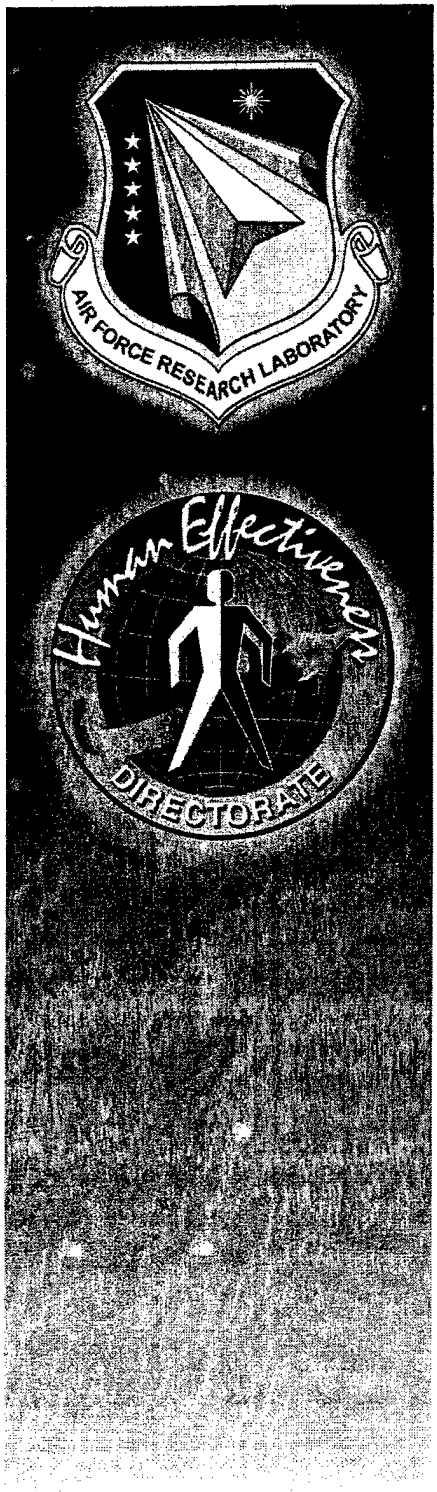
AFRL-HE-WP-TR-2002-0249

United States Air Force
Research Laboratory



PERFORMANCE ASSESSMENT OF A COTS SPEECH RECOGNITION SYSTEM ON THE N4 DATABASE

David T. Williamson

HUMAN EFFECTIVENESS DIRECTORATE
COLLABORATIVE INTERFACE BRANCH
WRIGHT-PATTERSON AFB OH 45433-7022

Robin A. Snyder Jr.

SYTRONICS, INC.
4433 DAYTON XENIA ROAD
DAYTON OH 45432

DECEMBER 2002

INTERIM REPORT FOR THE PERIOD AUGUST 2002 TO DECEMBER 2002

Human Effectiveness Directorate
Warfighter Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2002-0249

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

//Signed//

BRIAN P. DONNELLY, Lt Col, USAF
Deputy Chief, Warfighter Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MMM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From – To)
December 2002	Interim Report	August 2002 – December 2002

4. TITLE AND SUBTITLE

Performance Assessment of a COTS Speech Recognition System on the N4 Database

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

David T. Williamson
*Robin A. Snyder, Jr.

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
10

5f. WORKUNIT NUMBER
02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

*Sytronics, Inc.
4433 Dayton-Xenia Rd, Bldg 1
Dayton OH 45432

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

10. SPONSOR / MONITOR'S ACRONYM

AFRL-HE-WP-TR-2002-0249

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report discusses the evaluation of a commercially available speech recognition system on the NATO Native and Non-Native (N4) database. Using the statistical language modeling techniques, trigram language models were generated for each of three countries in the database, CA, NL, and UK. Due to time constraints, the DE database was not evaluated. For each of the countries, two factors were assessed. The first was overall word accuracy and the second was callsign accuracy. For this evaluation, only standard American English acoustic models were used. Results of each country evaluation are discussed.

15. SUBJECT TERMS

Speech Recognition, Command and Control, Unmanned Aerial Vehicles

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON:
a. REPORT	b. ABSTRACT	c. THIS PAGE			David T. Williamson
UNCLAS	UNCLAS	UNCLAS	UNLIMITED	15	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-7593

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

PAGES

LIST OF TABLES

EXECUTIVE SUMMARY

This report discusses the evaluation of a commercially available speech recognition system on the NATO Native and Non-Native (N4) database. Using the statistical language modeling techniques, trigram language models were generated for each of three countries in the database, CA, NL, and UK. Due to time constraints, the DE data was not evaluated. For each of the countries, two factors were assessed. The first was overall word accuracy and the second was callsign accuracy. For this evaluation, only standard American English acoustic models were used. Results of each country evaluation are discussed.

THIS PAGE INTENTIONALLY LEFT BLANK

INTRODUCTION

Commercially available speech recognition systems are finally reaching a level of maturity to be considered for various military applications [1] [2] [3] [4]. These applications range from ground-based command and control operations in an air operations center to tactical command and control in a high performance fighter aircraft. Another application that is of interest to the military is in the area of training. The use of speech recognition technology to act as synthetic players in training exercises promises to greatly reduce the manpower required to train personnel for various tasks, such as air traffic control, AWACS operations, and other communications tasks. A significant challenge for speech technology is to have it act as a performance assessment tool to automatically grade a student on their ability to correctly perform a given communications task. An additional challenge is if the student is trying to perform the communications task in non-native English. To see if commercial-off-the-shelf technology is up to this challenge, an evaluation was performed on the NATO Native and Non-Native (N4) database [5] consisting of students conducting naval communications training sessions from four different countries, Canada (CA), United Kingdom (UK), Netherlands (NL), and Germany (DE). Of particular interest was to see how well the COTS system would be able to recognize not only the individual words, but also how well it could recognize and identify the various callsigns spoken during the training sessions. This report discusses the development of the language models and the resulting word and callsign accuracy obtained from three of the countries represented in the database, CA, UK, and NL. Due to time constraints, the DE data was not evaluated.

PROCEDURE

Language Model & Callsign Interpretation Grammar Development

A separate statistical language model (SLM) was developed for each of the three countries. For each model, the transcripts were modified to replace specific callsign references with a generic Callsign grammar placeholder. A trigram SLM was generated from the modified training data. A unique callsign interpretation grammar was developed for each country based on an analysis of the format and frequency of occurance of callsigns. In addition to creating callsign grammars, several other grammars were developed to improve callsign detection accuracy. These included grammars for authentication codes and zulu time. The specific interpretation grammars for each country are outlined in Appendices A-C. Note that the nodes with a dotted line are optional nodes. For all three countries tested, the standard American English acoustic models provided with the system were used.

Data Preparation

Prior to the evaluation, several steps were necessary to prepare the source material. First, individual wav files were generated based on the transcription data provided. Next, each wav file was downsampled to 8KHz to match the requirements of the COTS system's acoustic model. Recognition testing was performed on each data set with several default parameters modified based on prior experience with this system on similar speech data. These parameters included

enabling a noise filtering process to improve the signal, reducing the rejection threshold to reduce rejection errors, and increasing the pruning value to improve accuracy. All recognition data was captured in log files for subsequent analysis.

RESULTS

The results for each country evaluation were parsed into two separate data sets. The first set contained the raw recognition text result returned by the system. The second set contained only a list of callsigns detected by the callsign interpretation grammars. These data sets were then formatted into spu_id input files for analysis by sclite, a NIST developed scoring program commonly used to score recognition testing.

Raw Text Transcription Results

The first metric of interest was how well the COTS system performed on the raw transcription task. The results for all three countries are presented in Table 1.

Performance Metric	**CA**		**NL**		**UK**	
Sentence Recognition Performance	count	(%)	count	(%)	count	(%)
Total Sentences	809		327		324	
Total Errors	767	94.8%	273	83.5%	229	70.7%
Substitutions	457	56.5%	255	78.0%	200	61.7%
Deletions	612	75.6%	96	29.4%	89	27.5%
Insertions	255	31.5%	110	33.6%	58	17.9%
Word Recognition Performance						
Total Words	11555		4520		4189	
Total Errors	3434	29.7%	1113	24.6%	924	22.1%
Substitutions	1005	8.7%	766	16.9%	438	10.5%
Deletions	2015	17.4%	172	3.8%	399	9.5%
Insertions	414	3.6%	175	3.9%	87	2.1%
Correct	8535	73.9%	3582	79.2%	3352	80.0%
Word Accuracy		70.3%		75.4%		77.9%

Table 1. Sentence and Word Error Rates for Transcription Task.

Callsign Detection Results

The second item of interest was how well the system could recognize and label callsign data within a given utterance. For purposes of scoring, each callsign was considered a single token or word. Also, a sentence was simply a sequence of callsigns detected in the original utterance. The results for all three countries are presented in Table 2.

Performance Metric	CA		NL		UK	
Sentence Recognition Performance	count	(%)	count	(%)	count	(%)
Total Sentences	809		321		324	
Total Errors	485	60.0%	246	76.6%	181	55.9%
Substitutions	330	40.8%	218	67.9%	154	47.5%
Deletions	73	9.0%	17	5.3%	12	3.7%
Insertions	222	27.4%	101	31.5%	39	12.0%
Word Recognition Performance						
Total Words	1217		554		519	
Total Errors	802	65.9%	438	79.1%	248	47.8%
Substitutions	381	31.3%	295	53.2%	173	33.3%
Deletions	106	8.7%	20	3.6%	27	5.2%
Insertions	315	25.9%	123	22.2%	48	9.2%
Correct	730	60.0%	239	43.1%	319	61.5%
Word (Callsign) Accuracy		34.1%		20.9%		52.2%

Table 2. Callsign Detection Results.

DISCUSSION

This database represented a significant challenge for evaluation. Not only was there a significant amount of disfluent speech present, but the addition of non-native English speakers proved very difficult for the COTS system. To be fair, the system's American English acoustic models were not very representative of much of the database. Also, very little fine tuning of pronunciation dictionaries was performed due to time constraints in the evaluation. This was a particular problem in the NL evaluation with many Dutch words interspersed among the English words. Additional performance benefits could be obtained if some adaptation was performed on the standard acoustic models and if dictionaries were tuned.

Another problem encountered in the evaluation was the length of several of the test utterances. The COTS system tested only accepts utterances under 30 seconds in duration. Many of the utterances exceeded this length. Appendix D shows the list of utterances for each country that could not be evaluated. Additional effort could be expended in splitting the utterances into smaller segments and then evaluating these segments against the COTS system.

REFERENCES

1. Williamson, D.T. and Barry, T.P. (2001). Speech Recognition in the Joint Air Operations Center – A Human–Centered Approach. In *Proceedings of the Human Computer Interaction International 2001 Conference*. New Orleans, LA.

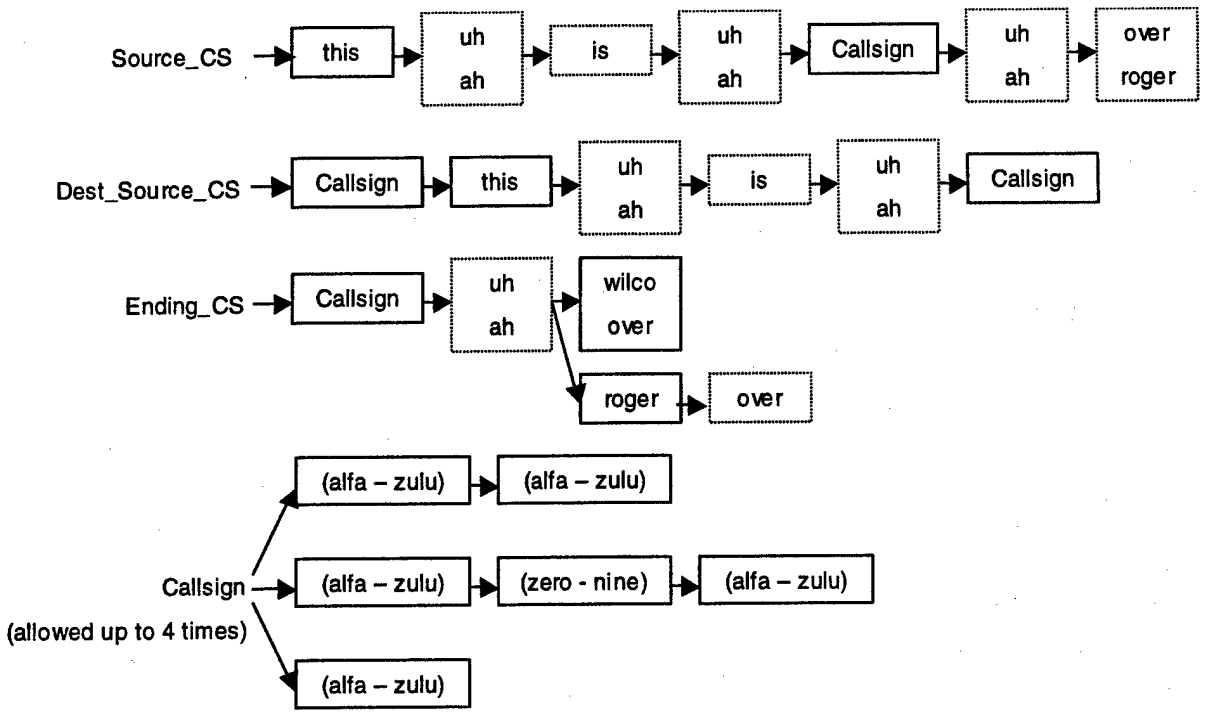
2. Williamson, D.T. and Barry, T.P. (2000). The Design and Evaluation of a Speech Interface for Generation of Air Tasking Orders. In *Proceedings of the International Ergonomics Association 14th Congress and the Human Factors and Ergonomics Society 44th Annual* Meeting. San Diego, CA: Human Factors and Ergonomics Society

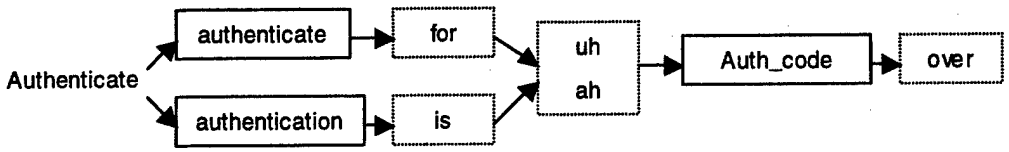
3. Williamson, D. T. (1997). Robust Speech Recognition Interface to the Electronic Crewmember: Progress and Challenges. In *Proceedings of 4th Human-Electronic Crewmember Workshop.* Kreuth, Germany.

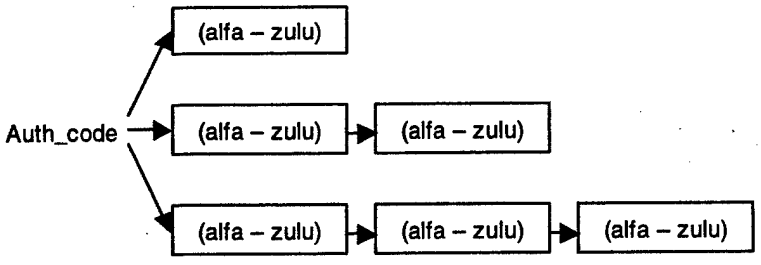
4. Barbato, G. J. "Integrating Voice Recognition and Automatic Target Cueing to Improve Aircrew-System Collaboration for Air-to-Ground Attack", In *Proceedings of the Research and Technology Organization Panel: Sensor Data Fusion and Integration of the Human Element*; RTO-MP-12 (pp. 24-1 to 24-11) published February 1999. System Concepts and Integration (SCI) Symposium, Ottawa, Canada, 14-17 September 1998.

5. Benarousse, L., Geoffrois, E., Grieco, J., Series, R., Steeneken, H., Stumpf, H., Swail, C., and Thiel, D. "The NATO Native and Non-Native (N4) Speech Corpus", In *Proceedings of the Workshop on Multilingual Speech and Language Processing;* published September 2001. Aalborg, Denmark.

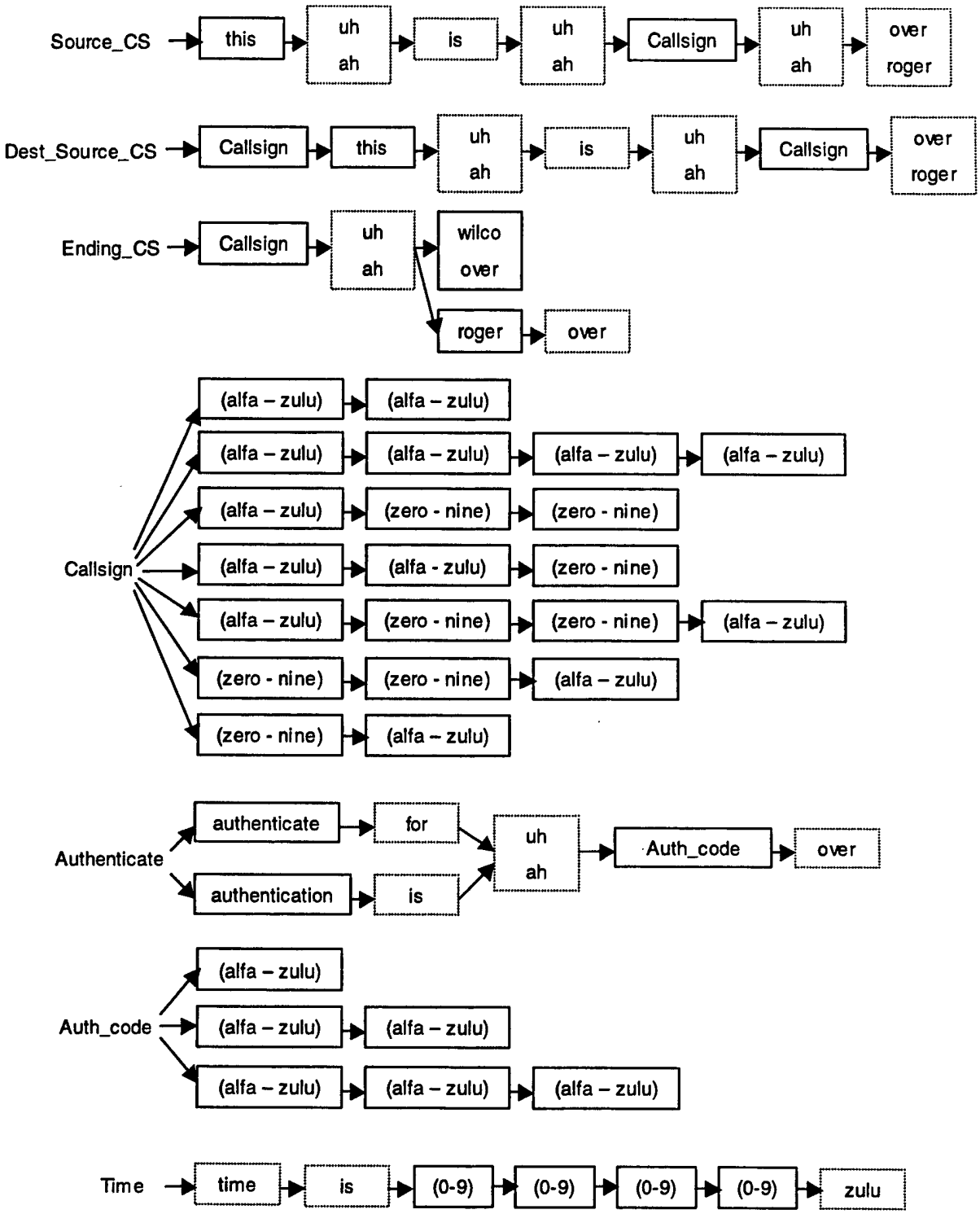
APPENDIX A:
CALLSIGN GRAMMARS FOR CA DATA





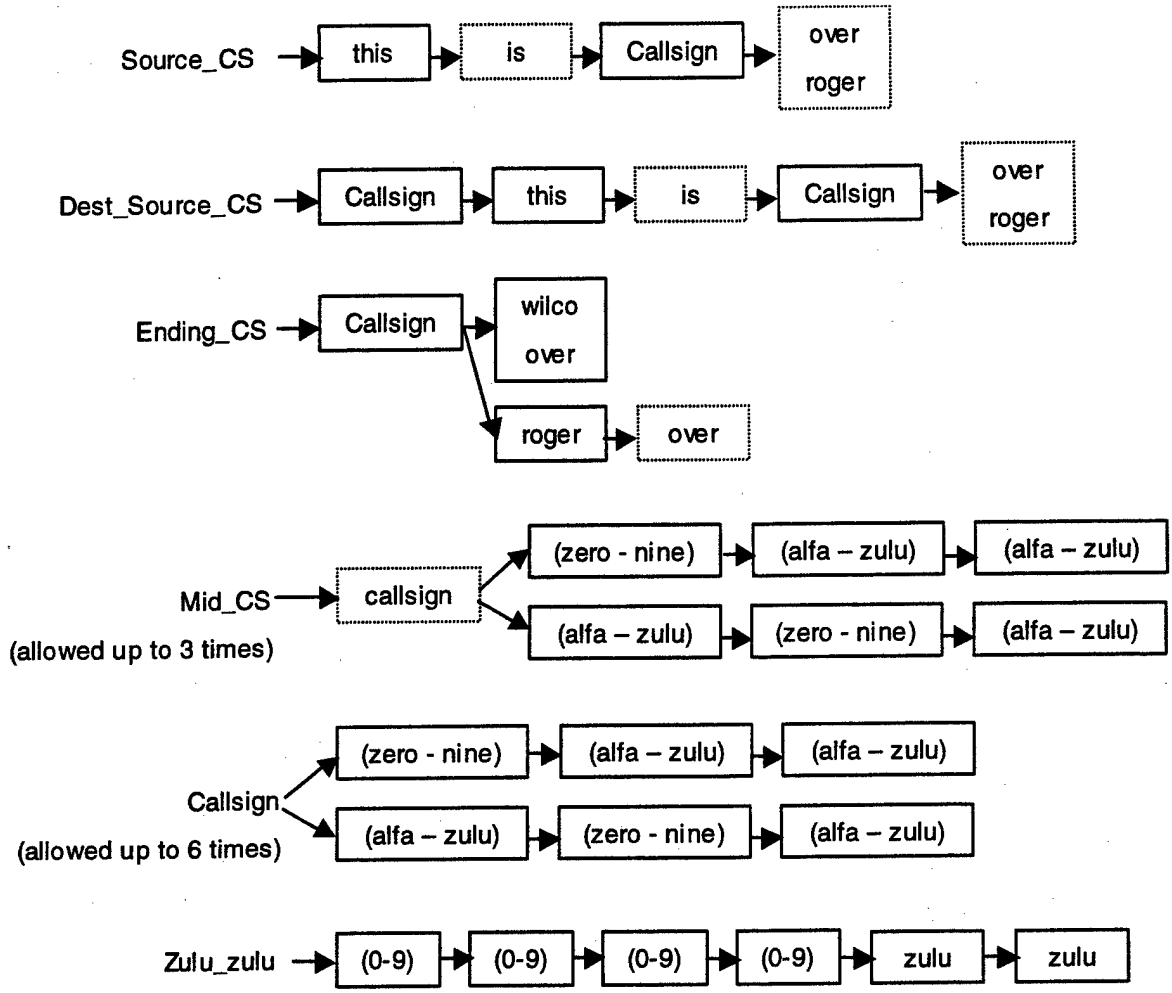


APPENDIX B:
CALLSIGN GRAMMARS FOR NL DATA



APPENDIX C:

CALLSIGN GRAMMARS FOR UK DATA



APPENDIX D:
UTTERANCES NOT EVALUATED

CA		NL		UK	
CA001-01-7	CA005-11-52	NL001-07-56	NL007-22-1	UK001-01-1	UK003-02-103
CA001-01-19	CA005-12-54	NL001-10-59	NL007-23-9	UK001-01-3	UK003-06-108
CA001-01-21	CA005-12-56	NL001-08-61	NL007-12-20	UK001-01-6	UK003-01-111
CA001-03-23	CA006-14-20	NL001-05-62	NL008-16-5	UK001-01-8	UK004-06-1
CA001-03-26	CA006-15-22	NL001-05-63	NL008-18-8	UK001-01-12	UK004-06-4
CA001-01-28	CA006-16-35	NL001-01-75	NL008-15-10	UK001-04-14	UK004-10-9
CA001-03-38	CA006-08-47	NL002-04-1	NL008-13-13	UK001-01-16	UK004-06-14
CA001-03-41	CA007-14-1	NL002-09-4	NL009-22-3	UK001-01-32	UK004-06-18
CA001-06-43	CA007-U-18	NL002-09-5	NL009-22-5	UK001-01-36	UK004-10-21
CA001-03-48	CA007-17-25	NL002-02-10	NL009-14-13	UK001-05-44	UK004-06-24
CA001-03-54	CA007-07-29	NL002-09-13	NL009-19-15	UK001-02-50	UK004-10-38
CA001-02-55	CA007-07-38	NL002-03-15	NL009-16-17	UK001-04-62	UK004-06-41
CA002-05-1	CA007-08-42	NL002-06-19	NL009-16-20	UK001-08-85	UK004-12-43
CA002-03-8	CA007-08-44	NL002-10-22	NL009-21-23	UK001-03-89	UK004-12-46
CA002-05-35	CA007-09-48	NL003-07-1	NL009-18-25	UK001-06-93	UK004-06-48
CA003-05-7	CA008-09-15	NL003-09-6	NL009-18-26	UK002-06-1	
CA003-02-12	CA008-08-18	NL003-09-10	NL010-18-3	UK002-03-12	
CA003-02-16	CA008-08-23	NL003-05-14	NL010-20-6	UK002-08-41	
CA003-02-19	CA008-13-26	NL003-02-23	NL010-20-9	UK002-08-44	
CA003-02-21	CA008-07-30	NL003-02-27	NL010-23-11	UK002-08-50	
CA003-02-22	CA008-07-32	NL003-11-31	NL010-12-12	UK002-08-52	
CA003-03-24	CA009-07-32	NL003-08-32	NL010-22-13	UK002-08-55	
CA003-04-26	CA009-12-38	NL003-01-35	NL010-17-15	UK002-08-59	
CA003-03-28	CA009-16-40	NL003-03-43	NL010-18-17	UK002-08-81	
CA003-02-35	CA009-14-46	NL003-03-45	NL010-17-18	UK003-06-2	
CA003-04-38	CA009-12-52	NL003-08-51	NL011-24-17	UK003-02-10	
CA003-02-40	CA009-07-79	NL003-10-53	NL011-26-20	UK003-12-20	
CA003-03-41	CA010-14-18	NL004-11-2	NL011-24-34	UK003-02-39	
CA003-04-44	CA010-10-45	NL004-09-4	NL011-27-39	UK003-12-44	
CA003-03-60	CA010-09-51	NL004-05-6	NL011-29-41	UK003-07-45	
CA003-05-63	CA010-08-55	NL004-02-11	NL011-26-46	UK003-10-49	
CA004-05-1	CA010-07-57	NL004-04-12	NL011-25-56	UK003-10-52	
CA004-02-7	CA010-07-63	NL004-XX-14	NL011-26-57	UK003-12-58	
CA004-02-9	CA011-21-1	NL004-08-25	NL012-30-1	UK003-07-60	
CA004-01-13	CA011-21-6	NL004-04-28	NL012-24-3	UK003-12-68	
CA004-03-17	CA011-21-23	NL004-10-31	NL012-28-9	UK003-12-73	
CA004-01-22	CA011-21-36	NL004-07-33	NL013-30-3	UK003-12-75	
CA004-01-24	CA011-21-81	NL005-19-11	NL013-24-20	UK003-12-83	
CA004-01-39	CA011-21-97	NL005-21-15	NL013-30-43	UK003-12-90	
CA004-02-47	CA011-19-99	NL006-13-1		UK003-12-94	
CA004-03-49	CA011-21-100	NL006-18-3		UK003-10-96	
CA005-U-2	CA011-21-142	NL006-17-14		UK003-01-98	
CA005-08-4	CA011-21-149	NL006-14-16		UK003-01-101	
CA005-10-10	CA011-U-156				
CA005-U-12	CA011-U-169				
CA005-15-19	CA011-21-191				
CA005-16-21	CA011-21-193				

CA005-U-23	CA011-21-230		
CA005-07-27	CA011-21-237		
CA005-U-29			
CA005-U-39			
CA005-10-44			
CA005-10-46			