.04	.05	.10
Discrimination			
Pearson's r	.12	.16	.29

Discussion

Writing all possible reasons for and against alternatives had no impact on choice or confidence. Writing all facts that could be recalled about the topic of each question resulted in confidence judgments that better discriminated between correct and incorrect answers, as compared with control and reasons conditions. Preliminary results from coding of the listed reasons and recollections (not shown), indicate that approximately the same amounts of information were generated under the two procedures, but that important qualitative differences in the protocols seem to exist. It appears that successful recall of facts that are relevant to the topic, but that do not constitute reasons for or against presented alternatives, is taken as critical evidence that chosen alternatives are correct.

Acknowledgments

This research was supported by NSF Grant SES-9911301.

References

Allwood, C. M., & Granhag, P. A. (1996). The effects of arguments on realism in confidence judgements. *Acta Psychologica*, **91**, 99-119.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, **24**, 411-435.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 107-118.

Perception Matters: Effects of Perceptual Richness on Categorization

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
Ohio State University, 208b Ohio Stadium East, Columbus, OH 43210, USA

Anna V. Fisher (fisher.449@osu.edu)
Center for Cognitive Science
Ohio State University, 208b Ohio Stadium East, Columbus, OH 43210, USA

Abstract

Research presented here examines effects of perceptual and non-perceptual information on categorization as a function of perceptual properties of presented information. In Experiment 1, triads of perceptually-rich and perceptually impoverished stimuli were calibrated to equate discriminability of both kinds of stimuli. In Experiment 2, participants were presented with categorization task under two conditions, one, in which stimuli were not labeled and another, in which linguistic labels were provided. In the first condition, participants relied solely on perceptual information, and there was no difference between perceptually-rich and perceptually-impoverished stimuli. However, in the second condition, where linguistic labels were provided, there were dramatic differences across different stimuli types: perceptually-impoverished stimuli elicited mostly label-based responses, whereas perceptually-rich stimuli elicited mostly perceptually-based responses.

Introduction

It has been often argued that similarity is an insufficiently constrained rule to guide categorization decisions and that there is often a dissociation between similarity and categorization. For example, Rips (1989) presented participants with stories about animals undergoing appearance transformation. Participants were asked to rate similarity of the transformed organism, and to determine its category. While they judged the transformed animal to be more similar to the new category, they considered the transformed animal to be more likely a member of the old category (i.e., reptile). However, there is evidence (Johnson & Mervis, 1997; Sloutsky, Lo, & Fisher, 2001) perceptual information plays an important role in categorization. Therefore, one might counter argue that the transformational studies used only verbal descriptions or perceptually-impoverished pictures to demonstrate that perceptual similarity is of secondary (if any) importance for categorization. At the same time, it is possible that the role of perception is not fixed, such that contribution of perceptual information to categorization varies with variance in perceptual richness of presented information. This possibility was addressed in the present research.

Experiment: Categorization and perceptual richness

A total of 125 undergraduate students took part in the experiment. The experiment had a 3 (stimuli type: Perceptually-rich vs. Perceptually-impoverished) by 2 (Labeling condition: Label vs. No-Label) by 2 (similarity ratio: Test items are equally similar to the Target vs. Test A is more similar than Test B to the Target) mixed design with stimuli type and labeling conditions as between-subject variables and similarity ratio as a within-subject variable. In the label condition, pictures were accompanied with artificial two-syllable linguistic labels that were presented as count nouns (e.g., a bala, a gula). Participants were presented with triads of stimuli and asked to select from the two bottom pictures (i.e. Test A and Test B) the one that was the same kind of animal as the upper picture in the center (i.e. the Target).

Proportions of categorization choices were subjected to a three-way (Stimuli Type by Labeling condition by Similarity ratio) mixed ANOVA. As expected there was a significant stimuli type by labeling interaction, F (1,119) = 4.08, MSE = 0.11, $p < .05$, indicating that there were no differences among stimuli types in the no label condition, while there were significant differences in the label condition. This interaction is the most critical as it indicates that differences in the label condition do not stem from different discriminability of perceptually-rich and perceptually-impoverished stimuli. The results clearly indicate that reliance on perceptual and non-perceptual information for categorization is mediated by perceptual richness of information.

Acknowledgments

This research has been supported by grants from the National Science Foundation (BCS # 0078945) to Vladimir M. Sloutsky.

References

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. Journal of Experimental Psychology: General, 126, 248-277.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (21-59). New York: Cambridge University Press.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar: Linguistic labels and the development of inductive inference. Child Development, 72, 1695-1709.

Investigating Cognitive Gain in a Logical Experiment

Adriana SOARES
Mestrado em Psicologia
Universidade Gama Filho
mespsi@ugf.br
T: 55 21 25997139 F: 55 21 25997139

Cabral LIMA
DCC/IM/UFRJ
clima@dcc.ufrj.br
T: 5521 25983168/F: 5521 25983156

Several researches have been performed in various domains of knowledge about learning differences between learners regarded as experts and those regarded as novices. In some previous researches, we have investigated the learning, area of interest of researchers involved with cognitive science and computer science (AbLi96, Lima96, Po&col02 and Ca&col02).

In [AbLi96], for example, we conducted an experiment to examine characteristics related to cognitive gains in problem solving and the criterion to decide if a logical reasoning used in the solution of a specific problem could be used by similarity for other classes of problems. A more theoretical contribution concerning logic theory and its learning is in [Lima96].

[Po&col02] examines mental-representation strategies applied to solve contextual problems involving mathematical calculus: the interest is centered on formal procedures, algorithms, and strategies that could be used by three groups of peoples: with specific mathematical knowledge (*the expert group*); without this knowledge (*the control group*); and without this specific knowledge but acquainted with the designed problem context (*the familiar group*).

[Ca&col02] investigates the human reasoning applied to the mathematical problem-resolution process: the approach is based on two main settings: *a.* the investigation of mental processes involved in the human reasoning applied to problem resolution; *b.* the analysis of differences in the categorization and resolution of mathematical problems by beginners and specialists

We are actually developing a new experiment involving the teaching of computational logic by a playful approach (Socratic situation involving logical reasoning). This experiment aims to verify cognitive gains in two learners groups: experts and novices.

By examining the logical sequenced steps done to solve a logical entertainment we intent to detect criteria of utilization of reasoning by absurd or probabilistic reasoning by these groups. The extended idea concerns in fact the identification of cognitive contexts able to improve rather the use of probabilistic reasoning than reasoning by absurd and vice-versa. Of course, in our experiment, it is assumed incontestably that experts solve complex problems considerably faster and more accurately than novices do. Those differences are commonplaces of everyday experience, yet only recently have we begun to understand what the expert does differently from the novice to account for this superiority. Nevertheless, some initial results of our investigation could be applied to the learning of computational logic. These results could be applied in an elaboration of ideal pedagogical guidelines in order to facilitate the utilization of a more appropriate kind of reasoning to solve a particular Socratic problem.

References:

[AbLi96] ABBOUD R., LIMA C.: *Problem solving difficulties: a comparison of expert and novice logic reasoning.* Advances in Database and Expert Systems, IIAS Editions, ISBN 0921836333, Windsor, Canada, pp 65-69, 1996.

[Ca&col02] CAETANO M., SOARES A., LIMA C.: *Human reasoning and logical-formal guidelines: an analysis of the mathematical problem-resolution process.* Paper accepted to be published in the proceedings of the 14th International Conference on Systems Research, Informatics, and Cybernetics. Baden-Baden, Germany, 2002.

[Lima96] LIMA, C.*: Can you decide when to use reduction to the absurd and probabilistic reasoning to solve problems in mathematical logic?* Advances in Artificial Intelligence and Cybernetics. IIAS Editions, ISBN 0921836-42-2, Windsor, Canada, pp. 66-70, 1996.
[Po&col02] POURBAIX M., SOARES A., LIMA C.: *Challenges and implications in cognitive sciences: on mental representation in mathematical problem resolution.* Paper accepted to be published in the proceedings of the 14th International Conference on Systems Research, Informatics, and Cybernetics. Baden-Baden, Germany, 2002.

Children's Developing Ability to Create External Representations: Separating *What* Information is Included From *How* the Information is Represented

Lara M. Triona (triona@cmu.edu) & **David Klahr (klahr@cmu.edu)**
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Many areas of education focus on teaching children how to produce specific kinds of external notations. Although most children appreciate the usefulness of creating such representations by their eighth year, their ability to actually *create* useful external representations is highly variable and task specific until adolescence (Cohen, 1985; Karmiloff-Smith, 1979; Lee & Karmiloff-Smith, 1996; Eskritt & Lee, 2001). This variability could result from children's failure to realize *what* information needs to be included in a good representation, or it could result from their inability to figure out *how* to create such a representation, even if they know what to include in it. It is likely that these two difficulties interact such that, when children are deciding on their method of representation, they fail to verify that less salient pieces of information will be included. The current study is a first step in evaluating the contingencies between (1) *what* information is included, (2) *how* the information is represented, and (3) children's ability to create useful external representations.

We asked first and second grade children to create external representations that could be used by others to replicate a complex sequence of actions in a simple puzzle (Klahr, 1985). Children generated the sequence of moves themselves within a well-defined problem to assure they had sufficient understanding of the action sequence to be represented. The closed structure of the problem made it possible to define what specific information needed to be included in the external representation. In addition, we categorized the method of representation both for the overall organization of the notation and specific for each piece of information. In our analysis, we examined how the adequacy of the external representation (i.e., can another person use the notation to replicate the sequence of actions) relates to *what* information is included and *how* the information is represented.

Our analysis of children's external representations distinguishes among (a) references to the object to be moved, (b) the location to move the object, and (c) information about order of the moves. In order to categorize the overall organization of children's notations, we coded their notations as either linguistic or figural. The adequacy of representations was evaluated on a seven-point scale, with 1 denoting no relation between the notation and the task and 7 indicating that all information that was required to replicate the sequence was explicitly represented.

Of the three types of information included, children were more likely to refer to the objects and locations than to sequential order, $F(1, 25) = 21.8$, $p < .001$. This result is consistent with prior findings that suggested children had difficulty including explicit sequence information (Bolger & Karmiloff-Smith, 1991; Lee & Karmiloff-Smith, 1996). Because information about move order was necessary for the notation to be rated as very adequate, the inclusion of sequential information in the notation was related to adequacy, $r = .82$, $p < .001$.

Similar to Lee & Karmiloff-Smith (1996), notational adequacy was associated with the overall organization of the notations, $F(1, 23) = 21.4$, $p < .001$, such that linguistic notations were more adequate than figural notations. The differential inclusion of information about sequence was highly related to notation type, $X^2(N = 24) = 12.2$, $p < .001$; all of the linguistic notations included information about sequence while a majority of the figural notations omitted this information. By categorizing the sequential information included as either explicit (i.e., words or numbers) or implicit (i.e., position on page such as top to bottom), we found that sequential information included in the linguistic notations was more likely to be implicit than in figural notations, $X^2(N= 13) = 5.3$, $p = .02$. However, we found an interaction between overall organization and method used to represent sequential information for notational adequacy, $F(1, 9) = 10.91$, $p = .01$, showing that figural notations with explicit sequential information did not have as high of an adequacy as linguistic notations.

The current study revealed a complex relationship between *what* information is included, *how* that information is represented, and notational adequacy. Future research should address why figural notations with explicit sequential information were not as adequate as linguistic notations.

References

Bolger, F. & Karmiloff-Smith, A., (1990). The development of communicative competence: Are notational systems like language. *Archives de Psychologie, 58*, 257-273.

Cohen, S. R. (1985). The development of constraints on symbol-meaning structure in notation: Evidence from production, interpretation, and forced-choice judgment. *Child Development*, *56*, 177-195.

Eskritt, M. & Lee, K. (2001). "Remember where you last saw that card:" Children's production of external symbols as a memory aid. *Developmental Psychology, 38,* 254-266.

Karmiloff-Smith, A. (1976). Micro- and macro-developmental changes in language acquisition and other representational systems. *Cognitive Science*, *3*, 91-118.

Klahr, D. (1985). Solving problems with ambiguous subgoal ordering: Preschoolers' performance. *Child Development, 56,* 940-952.

Lee, K. & Karmiloff-Smtih, A. (1996). The development of cognitive constraints on notations. *Archives de Psychologie, 64,* 3-26.

What Does it Take to Pass the False Belief Task? An ACT-R Model

Lara M. Triona (triona@cmu.edu)
Amy M. Masnick (masnick@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Bradley J. Morris (bjmorris@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

The false belief task is used to assess whether children have a theory of mind (i.e., whether they know that other people can hold different beliefs). One version is the *unexpected contents* task (Perner, Leekam, & Wimmer, 1987), in which a child is shown a box (e.g., a crayon box) and its unexpected contents (e.g., candy). After the box is closed, a child is asked, "What did you think was in the box?" While children under 4 tend to answer "candy," older children respond correctly with "crayons." In order to explain this age effect, it is important to understand what is needed to pass the false belief task.

A computational model is one means of specifying the processes required. We designed an ACT-R (4.0; Anderson & Lebiere, 1998) model of the *minimal* processes needed to simulate performance on the false belief task. The model consists of five productions: two that respond to the two control questions, two that respond to the false belief question, and one that stops the model.

Our model includes three types of declarative knowledge: (1) *goals* contain the information presented in current question (e.g., a closed crayon box); (2) *general knowledge* provides relevant prior knowledge (e.g., crayons are usually inside a crayon box); and (3) *objects* indicate object-specific information (e.g., there is candy in *this* crayon box).

The first two productions specify the processes by which children respond to control questions. The expected contents production accesses prior knowledge about those types of boxes and identifies the contents based on this general knowledge. The second production uses specific input about the contents of the package (e.g., candy in the crayon box) to update the object-specific knowledge.

A correct response to the false belief question requires only a modification of the expected contents production: (1) identify the current question as a special case, (2) ignore the content knowledge about the *specific* box, and (3) refrain from changing object-specific knowledge based on prior general knowledge. Our model does not need to consider mental representations or beliefs in order to respond that crayons would be expected in a crayon box despite knowing there is candy inside the crayon box.

If a child fails to recognize the false belief question as a special kind of question, we expect she will simply report the actual contents of the box. The final production stops the model after responding to the false belief question.

When the model is run, each of the control questions match only one production; thus the model always responds correctly. However, when the false belief question is posed, the model matches both the modified expected contents production and the report knowledge production. The developmental pattern in responses can be modeled by hypothesizing that the older children have had further experience with these special questions while the younger children have not. We modeled this by manipulating the parameter q – the probability that the production would achieve the goal. The pattern of results, shown in Figure 1, is similar to the pattern in children's responses. The model predicts that the reaction time for the correct response will be 500 milliseconds longer than an incorrect response.

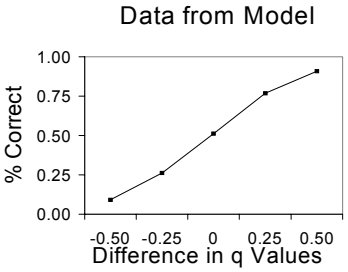


Figure 1: Data from model showing % of runs that pass the false belief question as the difference in q varies from report knowledge production being more successful to the modified expected contents being more successful.

The current model only requires distinguishing questions that require reporting current knowledge from ones that require ignoring current knowledge. Because Wellman, Cross & Watson's (2001) meta-analysis found the developmental pattern of results for different versions of the false belief task was robust, it is likely that the current model can be generalized to other variations. We consider the current model a first step in specifying alternative explanations for false belief performance. This use of computational modeling can be productive in refining our understanding of the development of children's theory of mind into a more specified, and therefore testable, theory.

References

Anderson & Lebiere (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Perner, J., Leekam, S.R., Wimmer, H. (1987). Three-year-olds' difficulty understanding false belief: Representational limitation, lack of knowledge or pragmatic understanding? *British Journal of Developmental Psychology, 5,* 125-137.

Wellman, H. M., Cross, D., Watson, J. (2001). Meta-analysis of Theory-of-Mind development: The truth about false belief. *Child Development, 72(3)*, 655-684.

The Grounding of Symbols in Affordances

William H. Vidal
Scientific and Philosophical Studies of the Mind Program
Franklin and Marshall College
Lancaster, PA 17604
Wh_vidal@fandm.edu

Affordances vs. Experience

In the past two decades, the development of artificial intelligence has received thorough criticism from many philosophers. In his paper, The Symbol Grounding Problem (1990), Harnad attempts to prove A.I. systems' inaptitude to ground computational symbols in experience. To determine whether experience is the only criteria for grounding, a non-representationalist framework to artificial systems must be applied. This paper is an attempt to demonstrate the ability of artificial systems to ground their symbols in the potential activity afforded by their environment. To illustrate the grounding of symbols in affordances the analysis is presented in terms of Marr's three descriptive levels (Marr, 1983). Within the computational level, I present an ecological model of perception, as well as, how the behavior of an A.I. system is intelligible in terms of affordances. The placement of an A.I. system's behavior within the broader context of its environment widens the potential for grounding and draws the focus away from inner formal-symbol operations. Following the establishment of the environment's relevancy in A.I. systems' perceptual mechanism is an examination of the representational level. The program, at the representational level, represents the bridge between a system's ecological perceptual mechanism and the implementation of stimulus information. I conclude the analysis with the implementation layer's implications on which symbols need grounding and the causal link between the reception of stimulus information and motor commands.

The symbol Grounding Problem

The problem theoretically arises because the symbols in an A.I. system's representation layer are manipulated formally according to preset programmed rules and do not have any causal connections with the exterior world. In other words, the symbol grounding problem highlights the lack of connectedness between the symbols within the programmed layer of an A.I. system and the exterior environment. Whether one is trying to prove A.I. systems can have intentionality, develop a potential humanoid, or examine the replication of human processes through artificial intelligence, the symbol grounding problem poses a barrier. How can one potentially make a machine that has meaningful thoughts, if its symbols are detached from any form of reality? According to Harnad, human mental symbols are grounded in our daily interactions with the exterior world. The association of symbols with memories leads Harnad to equate symbol grounding's constitutive element to experience and memories. (Harnad, 1990)

An Ecological Model

Within the bounds of his analysis, Harnad forwards valid criticisms of A.I. systems. Indeed, from a program layer investigation and a focus on internal abstract computation, A.I. systems do not have any connections with the exterior world. However, one is not claiming that a system's variables and design do not originate from a programmer and a system's general concept is the realization of a programmer's abstractions. These facts simply entail that the construction of an A.I. system is initially based on a designer's abstractions. Although the behavioral and program layer are the components we perceive and control, a grounding mechanism's processes reside within the implementation layer. An affordance structured analysis grounds an A.I. system's symbols by appealing to stimulus information, as opposed to the traditionalist appeal to causal energy connections. More precisely, the theory of affordances and direct perception enables a system to implement behavioral modifications through opportunities for actions specified in optic arrays. An alternate grounding mechanism must fulfill the three central tenets of symbol grounding theory: meaningful perception, purposeful action, and environment dependent symbols. Direct perception and Gibson's ecological approach forwards a dynamic model of action grounded in affordances that satisfies these three criteria. The symbol grounding problem is perspective dependent and is in nature, only a theoretical conflict.

References

Harnad (1990). The Symbol Grounding Problem. *Physica*, D 42: 335-346.

Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman and Co.

Flexible Use of Prospective and Retrospective Memories

Horatiu Voicu (hv2@duke.edu)
Department of Psychological and Brain Sciences, 9 Flowers Dr
Durham, NC 27708 USA

Experimental data show that when animals are interrupted while executing a task, they optimize the use of prospective memory and retrospective memory in order to improve the performance of the task. This work describes a computational model of information retrieval from prospective and retrospective memories. The model includes a mechanism of memory load optimization that selects which memory participates in the decision making process. Computer simulations show that the model produces behavior similar to that found in rats and pigeons. Preliminary experimental results concerning memory load optimization in humans show a similar behavioral pattern.

Imagine that you go to the grocery store to buy 12 items for a special recipe you want to prepare. While you gather the items in your cart you meet a friend that asks you to help jumpstart her car because she has a discharged battery. After you help your friend you return to the store. Because you already spent too much time with your friend and your cart is located far away from the entry point of the store, you decide to take a new cart and pick up the remaining groceries. Then, you go to the location of the first cart that you used and notice that you have the complete list of groceries. What are the variables that influenced your behavior? Two important variables are the amount of time you spent with your friend and the number of groceries gathered until the point of interruption. Other factors are the degree of familiarity with the grocery store and the 12 items you intended to buy. Then, an important question is whether the number of groceries gathered until the point of interruption and the duration of time spent with your friend have an influence on holding a complete list of groceries (no duplicates or missing items) at the time you reach the first cart.

Answers to similar questions have been provided in studies with animals. For example, Cook et al. (1985) designed a study to find what type of encoding rats use to solve a 12 arm radial maze. Subjects were trained to find food in the maze so that they have an internal representation of the maze. At the beginning of each trial all the arms of the maze are baited. Rats are placed in the maze and allowed to explore 2, 4, 6, 8, or 10 arms before they are removed from the maze. After a certain amount of time has elapsed they are placed back in the maze and allowed to explore the maze until all food is collected. The number of errors committed by the subjects during one trial measures their performance.

The results show that subjects can shift their coding strategy to optimize their memory load. An interruption after the 6th arm produces an error larger than that generated by an interruption close to either the beginning or ending of the task. Similar results have been obtained with pigeons (Zentall et al., 1990) in an analog task of the radial arm maze.

This work introduces a computational model that describes how animals can achieve memory load optimization. The model contains a retrospective memory, a prospective memory and a mechanism for deciding which memory is used for producing action.

Figure 1 shows the performance of the model when the point of interruption and the delay are varied. Figure 2 shows experimental data for interruptions of 15 and 60 minutes. These data are similar to those presented in figure 1 (see delay 8 and 25). Both theoretical and experimental studies suggest that while short delays affect only the performance interrupted halfway, long delays affect any performance interrupted after the 6th arm.

The computer simulations suggest that prospective memory decays faster than retrospective memory. This might happen because prospective memory is not used as often as retrospective memory. Prospective memory can be used only when a complete representation of the task exists in memory.

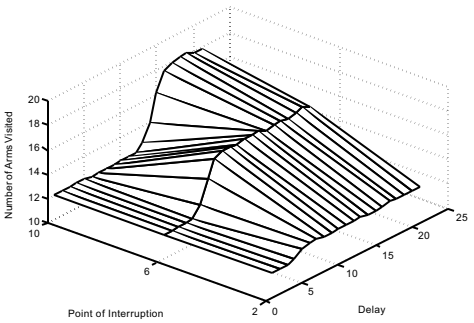


Figure 1. Performance of the model when exploration is interrupted after the 2nd, 6th and the 10th arm while the delay increases.

A B

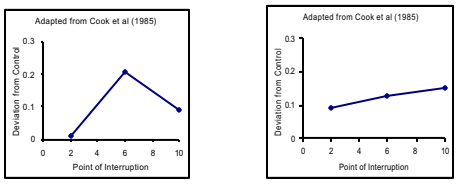


Figure 2. Experimental data that show the performance of rats when delay is 15 minutes (A) and 60 minutes (B).

Cook, R.G., Brown, M.F. and Riley, D.A. (1985). *Journal of Experimental Psychology,* 11, 653-469.

Zentall, T.R., Steirn, J.N., and Jackson-Smith, P. (1990). *Journal of Experimental Psychology,* Vol. 16, No. 4, 358-371.

Motivational Patterns During Hypermedia Learning

Regina Vollmeyer (vollmeye@rz.uni-potsdam.de)
Falko Rheinberg (rheinberg@rz.uni-potsdam.de)
Institut für Psychologie, Universität Potsdam, Postfach 601553
14415 Potsdam, Germany

Bruce D. Burns (burnsbr@pilot.msu.edu)
Department of Psychology, Michigan State University
East Lansing, MI 48824-1117

Our aim was to test Vollmeyer and Rheinberg's cognitive-motivational process model (2000) in the context of learning a hypermedia program. The model assumes that initial motivation affects learning through the mediating variables *motivational state* during learning and *strategy* used for learning. Initial motivation contains four factors: *probability of success*, i.e., learners take into account their ability and the perceived difficulty of the task; *anxiety,* i.e., learners think about failing the task; *challenge,* i.e., whether learners accept the situation as an achievement situation; and *interest*, i.e., whether the topic of the learning material is important.

Method

The hypermedia program described the outbreak of World War I (Vollmeyer & Burns, 2002) on 51 pages. The 42 participants read the instructions saying that after about 25 minutes of learning they will answer a questionnaire. Before starting they answered the QCM (*Q*uestionnaire of *C*urrent *M*otivation; Rheinberg, Vollmeyer, & Burns, 2001) measuring anxiety, probability of success, interest, and challenge. When working with the program we measured two mediating variables: (1) the motivational state (three items e.g., "The task is fun"), and (2) strategy as time per page. The dependent variable was knowledge assessed with 34 questions. To take speed as well as accuracy into account we calculated knowledge as the product of correct answers in the knowledge test and number of pages looked at.

Results and Interpretation

The four factors of initial motivation have some intercorrelation (Rheinberg et al., 2001), so we looked for common patterns in initial motivation through a hierarchical cluster analysis. We found 2 groups: In the first group interest and probability of success were significantly higher, in the second group anxiety and challenge (see Table 1). Thus we interpreted the first group as interested and believing in success (I/P), the second as challenged, but fearing failure (A/C). Table 1 shows how differently these two groups learned the program. The I/P-group had a higher motivational state during learning, they looked at more pages and thus spent less time per page. Knowledge, measured as the product of number of pages * correct answers, was also higher for the I/P-group.

Table 1: Means for I/P- (n = 20) and A/C-group (n = 22).

	I/P	A/C	p
interest	5.34	4.58	.013
probability of success	5.94	4.74	.001
challenge	4.18	4.69	.065
anxiety	1.54	3.35	.001
motivational state	5.60	4.46	.004
time per page	30.55	38.81	.013
number of pages	43.70	35.91	.008
correct answers	15.05	12.86	.082
number * correct answers	653.00	460.82	.002

To test if initial motivation affects learning through the motivational state and strategy we calculated regression analyses. First we found that initial motivation correlated with motivational state, r = .44, time per page, r = -.38, and knowledge, r = .46. The mediating variables also correlated with knowledge (motivational state: r = .37, time per page: r = -.58). However, in a regression analysis, in which the predictor (I/P-, A/C-group) is controlled, only time per page is significant, β = -.46, t = 3.52, p = 0.001. In conclusion, we found that initial motivation affected learning only via strategy: More anxious and challenged learners spent more time per page but they learned less of the page's content compared to success-oriented and interested learners.

Acknowledgments

This research was supported by the DFG (German Research Foundation, Vo 514/10 to Vollmeyer and Rheinberg).

References

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47,* 57-66.

Vollmeyer, R., & Burns, B. D. (2002). Goal specificity and learning with a hypermedia program. *Experimental Psychology, 49*.

Vollmeyer, R., & Rheinberg, F. (2000). Does motivation affect performance via persistence? *Learning and Instruction, 10,* 293-309.

Preschool Children's Use of Auditory Information in Drawing Inferences about Animal Kinds

Winnie H.K. Wai (h9910031@hkusua.hku.hk)
Benise S.K. Mak (benise@hku.hk)
Department of Psychology, The University of Hong Kong,
Pokfulam Road, Hong Kong, PR China

Infants are born with five senses: sight, hearing, taste, smell and touch, through which they come to understand the world around them and interact with the environment effectively. However, the role of visual information in early conceptual development has been stressed. In this paper, we would argue that auditory information should also play a part. For instance, the sound of an animal can be one of the essential cues for us to determine what the animal is. This study, therefore, was an attempt to examine the extent to which preschool children use auditory information in drawing inductive inferences about animals.

In a recent study by Wong (dissertation), 4-year-old children's use of auditory and visual information has been compared. They were found to be more likely to draw inferences about natural kinds based on the similarity of auditory information than on shape similarity. The importance of auditory information has been shown.

However, some would argue against the importance of perceptual information no matter whether it is visual or auditory. A series of studies by Gelman et al. (e.g, Gelman & Coley, 1990) have shown that children as young as 3 years of age were able to go beyond perceptual cues and use conceptual information, such as category labels, to make judgments. They tended to rely more on labels than on perceptual appearance to draw inferences about animals.

Despite of these findings, the role of category labels has been questioned. A study by Mak, Vera and Lo (under preparation) has shown that young children's use of labels seems to be rather limited. Preschool children tended to use motion information more often than category labels to make categorical judgments.

Following this line of argument together with the evidence that infants are not only sensitive to sound but also show good ability in sound detection, discrimination and localization (e.g., Leventhal & Lipsitt, 1964; Wormith, Pankhurst, & Moffitt, 1975), it is reasonable to believe that preschool children would be more likely to use the sound of animals than category labels to draw inductive inferences.

The present study was a 2 (similar & different sound) x 2 (similar & different label) between-subject factorial design. Two hundred and forty 4-year-old children participated. In the experiment, each child was tested individually. Children were presented with 2 pairs of animal stimuli ("cat" and "dog"), one pair at a time (a target and a test animals which shared similar appearance). The children were first taught a new property about the target and were then asked to infer if the property was also true for the test animal.

Results provide supports for our hypothesis, showing that 4-year-olds tended to use auditory information significantly more often than verbal labels to draw inductive inferences about the animal stimuli. This finding is consistent with those in Mak et al.'s study, suggesting that the role of category labels may not be as important as Gelman et al. have suggested and that of auditory information cannot be ignored.

Although children's use of perceptual and conceptual information has been compared in this study, we are not suggesting that they play distinct roles in children's conceptual development. On the contrary, we do believe that information of shape, sound and labels interact to guide children's categorical judgment. This, however, remains to be determined in future studies.

References

Gelman, S.A., Coley, J.D. (1990). The important of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, 26 (5). 796-804.

Leventhal, A., & Lipsitt, L.P. (1964). Adaptation, pitch discrimination and sound localization in the neonate. Child Development, 35, 74-89.

Mak, B.S.K., Vera, A.H. & L.Y., Lo (under preparation). The role of category labels in 3-year-old children's inferences about animal kinds.

Wong, J.T.H. (1999). Cognitive Development: The effect of sound on categorization tasks. Dissertation in the University of Hong Kong, Psychology Development.

Wormith, S.J., Pankhurst, D., & Moffitt, A.R. (1975). Frequency discrimination by young infants. Child Development, 46, 272-275.

If Only I Had Acted Differently: Reasons and Actions in Counterfactual Thinking

Clare R. Walsh (cwalsh@tcd.ie)
and Ruth M.J. Byrne (rmbyrne@tcd.ie)
Psychology Department, University of Dublin, Trinity College, Dublin 2, Ireland

Counterfactual Thinking

Suppose an action of yours leads to a bad outcome. You are plagued by thoughts of 'if only I hadn't acted'. But suppose you are reminded of a very good reason why you acted. Will the reason diminish your tendency to think 'if only' about your action? Our aim is to report experimental results on how the *reasons* for actions can influence 'if only' thoughts.

Generating counterfactual thoughts about what might have been may be central and pervasive in human cognition (e.g., Byrne & McEleney, 2000). Counterfactual thoughts follow certain regularities. Most importantly, for our purposes, people think 'if only' about controllable events (e.g., stopping for a beer) rather than uncontrollable ones (e.g., sheep crossing the road) (Girotto, Legrenzi & Rizzo, 1991; McCloy & Byrne, 2000). The focus on controllable actions has been demonstrated repeatedly in real life as well as laboratory studies. The ability to imagine that a person could have acted differently may be central to our concepts of freedom and responsibility and may underlie emotions such as guilt and regret.

Why do people think 'if only' about controllable actions? The answer may be that they are perceived to be independent of external causes (Girotto et al., 1991). But most actions depend on a *reason*. We examine how reasons for acting can influence 'if only' thoughts.

There are many different sorts of reasons (Walsh & Byrne, 2002). Reasoners may view some reasons as necessary, and may imagine that without the reason the action would not have occurred. Other reasons may be viewed as non-necessary and reasoners may imagine alternative reasons for acting. We suggest necessary reasons are generally *enduring*, oriented towards longer term plans whereas non-necessary reasons tend to be *immediate*, satisfying current desires and short-term goals. We expected that reasoners' would generate fewer 'if only' thoughts about actions for which there were necessary reasons compared to non-necessary reasons or no reason.

Imagining Counterfactual Alternatives

In one experiment, we constructed three versions of a scenario about an individual, Tom, who is delayed by several events on his way home from work, only to find he is too late to save his dying wife. In one version the action (going to the gym) was preceded by a non-necessary reason, an immediate desire to act 'for its own sake' (Tom really likes to go to the gym'); in a second, the action was preceded by a necessary reason, an enduring long-term plan ('Tom is trying hard to lose weight'), and in the third version no reason was given.

We assigned 194 students to one of the three groups. They listed four completions for the following counterfactual thought:

> As commonly happens in such situations, Tom often thought, "if only..."

The results showed that a necessary reason reduced participants' tendency to think 'if only he hadn't acted' (75%) compared to a non-necessary reason (89%, chi^2 = 3.86, $p < .025$), and compared to no reason (88%, chi^2 = 3.32, $p = .03$). A necessary reason shifts some 'if only' thoughts from the action to the reason instead. The possibility in which the reason does not occur and the action occurs anyway is ruled out and so a counterfactual can be generated of the form, "if only the reason had not happened, the action would not have happened." In contrast, for non-necessary reasons there is the possibility that the reason does not occur and the action occurs anyway and so a counterfactual cannot be generated of the form, "if only the reason had not happened, the action would not have happened."

The focus of counterfactual thoughts on controllable actions may arise in part because controllable actions seem to be independent of any external constraint. Once such constraints are made apparent, in the guise of the provision of reasons for acting, the tendency to think 'if only' about controllable actions is reduced.

Acknowledgements

The research was supported by Enterprise Ireland, the Irish Research Council for the Humanities and Social Sciences, and Dublin University.

References

Byrne, R. M. J. & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 26, 1318-1331.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica, 78*, 111-133.

McCloy, R. & Byrne, R.M.J. (2000). Counterfactual thinking about controllable actions. Memory and Cognition, 28, 1071-1078.

Walsh, C.R. & Byrne, R.M.J. (2002). Counterfactual Thinking about Reasons for Acting. *In submission.*

The Interaction Effect of Medium and Pedagogy on Semantic Knowledge Structure

Alex LI Wang-on (h9722776@hkusua.hku.hk)
Department of Psychology, The University of Hong Kong

John A. SPINKS (spinks@hku.hk)
Department of Psychology, The University of Hong Kong

1. Background

Enthusiastic educators digitize different sorts of learning materials in the hope of enhancing learning outcome. The learning outcomes from this trend, however, remain inconclusive because detailed evaluation is often lacking.

One difference between the traditional printed medium and the digital medium is the arrangement of presentation of the concepts to be learnt. Digital medium allows the concepts to be arranged in a networked structure with the aid of hyperlinks which is not possible in traditional printed medium. The advantage of networked structures lies in encouraging the active exploration of information in the absence of any predefined structure. This encourages an active knowledge construction process, which is crucial in learning. A concomitant disadvantage is that learners may get easily lost in the network (Dix, Finlay, Abowd & Beale, 1998).

The digital medium may therefore complement the strengths of problem-based learning pedagogy, as they share the same objective in encouraging the learners to explore. The disadvantage of the digital medium may on the other hand be overcome through providing learners with browsing objectives, such as instructions to solve particular structured problems within a PBL learning framework.

2. Method and Results

To assess the interaction effects of problem-based learning and digital medium, techniques in learning outcomes assessment were further explored. These evaluated the learners' semantic differentiation with the help of multidimensional scaling. Chi, Feltovich and Glaser (1981) argued that experience level determines the differentiation of a problem. A number of other researchers have also suggested that expertise affects concept differentiation, experts tending to differentiate concepts more finely and more systematically (*c.f.* Fisher, 2000).

2.1 Construction of Expert Model

A group of research postgraduate students with biological psychology background was asked to make pair-wise difference comparisons on some biological psychology terms, which constitute the dissimilarity matrices. ALSCAL (Alternating Least squares SCALing) solutions of them was constructed, and regarded as the experts' model.

2.2 Pedagogy and medium interaction effect

A 2x2, medium (digital, printed) x pedagogy (directed learning, problem-based learning), research design was run in an authentic biological psychology learning environment. Dissimilarity matrices data were collected after the learning sessions. Preliminary data analysis suggested that multidimensional scaling solutions were problematical in terms of the groups, because of the large variance within group. Individual ALSCAL solutions were, therefore, constructed, from which parameters were derived for further analysis.

2.2.1 Tuncker's Congruence Coefficient

Tuncker's congruence coefficients between individuals' and the experts' dissimilarity matrices were calculated based on the first stage results. A 2x2 ANOVA (pedagogy x medium) showed that there is a significant pedagogy main effect ($F = 17.414$, $p < 0.01$). The knowledge differentiation criteria by the individual in the directed learning group are more similar to the experts' group model than to the problem-based learning group.

2.2.2 Further analysis

Further analyses using parameters from a cluster analysis and property vector fitting revealed no significant difference between groups.

3. Discussion

The results indicated that the effect of medium shift is not as dramatic as the effect of pedagogy application. Educators should consider putting more effort into implementing useful pedagogy rather than digitizing learning materials alone.

The pedagogies investigated in this study, through directed learning, which mimics the nature of traditional teaching, and problem-based learning, yield different semantic knowledge structures. This mirrors the objective of problem-based learning, in that it leads the learners to construct their own understanding. The individual semantic models that are derived as outcomes from this process are, therefore, less congruent with those of the experts. Whether this is beneficial or not depends on the teaching objective.

Finally, this research only explores one area of digital medium based instruction. The effects of more sophisticated digital functions, for example animation, intelligent tutor and online-collaboration, should be further explored.

4. References

Chi, M. T. H., Feltovich, P. J. & Glaser, R., (1981). "Categorization and Representation of Physics Problems by Experts and Novices." *Cognitive Science*, pp.121-152.

Dix, A., Finaly, J., Abowd, G. & Beale, R. (1998). *Human-computer interaction*. Essex: Prentice Hall Europe.

Fisher, K. M. (2000). "SemNet Software as an Assessment Tool." In Mintzes J. J., Wandersee J. M. & Novak J. D. (Eds). *Assessing Science Understanding: A Human Constructivist View.* Acad. Press..

The Neural Instantiation of Number

John W. Whalen (whalen@udel.edu)
Frank Morelli (fmorelli@udel.edu)
Department of Psychology, University of Delaware
Newark, DE 19716 USA

Introduction

Fundamental to calculation and arithmetic competency is the ability to abstractly represent numerical quantity. While much is known about the psychophysics of human quantity representation, little is known about the neural instantiation of this key ability. Through ERP (event-related potentials), we characterize the nature of multidigit numerical representations. Dehaene (1996) revealed that bilateral regions of parietal cortex are involved in judging the largest of single digit numbers (1,4,6, and 9) from the number 5. Our research replicates his findings and also considers the nature of the neural representation of the magnitudes themselves. Preliminary studies have revealed systematic variations in ERP signature in response to the presentation of small numerical quantities (range: 1 – 16). This activation is localized to bilateral inferior parietal regions and occurs 220 ms after stimulus onset (Whalen, West, & Cook, 2002). The present work investigates the neural representation of larger quantities, and the mapping of multidigit numerals to neural quantity representations.

Methodology

Participants were shown Arabic digits ranging from 0 to 99. Stimuli were randomized and presented individually for 500 ms. Because Arabic numerals are known to automatically elicit representations of numerical magnitude (Lefevre et al., 1988; Naccache & Dehaene, 2001), no response was required of participants. Participants were also periodically asked to compare the relative numerical magnitude of two sequentially presented numbers (participants judged whether the second number in the pair was “smaller” or “larger” than the first). Event related potentials were recorded using a 128 channel EGI (Electrical Geodesics, Inc.) Sensor Net. ERPs were collected for each tens quantity (e.g, 3 of 34), units quantity (e.g, 4 of 34), and overall magnitude (e.g., 34) for 250 ms prior to stimulus onset to 500 ms post onset.

Results

Using single and double-digit Arabic numerals ranging from 0 to 99, we discovered functionally distinct neural representations for the individual numerals that compose a multidigit number (e.g., the "3" and "4" of the multidigit numeral "34") and for overall numerical magnitude. Regression analysis over 20 ms intervals at each electrode site revealed systematic linear variation in neural voltages relative to the numeral presented. This included distinct representations for both the tens and units quantities, and the overall magnitude. Overall magnitude was represented bilaterally in inferior parietal regions, while tens and units representations were localized to the right superior parietal gyrus. While the ERP signature for each quantity was localized to unique regions of parietal cortex, the onset of the representations were nearly simultaneous, commencing approximately 220 ms after presentation of the numeral. The linear changes in ERP signatures relative to the magnitude presented suggests that number is represented linearly in parietal cortex, a finding consistent with studies of quantity representations in non-humans (Gallistel & Gelman, 2000).

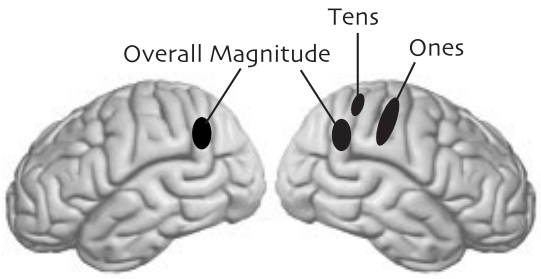


Figure 1: Source Localization for Tens, Units and Overall Magnitudes.

Acknowledgments

This research was supported by NSF-MRI Award Number 9977628.

References

Dehaene, S. (1996). The organization of brain activations in number comparison: Event-related potentials and the additive-factors method. *Journal of Cognitive Neuroscience, 8*(1), 47-68.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: from reals to integers. *Trends in Cognitive Sciences, 4*(2), 59-65.

Lefevre, J., Bisanz, J., & Mrkonjic, L. (1988). Evidence for obligatory activation of arithmetic facts. *Memory & Cognition*, 16, 45-53.

Naccache, L., & Dehaene, S. (2001). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex*, 11(10), 966-974.

Whalen, J., West, V., & Cook, B. (2002). Why you shouldn't count on subitizing: Evidence from estimation and counting. *Manuscript submitted for publication*.

Partial Analogical Transfer in Problem Solving: Roles of Centrality and Order

Tsunhin J. Wong (thjwong@hkusua.hku.hk)

Albert W. L. Chau (awlchau@hkucc.hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong

Antonietti (1991) first presented the idea of partial analogies in solving an ill-defined problem. Prior to that, studies looking into the role of analogies in problem solving used complete analogies. Antonietti concluded that partial analogies helped problem solving only when all the cues were present and presented in the correct order.

The conditions for partial analogy to work as identified by Antonietti seem to be overly stringent. Partial analogy should be effective in most situations as most analogical cues available in daily life are partial in nature. This study therefore looks into the role of analogical transfer of partial analogies.

Our study differs from Antonietti (1991) in a few ways. First, as it is logical to assume that some analogical cues are more crucial than the others, the notion of centrality of an analogy was examined. Second, we also revisited if analogical cues have to be presented in the exact order in order to be effective. Finally, we presented the partial analogies as problems for participants to solve rather than disguised them as arithmetic problems.

Method

Forty undergraduates at the University of Hong Kong participated in the experiment as part of a course requirement. None of them had been exposed to the problems used in the study.

Every participant completed the experiment on a computer. The problems were written in Flash with both texts and diagrams.

Procedure

Two types of problems were used in the experiment. The analogy problems were concerned about how to direct water to a target location. They were used to prime participants to use the two strategies which are necessary for solving the target problems: divergence which is dividing the flow to avoid overload and convergence which is pulling together the divided flows to achieve the intensity needed. The target problems were the Fortress problem and Duncker's radiation problem. The first problem deals with how to organize soldiers to conquer and castle while the other deals with how to use X-ray to destroy a tumor.

Each participant was first told some basic concepts subjects in fluid dynamics. S/he then proceeded to solve one of the four versions of the analogy problem depending on the experimental condition to which s/he was assigned: i) *partial analogies* presented in the *correct* order (divergence then convergence), ii) *partial analogies* presented in the *reverse* order (convergence then divergence), iii) *complete* analogy; and iv) *unrelated* analogy. Finally the participant was asked to solve the two target problems. If s/he could not solve a problem within 5 minutes, s/he was then told that the problems they had tackled earlier might help them. A total of 10 minutes were allowed to solve each problem.

Results and Discussion

43.9% of the subjects solved the Duncker's problem without being prompted to use the previous problems, with the highest in the *partial-correct* condition (81.8%) and the lowest in the *unrelated* condition (20%).

Solving the Fortress problem in the *unrelated* condition required more time than in all the other conditions. Time required to solve the Duncker's problem was in the decreasing order of: *partial-correct, partial-reverse, complete, and unrelated.* Besides, the *partial-correct* condition required the least prompting for using the partial analogy condition.

Contrary to Antonietti's study, participants in the *partial-reverse* condition were able to solve the target problems spontaneously. In other words, exposure to partial analogies is sufficient for priming to occur though the exact temporal order can provide additional facilitation. Second, it was also found that partial analogies were more effective than *complete* analogy in priming the participant to solve the target problem. This is possibly due to the complexity involved in solving the complete analogy problem. Third, the present findings suggested that convergence plays a more central role in analogical transfer. This is in line with Pedone, et al (2001) who found that convergence alone was sufficient to trigger spontaneous analogical transfer. The second and third findings together suggested that convergence is the more central partial analogy. Only it should be presented to achieve the greatest priming or facilitatory effect in problem solving. Presenting a less useful partial analogy (divergence) lowers instead of enhances the transfer in problem solving performance. This may be explained by introducing a weight system in the branches of the structural mapping theory (Gentner, 1983) or constraints in the multiconstraint theory (Holyoak & Thagard, 1980).

References

Antonietti, A. (1991). Effects of partial analogies on solving an ill-defined problem. *Psychological Reports, 68*, 947-60.

Pedone, R., Hummel, J. E., & Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. *Memory & Cognition, 29(2)*, 214-221.

Gentner, D. (1983). Structural-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Holyoak, K. J., & Thagard, P. (1980). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Mental Metalogic and its Initial Empirical Justifications: The Case of Reasoning with Quantifiers and Monadic Predicates

Yingrui Yang and Selmer Bringsjord
yangyri@rpi.edu • selmer@rpi.edu
Department of Cognitive Sciences
Rensselaer Polytechnic Institute, Troy, NY 12180 USA

In psychology of human deductive reasoning, mental logic theory claims that people reason by applying inference schemas (e.g., Braine & O'Brien, 1998; Rips, 1994), and mental models theory claims that people reason by constructing mental models (e.g., Johnson-Laird and Byrne, 1991). There is a great deal of empirical evidence supporting each theory. The authors have proposed a mental metalogic theory (Yang & Bringsjord, 2001) studying the interactions between applying inference schemas and constructing mental models based on the current theories of mental logic and mental models. We report a set of experiments designed to examine possible interactions of this kind. Mental metalogic suggests ways of modeling reasoning strategies.

Our strategy for constructing experimental problems was to integrate one problem type used in mental logic research (Yang, et al. 1998) and another problem type used in mental model research (Yang & Johnson-Laird, 2000). Below is a resulting sample problem used the experiment.

> The premises given below are either all true or all false:
> All the beads are wooden or metal.
> The wooden beads are red.
> The metal beads are green.
> The square beads are not red.
> Is possible that the square beads are green?

This new problem type can be used to manipulate two independent variables. The first independent variable is about the set of premises. For a given problem, it can have the set of original premises, or the denials of these premises. The second independent variable is how a question is presented. It can take the form, "Is it possible that …" or "Does it necessarily follow that …". Thus, by manipulating these two independent variables, four types for a given problem are produced. The first experiment used a 2x2 between-subjects design to manipulate two independent variables in four conditions according to the 4 problem types explained above. 18 original multi-step problems similar to the example above were carefully selected from Yang, et al. Their task was to choose among the given responses (i.e., Yes, No, or Can't tell). The mean accuracy for the original/necessity problems was 45.5%, for the original/possibility problems 91%, for the denials/necessity problems 83%, and for the denials/possibility problems 60%. (N=40 for each problem type). The results are clear-cut. For the problems using original premises, the problem type of possibility was evaluated significantly more accurately than the problem type of necessity (Mann-Whitney $Uz = 5.17$, $p < .001$). For the problems using the denials of the original premises, the problem type of necessity was evaluated significantly more accurately than the problem type of possibility (Mann-Whitney $Uz = 5.14$, $p < .001$). In addition, there was a reliable interaction. The difference between problem types of necessity and possibility for the problems using original premises was greater than for the problems using the denials of the original premises (Mann-Whitney $U = 44$, $p < 0.01$). The similar results were obtained from a second set of experiments using dyadic predicate problems parallel to the monadic predicate problems used in the first experiment A 2x2 within-subjects design was used (N=140, individually tested). This time the latency data were also collected, and the results showed that an answer took significantly longer time when two cases (both "all true" and "all false" situations) needed to be considered than when only one case (i.e., "all the premises are true") needed to be considered. For the problems with original premises and necessity questions, 55% subjects answered yes, which was an illusion because they failed to consider the "all false" case. But they could apply inference schemas in the local situation of "all true". However, another fairly large portion of participants (45%) responded "No" to the problems of this type, and would have needed to consider the "all-false" case, which took longer time. In this local situation, there are no inference schemas currently available to deal with the denials of the original premises, and reasoners may likely construct mental models.

There are long-standing controversies between mental logic and mental model theories, as well as other emerging controversies between the mental logic/model paradigm, mental metalogic, and other approaches in reasoning. Deduction is core to human cognition. These issues deserve open discussions and debates, which have been the ways for different theories to grow in this field.

(Note. References are available upon request.)

"If" is Easier Than "Or" in the GRE

Yingrui Yang (yangyri@rpi.edu)
Department of Cognitive Sciences, Rensselaer Polytechnic Institute
Carnegie Hall, Troy, NY 12180 USA

Philip N. Johnson-Laird (phil@princeton.edu)
Psychology Department, Princeton University
Princeton, NJ 08544 USA

In logic, a conditional, such as: "If the trend continues then a decline will occur" is equivalent to a disjunction: "Either the trend stops, i.e., doesn't continue, or a decline will occur". Both assertions are compatible with the following three possibilities, where "¬" denotes negation:

Trend Decline
¬ Trend Decline
¬ Trend ¬ Decline

The equivalence may break down as a result of the specific content or context of assertions (Johnson-Laird and Byrne, 2002). But, where the two assertions are equivalent, the conditional has the mental models:

Trend Decline
. . .

in which the first model represents the possibility in which the antecedent is true, and the second wholly implicit model represents the possibilities in which the consequent is false. The disjunction has the mental models:

¬ Trend
Decline
¬ Trend Decline

It follows that reasoning should be easier with the conditional than with the disjunction.

We conducted three experiments to test this prediction using "logical reasoning" problems from the Graduate Record Examination (the GRE, devised by Educational Testing Services, Princeton. In Experiment 1, 20 participants carried out either a conditional version or a disjunctive version of 8 GRE problems, e.g.: the conditional version:

Because the number of surgeons is growing faster than the number of operations and because noninvasive medical therapies are increasingly replacing surgery, the average annual number of operations per surgeon has fallen by one-fourth in recent years. It can be concluded that, if these trends continue, a dangerous decline in the level of surgical skill will occur.

The argument is based on which of the following assumptions?

(A) Surgeons now spend a large percentage of their time performing noninvasive medical procedures.

(B) A surgeon's skill cannot be properly maintained unless the surgeon performs operations with a certain minimum frequency.

Option (B) is the correct answer. The disjunctive version included instead the following final assertion:

It can be concluded that, either these trends stop, or a dangerous decline in the level of surgical skill will occur.

The participants had to select the correct response from the pair of assertions, which were the correct conclusion and the most frequently chosen foil (according to ETS). The accuracy of responses did not differ, but the participants were reliably faster to solve the conditional problems (mean 1.76 min.) than the disjunctive problems (mean 2.06 min.).

Experiment 2 was a replication but in which the two response options were conditionals (for the conditional problems) and disjunctions (for the disjunctive problems). The participants were reliably more accurate and faster with the conditional problems (73% correct, 0.8 min.) than with the disjunctive problems (61% correct, 1.17 minutes). The use of a sentence containing a given connective in both the text and the two response options evidently amplified the difference between conditionals and disjunctions.

Experiment 3 used conditional and disjunctive texts with conditional and disjunctive response options in all four combinations. The results showed that the nature of the response options was decisive. The 40 participants were faster and more accurate with problems that had conditional responses than with problems that had disjunctive response options.

We conclude that the model theory's predictions about the different representations of conditionals and disjunctions extend to realistic problems based on the GRE. Theories based on formal rules of inference (e.g., Braine and O'Brien, 1998; Rips, 1994) make no predictions about this difference.

The research was a part of the project, eWriter, which was supported by a grant from ETS and the GRE Board to S. Bringsjor, Y. Yang, P.N. Johnson-Laird, and M. Bauer.

References

Braine, M.D.S., and O'Brien, D.P., Eds. (1998) Mental Logic, Mahwah, NJ: Erlbaum.

Johnson-Laird, P.N., and Byrne, R.M.J. (2002) Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, in press.

Rips, L.J. (1994). The Psychology of Proof. Cambridge, MA: MIT Press.

A Computerized Lexical Database of Cantonese

Michael C. W. YIP

School of Arts & Social Sciences, The Open University of Hong Kong

myip@ouhk.edu.hk

Introduction

Lexical databases are now available for many languages over the world, for example, CELEX and Euro WordNet. However, there are not any comparable data of this kind for Cantonese. The necessity of this database cannot be underestimated since not much progress on the Chinese (Cantonese) psycholinguistic research can be made without the succor of these kinds of precise lexical information (Li & Yip, 1996, 1998; Yip, 2000). This project aims at making a large-scale collection of digital tape recordings of Cantonese speech and establishing an archive of Cantonese texts based on transcriptions of these recordings. A corpus of Cantonese syllables and words together with other polysyllabic Chinese expressions will be constructed. In the database, much useful lexical information can be generated: for example, spoken word frequency, frequency information of phoneme occurrence and phoneme co-occurrences, speech errors. In this project, a large-scale lexical database of Cantonese will be established in two phases. In the first phase, the target is to have a database of 300,000 Cantonese words. In the second phase, the target will be up to one million Cantonese words. It is hoped that the subsequent psycholinguistics research in the Chinese language can benefit from this computerized lexical database.

Main Objectives

Three main objectives of the project here are:

(1) Generation of relevant lexical information of Cantonese Chinese speech: the database can generate such kinds of useful lexical information for psycholinguistic research as (a) the frequency information of spoken Cantonese word (cf. Yip, 2001); (b) the probabilistic phonotactic information of Cantonese speech (Yip, 2000)

(2) Determination of the processing and production unit of Cantonese speech: from the data of speech error collected in the database, we can closely monitor if the speech error of Cantonese involved a whole syllable replacement or other sub-syllabic components interchanging (cf. Chen, 2000), and then inferred to the functional units of Cantonese speech (Chen & Yip, 2001; Yip, Song, & Chen 1999)

(3) Estimation of the code-switched situation in Hong Kong: from the database, we can estimate the size of code-switching and types of code-switchers in the bilingual situation of Hong Kong (Chan, 1992)

Methodology

This project is designed to construct a computerized lexical database of Cantonese. The database can be used to generate several different kinds of lexical information of Cantonese natural speech. It is based on the large-scale collection of digital tape recordings of natural Cantonese speech. Sources of the natural Cantonese speech include dialogues of Radio call-in programs (Chen, 2000), conversations of TV programs, casual chatting among the students in canteen. Collecting the Cantonese speech from different sources of naturalistic settings guaranteed the ecological validity of the lexical information generated from the database. Because the data gathered to the database is entirely came from the real and natural cases which obviously are psychologically real as well as can reflect the lexical information embedded in our mental lexicon.

Expected Results

The result of this project will be summarized in a computerized lexical database of Cantonese that is significant to research as well as to language teaching and learning. In terms of research, we believe that a more solid rigorous set of lexical information of Cantonese speech can be derived and it can have a wide range of applications to linguistic as well as psycholinguistic researches, especially lexical research centering on spoken language processing. In terms of language pedagogy, it will provide empirical ground for designing the most appropriate language learning methods to students according to the patterns of the prominent processing and production units of native Cantonese speakers. Meanwhile, it will also provide useful information of the pervasive code-switching situation in Hong Kong that clearly confounded the traditional language teaching methods in Hong Kong education sector.

References

Chan, H.-S. (1992) *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*. MA thesis, Chinese University of Hong Kong.

Chen, H.-C. & Yip, M. (2001). Processing Syllabic and Sub-syllabic information in Cantonese. *Journal of Psychology in Chinese Societies, 2*, 199-210.

Chen J. -Y. (2000) Syllable errors from Naturalistic Slips of the Tongue in Mandarin Chinese, *Psychologia*, 15-26.

Li, P., & Yip, M. (1996) Lexical Ambiguity and context effects in spoken word recognition: Evidence from Chinese. In G. Cottrell. (ed.). *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society.* (pp. 228-232). Mahwah, New Jersey: Lawrence Erlbaum.

Li, P., & Yip, M. (1998) Context effects and the processing of spoken homophones. *Reading and Writing, 10,* 223-243.

Yip, M. (2000) Recognition of Spoken words in continuous speech: Effects of transitional probability. In B. Yuan, T. Huang, & X. Tang. (Eds.), *Proceedings of the ICSLP'2000,* 758-761. Beijing: China Military Friendship Publish.

Yip, M. (2001) A preliminary study of subjective frequency estimates of spoken-words in Cantonese. *Psychological Reports, 88,* 1253-1258.

Yip, M., Song, H., & Chen, H. -C. (December, 1999) Cognitive Processing of Speech: The case of Chinese. Paper presented at the International Language in Education Conference, Chinese University of Hong Kong.

Author Index

Sylvia Acchione-Noel....703
Woo-kyoung Ahn....590
Anu Airola....566
Eleonora Albano....997
Elizabeth Albro....54
Martha Alibali....59, 661
James Allen....867
Richard Alterman....26
Erik Altmann....65
John Anderson....387, 1029
Janet Andrews....584
Bernard Ans....71
Pablo Arantes....997
Rita Ardito....77
Zippora Arzi-Gonczarowski....989
Ivan Ash....83
Kevin Ashley....268
Michael Atherton....89
Marios Avraamides....28, 95
Ryan Baker....990
Linden Ball....101
Bruno Bara....77
Lawrence Barsalou....30
Renato Basso....997
Christopher Bearman....101
James Bednar....107
Sieghard Beller....113
Maureen Below....884
Franck Tarpin Bernard....626
Marie Bienkowski....23
Enrico Blanzieri....77
Rens Bod....119
Deborah Boehm-Davis....27, 41
Guido Boella....125
Tara Booth....59
Ronald Boring....932
Anne-Louise Bornstein....780
Lera Boroditsky....131, 136, 994
Kristin Branson....238
Sarah Brem....23
Selmer Bringsjord....1054
Andrew Brook....142, 872
Sarah Brown-Schmidt....148
Russell Burnett....774
Bruce Burns....991, 1048
Kevin Burns....154
Michael Byrne....7
Ruth Byrne....160, 1050
Manoel Caetano....992
Jonathan Cagan....1023
Andrew Calder....238
Dustin Calvillo....993
Ellen Campana....148, 867
Christopher Campbell....602
Angelo Cangelosi....33
Stuart Card....13
Richard Carlson....2, 28
John Carroll....29
Daniel Carruth....36
Daniel Casasanto....994
Daniel Cassenti....28
Richard Catrambone....166, 250
Sergio Chaigneau....30
Lama Chandrasena....578
Suzanne Charman....172
Nick Chater....720
Albert Chau....1015, 1053
Alan Chauvin....322
Anthony Chemero....20
Peter Cheng....18, 530
C. M. Chewar....995
Michelene Chi....31, 655, 1001
Christine Chiarello....524
David Chin....999
Seth Chin-Parker....50, 178
Kwangsu Cho....184
Yoonsuck Choe....190
Ronald Chong....4, 21, 41
Yu-Ju Chou....996
Eric Chown....202
Morten Christiansen....220, 596
Timothy Clausner....208
John Clement....32
Orlando Bisacchi Coelho....997
Eliana Colunga....214
Louise Connell....998
Christopher Conway....220
Celestine Cookson....1007
Albert Corbett....990
Andrew Corrigan-Halpern....226
James Corter....1021
Fintan Costello....232
Garrison Cottrell....238
Kimberly Cottrell....244
Kenny Coventry....33
L. Andrew Coward....34
David Latch Craig....250
Valerie Crawford....23
Mathias Creutz....566
Martha Crosby....999
Géry d'Ydewalle....280, 914, 1000
Walter Daelemans....637
Jody Daniels....310
Mehdi Dastani....256
Neil Davey....435
Elizabeth Davis....966
Fabio Del Missier....262
Andreas Demetriou....35, 756
Wim De Neys....914, 1000
Ravi Desai....268
Stephen Deutsch....274
Kristien Dieussaert....280

Melanie Diez 27
David Diller 21
Stephanie Doane 36
Susan Dumais 11
Jeff Elman 24
Randi Engle 1001
Lindsey Engle 286
Carlos Espinel 1002
Zachary Estes 1003
Igor Farkaš 24, 45
Nicholas Fay 441
Aidan Feeney 292
Alex Feinman 26
Anna Fisher 1041
Eric Fleischman 298
Kenneth Forbus 554
Harry Foundalis 304
Donald Franceschetti 37, 708
Edson Françozo 997
Jerry Franke 310
Michael Freed 3, 649
G. Freedman 18
Robert French 71, 316, 322
Daniel Freudenthal 328, 334
Danilo Fum 262
Joachim Funke 1024
Maggie Gale 340
Max Garagnani 345
David Gardiner 292
Simon Garrod 441
Wilson Geisler 11
Silvia Gennari 351
Dedre Gentner 976
Peter Gerjets 798, 810, 004
Robert Gibby 890
Yolanda Gil 357
Alastair Gill 363
Steven Gillis 637
Kevin Gluck 21
Fernand Gobet 328, 334
Susan Goldman 23
Timothy Goldsmith 884
Robert Goldstone 369
Emilio Gomez 750
Avelino Gonzalez 459
Cleotilde Gonzalez 1005
Andrew Gordon 375
Michael Gorman 1006
Sydney Gould 602
Arthur Graesser 23, 37, 708
William Gregory Sakas 786
Thomas Griffiths 381
Stephanie Guerlain 38
Glenn Gunzelmann 387
Frank Guo 393
Prahlad Gupta 39
Todd Gureckis 399
York Hagmayer 405
Ulrike Hahn 411
Wendy Ham 136
Jeffrey Hansberger 27, 41
Andreas Hansson 417
Harlan Harris 423
Anthony Harrison 1007, 1008
Uri Hasson 429, 1009
Jenny Hayes 435
Linli He 482
Patrick Healey 441
Mary Hegart 18
Mary Hegarty 40
Julie Heiser 57, 447
Joshua Hemmerich 453
Amy Henninger 4, 459
Jon Hicks 465
Kazuo Hiraki 548
David Holliway 471
Robert Holt 27, 41
Keith Holyoak 286, 393
Andrew Howes 172, 476
Xiangen Hu 37
Curtis Ikehara 999
Josh Introne 26
Thomas Ioerger 482
Norio Ishii 667
Jesse Itzkowitz 47
Linden J. Ball 340
Brijnesh Jain 488
Jerzy Jarmasz 494
Vikram Jaswal 500
Luke Jerzykiewicz 142
Bonnie John 3, 649
Todd Johnson ..... x, 506, 920, 970
Philip Johnson-Laird 1009, 1055
P. N. Johnson-Laird 845
Gary Jones x
Matt Jones 1010
Randolph Jones x,4, 298
Catholijn Jonker 512
Desmond Jordan 43
Dan Joyce 33
Peter Juslin 518, 714
Natalie Kacinik 524
Ashish Karnavat 708
Yasuhiro Katagiri 896
Irvin Katz 530
David Kaufman 1011
Mark Keane 998, 1020
Christopher Kello x,1014
Alla Keselman 536
Jihie Kim 357
Alexandra Kincannon 1006, 1012
Walter Kintsch 750

Susan Kirschenbaum 18
David Klahr 673, 1044
Stefan Kleinbeck 1004
Alexander Klippel 1017
Kenneth Koedinger 542, 990
Takatsugu Kojima 1013
Janet Kolodner 42
Takanori Komatsu 548
Dave Koons 602
Hevin Korb x
Kenneth Kotovsky 1023
Amy Kruse 1038
Tate Kubose 43
Sven Kuehne 554
Darcie Kunder 1007
Emily Kuschner 584
Takashi Kusumi 1013
Christophe Labiouse 316
David Lagnado 560, 828
Michael Lagoudakis x
Krista Lagus 566
Thomas Landauer 44
Seth Landsman 26
Yiannis Laouris 810
Laura Leach 1014
Christian Lebiere 5, 21
Frank Lee 572
John Lee 441
Michael Lee 578, 685
Paul Lee 57, 1017
Terence Lee 1015
Yuh-shiow Lee 1016
Alan Lesgold 184
Leonardo Lesmo 125
Ping Li 24, 45, 950
Cabral Lima 1042
Han-yu Lin 1016
Alexandre Linhares 1018
John Lipinski 39
Hsi-wen Liu 1019
Kenneth Livingston 584
Lap Yan Lo 620
Deborah Lord 482
Max Louwerse 37
Bradley Love 21, 399
Marsha Lovett 46
Shenghua Luan 47
Christian Luhmann 590
Gary Lupyan 596
Dermot Lynott 1020
Brian MacWhinney 24
Paul Maglio 602, 608
James Magnuson 614
Benise Mak 620, 1015, 1049
Halima Habieb Mammar 626
Denis Mareschal 322
Arthur Markman 1021
Ellen Markman 500
James Marshall 631
Evelyn Martens 637
Amy Masnick 643, 1045
Michael Matessa 3, 649
Teenie Matlock 602
Toshihiko Matsuk 1021
D. Scott McCrickard 995
Daniel McFarlane 310
Mark McGregor 655
Danielle McNamara 244, 726
Nicole McNeil 661
Ryan Mears 890
Martial Mermillod 322
Risto Miikkulainen 107
Sy Miin Chow 196
Kelli Millwood 23
Kazuhisa Miwa 667
Naomi Miyake 48, 1039
Padraic Monaghan x
Stephen Moore 932
Frank Morelli 1052
Bradley Morris 643, 673, 679, 1045
Julie Bauer Morrison 1022
Jarrod Moss 1023
Vincent Müller 762
Victoria Murphy 435
Serban Musca 71
Julien Musolino 744
Ryuichi Nakaike 667
Daniel Navarro 578, 685
Stefani Nellen 1024
Josef Nerb x
Nancy Nersessian 250
John Nesselroade 196
Hansjörg Neth 691
Lars Niklasson 417
Eiji Nishimoto 786
Sourabh Niyogi 697
Ron Noel 703
Kent Norman 1025
Tenaha O'Reilly 726
Jon Oberlander 363, 441, 465
Hidemi Ogasawara 1026
Stellan Ohlsson 226
Takehiko Ohno 1026
Natsuki Oka 548
Brent Olde 37, 708
Andrew Olney 37
Anna-Carin Olsson 518
Henrik Olsson 518, 714
Luca Onnis 720
Daniel Oppenheimer 1027
Klaus Opwis 780
Magda Osman 732

Pierre-yves Oudeyer 738
Thomas Palmeri 590
Anna Papafragou 744
T. Park 1028
Vimla Patel 43, 536, 970, 1011
Philip Pavlik 1029
Stephen Payne 476, 691, 1032
Natalie Person 37
Célia Lúcia Gomes Pessanha 1030
Lorna Peters 435
Richard Pew 21, 274
Julian Pine 328, 334
Raedy Ping 966
Zygmunt Pizlo 1037
David Poeppel 351
Emmanuel Pothos 411
Maridelma Pourbaix 1031
Mercè Prat-Sala 411
Dennis Proffitt 840
José Quesada 750
Paul Quinn 322
Winston R. Sieck 1040
Jeroen Raaijmakers 11
Athanassios Raftopoulos 756, 762
Michael Ramscar 136, 768, 956
Mary Jo Rattermann 20
William Reader 1032
Douglas Reece 459
Bob Rehder 774
John Rehling 1033, 1034
Torsten Reimer 780
Roger Remington 3, 649
Alexander Renkl 49, 816
Russell Revlin 993
Falko Rheinberg 1048
Lynn Richards 33
Juliet Richardson 476
Frank Ritter x
Matthew Roberts 720
Laudino Roces 997
Brian Rogosky 369
Brian Ross 50, 178
Sandrine Rossi 902
Stéphane Rousset 71
Hitomi Saito 667
Dario Salvucci 792
Alexei Samsonovich 1035
William Sandoval 23
Lelyn Saner 1007, 1036
Walter Schaeken 280, 908, 914, 1000
Michael Scheessele 1037
Katharina Scheiter 798, 810, 1004
Ute Schmid 1004
Dylan Schmorrow 1038
Walter Schneider 6
Wolfgang Schoppek 804
Tina Schorr 810
Walter Schroyens 902
Christian Schunn 65, 184, 679, 1007, 1008, 1036
Silke Schworm 816
Sam Scott 822
Priti Shah 18
Lokendra Shastri 51, 345, 926
Richard Shiffrin 9
Richard Shillcock 996
Atsushi Shimojima 896
Hajime Shirouzu 48, 1039
Edward Shortliffe 970
Hua Shu 950
Thomas Shultz 24
Winston Sieck 1010
Peter Slezak 52
Steven Sloman 560, 828
Vladimir Sloutsky 429, 1041
Linda Smith 53, 214, 962, 966
Pamela Smith 435
Adriana Soares 992, 1030, 1031, 1042
Seika Soraku Kyoto 896
Robert Sorkin 47
John Spinks 1051
Christiane Spitzmüller 890
Justin Starren 1011
John Stasko 166
Mark Steedman 834
Jeanine Stefanucci 840
Nancy Stein 54
Eugenia Steingold 845
Lisa Stevenson 28
Mark Steyvers 11, 381
James Stigler 286
Ron Sun 850, 861
Yanlong Sun 856, 890
Masaki Suwa 55
Kentaro Suzuki 548
Mary Swift 867
Niels Taatgen 572
Helena Taelman 637
Michael Tanenhaus 148, 614, 867
Heike Tappe 1017
Yvette Tenney 21
Atsushi Terao 542
Chris Terry 850
Roger Thompson 20
Akifumi Tokosumi 56
Edina Torlakovic 872
Eva Toth 23
J. Gregory Trafton 18, 878
Jan Treur 512
Susan Trickett 878
Lara Triona 1044, 1045
David Trumpower 884
Pamela Tsang 482

Barbara Tversky 57, 447
Ryan Tweney 856, 890
Kazuhiro Ueda 548
Ichiro Umata 896
Jean-Baptiste Van de Henst 902, 908
Leendert van der Torre 256
Matthew Ventura 37
Alonso Vera 3, 620, 649
Niki Verschueren 914
William Vidal 1046
Horatiu Voicu 1047
Regina Vollmeyer 1048
Winnie Wai 1049
Michael Waldmann 405
Clare Walsh 160, 1050
Hongbin Wang 506, 920
Yue Wang 506, 920
Alex Li Wang-on 1051
Edward Wasserman 20
Carter Wendelken 345, 926
Michael Wenger 608
Robert West 932
John Whalen 1052
Jennifer Wiley 23, 83, 453
A .J. Wills 938, 982
Phillip Wolff 944
Tsunhin Wong 1053
Scott Wood 4, 7
David Woods 14
Robert Wray 4, 21
Fritz Wysotzki 488
Xiaoming Xi 530
Jun Xiao 166
Hongbing Xing 950
Yingrui Yang 1054, 1055
Daniel Yarlett 956
J. Frank Yates 1040
Michael Yip 1056
Hanako Yoshida 962, 966
Norikazu Yoshimine 56
Richard Young x
Wayne Zachary 21
Jeffrey Zacks 57
Thomas Zentall 20
Matthew Zettergren 944
Jiajie Zhang 506, 920, 970
Xi Zhang 861
Sergey Zharikov 976
Jan Zwickel 982

George Mason University

Department of Psychology

MS 3F5, 4400 University Dr., Fairfax, VA 22030 Dr. Robert F. Smith, Chair

---

Graduate Programs | Undergraduate Programs | Admission Applications | Course Catalogs

---

Center for Cognitive Development

The GMU Psychological Clinic

Applied Research in Cognition and Human Factors (ARCH) Lab

Undergraduate Psychology Honors Program

Center for Behavioral and Cognitive Studies

Faculty & Staff

Faculty Positions

More Information...

Mail comments to:
dcerri@gmu.edu

Last update:
6/12/02