REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 7, 1994	Final

4. TITLE AND SUBTITLE

Representing Text Meaning for Multilingual Knowledge-Based Machine Translation

5. FUNDING NUMBERS

6. AUTHOR(S)

Lynn Carlson Ronald Dolan
Elizabeth Cooper Steve J. Maiorano

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

Federal Research Division
Library of Congress
Washington, DC 20540-5220

8. PERFORMING ORGANIZATION REPORT NUMBER

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

U.S. Department of Defense
Fort Meade, MD 20755

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Prepared under an Interagency Agreement

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

An interlingua capable of representing the meaning of natural language texts is a crucial component of a knowledge-based machine translation system. In the Mikrokosmos project, researchers are defining the components of such an interlingua, or text meaning representation (TMR) language, through extensive analysis of Japanese and English texts in the domain of join business ventures. This paper describes the components of the TMR, providing examples of how certain phenomena are represented. The authors discuss their experience in analyzing the Japanese joint ventures corpus and its effect on TMR development. Dr. Ronald E. Dolan, a member of the Federal Research Division of the Library of Congress, is coauthor of the paper along with Lynn M. Carlson and Elizabeth L. Cooper of the U.S. Department of Defense, Fort George G. Meade, Maryland, and Steven J. Maiorano of the Office of Research and Development, Washington, D.C.

14. SUBJECT TERMS

Machine translations
Multilingual
Japanese-language

15. NUMBER OF PAGES

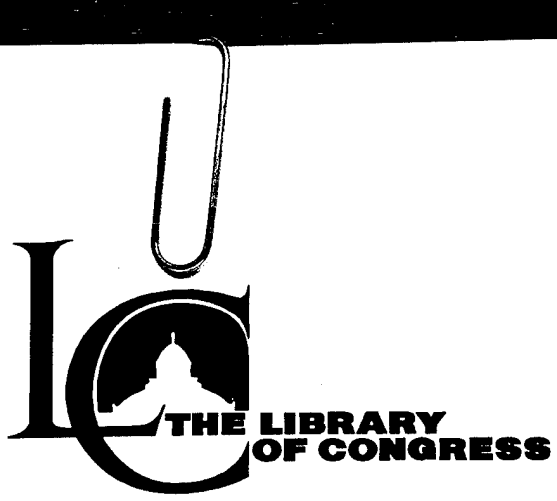
8

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

NSN 7540-01-280-5500

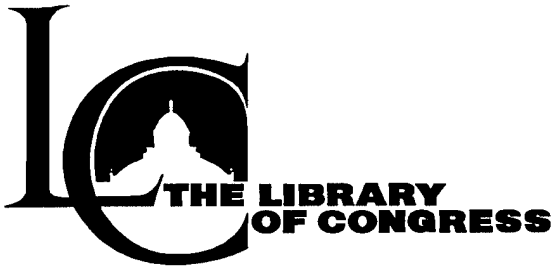
Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std 239-18
298-102



REPRESENTING TEXT MEANING FOR MULTILINGUAL KNOWLEDGE-BASED MACHINE TRANSLATION

A paper presented at the panel
The Voices of Experience: MT in Operational Settings
Association for Machine Translation in the Americas Annual Conference
Columbia, Maryland
October 7, 1994

Authors: Lynn M. Carlson
Department of Defense
Elizabeth L. Cooper
Department of Defense
Ronald E. Dolan
Library of Congress
Steve J. Maiorano
Office of Research and Development



REPRESENTING TEXT MEANING FOR MULTILINGUAL KNOWLEDGE-BASED MACHINE TRANSLATION

A paper presented at the panel
The Voices of Experience: MT in Operational Settings
Association for Machine Translation in the Americas Annual Conference
Columbia, Maryland
October 7, 1994

Authors: Lynn M. Carlson
Department of Defense
Elizabeth L. Cooper
Department of Defense
Ronald E. Dolan
Library of Congress
Steve J. Maiorano
Office of Research and Development

REPRESENTING TEXT MEANING FOR MULTILINGUAL KNOWLEDGE-BASED MACHINE TRANSLATION

Lynn M. Carlson
Elizabeth L. Cooper
U.S. Department of Defense
Ft. George G. Meade, MD 20755
{lmcarls, elcoope}@afterlife.ncsc.mil

Ronald E. Dolan
Federal Research Division, Library of Congress
Washington, D.C. 20540
frds@loc.gov

Steven J. Maiorano
Office of Research and Development
Washington, D.C. 20505
71045.26@compuserve.com

An interlingua capable of representing the meaning of natural language texts is a crucial component of a knowledge-based machine translation system. In the Mikrokosmos project, researchers are defining the components of such an interlingua, or text meaning representation (TMR) language, through extensive analysis of Japanese and English texts in the domain of joint business ventures. This paper describes the components of the TMR, providing examples of how certain phenomena are represented. The authors discuss their experience in analyzing the Japanese joint ventures corpus and its effect on TMR development.

1. Introduction

One of the central issues in the Mikrokosmos knowledge-based machine translation project (developed jointly by researchers at New Mexico State University, Carnegie Mellon University and various U.S. government agencies), is to develop an expressive language for representing the meaning of natural language texts. This language, an interlingua, should be language-neutral and facilitate computer processing. Using this language, the MT program specifies the meaning of input texts, or text meaning representations (TMRs). The TMR represents the result of analysis of a given input text in any one of the languages supported by the system, and serves as input to the generation process. Elements of the TMR language must be interpreted in terms of an independently motivated model of the world (or ontology). The link between the ontology and the TMR is provided by the lexicon, where the meanings of most open class lexical items are defined in terms of their mappings into ontological concepts and their resulting contributions to TMR structure (see Carlson and Nirenburg, 1990; Meyer, et al., 1990; Onyshkevych and Nirenburg, 1991, for a description of these static knowledge sources). Information about the nonpropositional components of text meaning — pragmatic and discourse-related phenomena such as speech acts, speaker attitudes and intentions, relations among text units, deictic references, etc. — is also derived from the lexicon, and becomes part of the TMR.

To test the Mikrokosmos TMR and to help in developing our methodology for massive acquisition of the ontology and the lexicons required for the support of automatic production of the TMRs in Mikrokosmos, we have undertaken extensive linguistic field work, analyzing and manually annotating a number of texts from the Japanese and English joint ventures (JV) corpora collected in the framework of the Tipster Text Program. (The analysis is being extended to Spanish and French, using texts that are similar in style and content to the Tipster corpora.) The annotation task is, in fact, manual translation of these texts into the lan-

guage of TMRs. This field work has facilitated the improvement of the specification of the interlingua as well as the lexicons and the ontology. In this paper, we present some of the preliminary results of the Japanese component of our field investigation.

2.Components of a TMR

2.1 What is a TMR?

A TMR is derived by syntactic, semantic, and pragmatic analysis of the text. Because the TMR is intended to be language neutral, it also avoids syntactic terminology (e.g., notions such as *clause*, *tense*, etc.) In addition to providing information about the lexical-semantic dependencies in the text, the TMR represents stylistic factors, discourse relations, speaker attitudes, and other pragmatic factors present in the discourse structure. In doing so, the TMR captures not only the meaning of individual elements in the text, but also the relations between those elements, while taking into account both propositional and nonpropositional components of textual meaning.

2.2 TMR Structure

The current version of the Mikrokosmos TMR is divided into several components which combine to convey the overall meaning of the original text. These include heads (roughly, the predications), speech acts, attitudes, and stylistic factors, as well as temporal, coreference, textual and domain relations. In the TMR, the results of analysis of an input text are represented in a frame-oriented notation (see Nyberg, 1988; Brown, 1994). A frame can represent an instantiated ontological concept, speech act, relation among frames in a TMR, speaker attitude, etc. Prefixes on symbols in the TMR have the following meanings:

```
%     instantiated ontological concept or meta-ontological TMR
construct (%company, %attitude)
$     named instance ($"Ajinomoto Dannon",$Japan)
&     symbolic constant (&red, &blue)
*     concept in the ontology (*company)
*x*   special variable (*author*, *unknown*)
```

Concepts in the ontological world model include objects, events, and their properties, arranged in a IS-A hierarchy. In producing a TMR, ontological concepts can be instantiated, so that `%company_34` would indicate a particular mention in a text of the ontological concept `*company`.

TMR representations have been developed based on preliminary analysis of the Japanese JV corpus. This corpus contains approximately 1300 on-line newswires (up to two pages in length) from four sources, reporting on international joint business ventures. These articles discuss the formation, expansion or dissolution of an agreement between two or more entities involved in economic activities such as manufacturing, sales, research or finance (see Tipster Text Program, 1993). The style of these texts is neutral with respect to such indicators as formality, politeness, color, etc. (see Hovy, 1988), and the texts do not contain a rich variety of speech acts or textual relations. In the discussion below, we concentrate on more frequently occurring phenomena in the corpus, giving examples to illustrate representations for heads, attitudes and relations.

In a TMR, a natural language clause is typically represented in a frame by instantiating an `EVENT` or `PROPERTY` concept from the ontology; this concept is referred to as an interlingual head in the TMR, and contains a number of modifying properties (such as case and circumstantial roles) that further define it. All

heads must have TIME, ASPECT (a combination of PHASE, ITERATION, DURATION; see Nirenburg and Pustejovsky, 1988), and POLARITY (positive/negative). Information about the head is given in a slot-filler format, with the slot representing a property and the filler, its value. Fillers are suffixed by an instance number, so that in a given text each occurrence of a concept has a unique number. The frame below represents the clause "Ajinomoto decided to underwrite...":

```
%decide_1
     agent       %company_1        ;Ajinomoto
     theme       %underwrite_1
     time        %time_1
     aspect      %aspect_1
     polarity    &positive
```

EVENT heads can have other slots (e.g. COTHEME, ACCOMPANIER, BENEFICIARY, PURPOSE, MANNER, ATTITUDE, LOCATION, FOCUS, etc.), as needed to convey the meaning of the original text.

2.3 Representing Attitudes

Attitudes are used to reflect the way elements in the text are perceived by an intelligent agent (typically the speaker/writer of the text). At present the following six attitudes are used in TMRs (this list may be expanded after further analysis of the corpus):

a. *Epistemic* - someone believes it is true/false

b. *Deontic* - someone believes someone must/must not

c. *Volition* - someone desires/does not desire

d. *Expectation* - someone expects/does not expect

e. *Evaluative* - someone believes it is best/worst

f. *Potential* - someone believes it is/is not possible

Attitudes are defined in terms of the following properties: TYPE, ATTRIBUTED-TO, SCOPE, TIME, and VALUE. The TYPE slot is filled with one of the attitude types listed above. ATTRIBUTED-TO is filled by the agent or entity who possesses the attitude. SCOPE identifies the segments of the TMR (and corresponding text) covered by the attitude, and TIME is the time at which the attitude holds. VALUE is assigned on a scale of 0 to 1.0, with 0 being negative, 1.0 being positive, and values or ranges in between showing qualification. Attitudes may be combined to capture a particular meaning in a text. For example, in representing the meaning of the Japanese input translated as "There is also concern[1] that ... licensing and know-how disputes will occur", an epistemic attitude reflects the belief that the situation may occur, while an evaluative attitude captures the less than positive feeling about the event taking place.

```
%attitude_4                    %attitude_5
   type          epistemic        type            evaluative
   attributed-to *author*         attributed-to   *author*
   scope         %occur_1         scope           %occur_1
   time          %time_16         time            %time_16
   value         >0.5             value           <0.4
```

1. The Japanese *osore* means "concern, danger, fear" and has a stronger negative connotation than the English translation, hence the negative value on the evaluative attitude.

2.4 Representing Relations

Relations of various types are used in TMRs to represent the connection between the content of two or more textual elements. Each has its own format, and may be further divided into subtypes. Below we give examples of domain and temporal relations.

Domain relations represent connections between events, states or objects in the text. These connections can be quite general, scoping over large portions of text, or more specific, and limited in scope (e.g. linking consecutive heads). Domain relations in the TMRs are classified into four categories, each of which may have several subtypes; further analysis may result in adding new and/or combining existing ones (such as in Hovy, 1994):

a. Causal - relations of dependency among events, states, and objects in the TMR

b. Conjunction - relations of adjacency between events, states, and objects in the TMR

c. Elaboration - relations between TMR elements, one of which expands on or refines the other

d. Alternation - relations that are used in situations of choice; either/or

Domain relations are represented with the slots TYPE, ARG_1, and ARG_2. TYPE is filled with the appropriate domain relation type, selected from one of the above categories, or a subtype; ARG_1 and ARG_2 are filled with the TMR elements between which the relation exists. Examples of a CONDITION causal relation and a PARTICULAR elaboration relation from the corpus follow:

"For example, **if** someone who subscribed at age 40 pays in approximately 20,000 yen every month, and 12,000 yen at bonus times, he could receive 84,000 yen every 3 months for 10 years..."

```
%domain-rel_5
      type        *condition
      arg_1       %deposit_1
      arg_2       %receive_2
```

"Autovax Seven announced a business tie up with Yaohan Department Store concerning setting up branch stores specializing in general auto supplies. **Specifically**, they plan to sell auto supplies..."

```
%domain-rel_1
      type        *particular
      arg_1       %plan_1
      arg_2       %establish_1
```

Temporal relations indicate the relative timing of one event in the text in relation to another. In the TMR, temporal relations are represented using the slots TYPE, ARG_1, and ARG_2, where fillers for ARG_1 and ARG_2 are times (e.g. %time_1), and TYPE indicates the relation between the two times, filled by one of the values at, after or during. A temporal relation may also have a VALUE slot to indicate the relative distance between two times. If %time_2 occurred **just after** %time_1, the temporal relation would look like this:

```
%temporal-relation_2
      type        &after
      arg_1       %time_2
      arg_2       %time_1
      value       <0.2
```

3. TMR Experience and Development Methodology

After developing a basic set of TMR notation, we carried out further analysis of the JV corpus, in order to test the adequacy of the representation. In this section, we discuss some representation issues we encountered, and the preliminary results of our analysis.

One problem that is prevalent in language and occurred repeatedly in analyzing the JV corpus was that of how to treat noun-verb pairs, such as *to tie up/a tie up*, *develop/development, construct/construction.* One of our goals in designing a language-independent ontology is to achieve economy of representation across languages, by avoiding a one-to-one correspondence between the lexical items of any one language and the constructs posited as ontological entries. Therefore, it would be desirable not to create, for example, an OBJECT concept for *development* and a corresponding EVENT concept for *develop*. One of our initial investigations was to see if we could achieve adequate representation of noun-verb pairs with a single ontological concept — either an EVENT or an OBJECT.

To study the issue, an analysis of the word *teikei* ("tie up") was carried out. *Teikei*, which occurs frequently in the Japanese JV corpus, refers to a broad range of business agreements between companies, and is used both nominally and verbally. Initially we tried to represent *teikei* as an EVENT (`tie-up`); however, this proved to be inadequate for representing properties of a tie up, which lend themselves to modification of an object concept:

"Tobishima Kensetsu established a technology tie up with Ellis Donne in the area of construction technology."

```
%tie-up_1
      agent          %company_1          ;Tobishima Kensetsu
      accompanier    %company_2          ;Ellis Donne
      time           %time_2
      aspect         %aspect_2
      polarity       &positive
```

The above notation, with `tie-up` as an EVENT, made it difficult to convey the fact that the `tie-up` was a "technology tie up in the area of construction technology." A keyword in context search on the English and Japanese JV corpora revealed that *tie up* and *joint venture* in English, and *teikei* in Japanese, often contained complex modification that was more easily captured by properties that describe objects, not events: *manufacturing and sales tie up, tie up for credit card business, mutual technology tie up, corporate group tie up, business tie up for production, etc.*

We concluded that a more efficient way of representing tie ups would be to consider `tie-up` an OBJECT and posit a concept such as `create` to account for the EVENT. The resulting representation of the above example captures the modification of `tie-up`:

```
%create_1
      agent          %company_1          ;Tobishima Kensetsu
      theme          %tie-up_1
      accompanier    %company_2          ;Ellis Donne
      time           %time_2
      aspect         %aspect_2
      polarity       &positive
```

```
%tie-up_1
      tie-up-type       %technology_2
      scope             %technology_1

%technology_1
      technology-type         %construction_1
```

One guiding principle of our methodology is to avoid ambiguity of representation in TMRs; thus one of our objectives is to arrive at a uniform treatment of noun-verb pairs that can accommodate the analysis of texts from multiple languages. However, further analysis of the corpora is needed before a final recommendation can be made.

Another outcome of the corpus analysis of *teikei* was a list of typical verbs that take *teikei* as an argument. Many of these verbs convey the same or similar sense. By clustering these into closely-related senses similar to WordNet "synsets" (see Miller, 1990), and proposing a single ontological concept to cover them, we can avoid a one-to-one correspondence between lexical items and ontological concepts, and achieve a more efficient ontology. Subtle nuances can then be captured by constraints defined in the lexicon on the ontological properties of the concept to which the word maps (see Onyshkevych and Nirenburg, 1991). Some examples of these overlapping verb senses follow:

teikei o staato suru - "start a tie up"
teikei o hajimeru - "start a tie up"
teikei ni fumikiru - "venture into a tie up"
teikei ni fumidasu - "venture into a tie up"
teikei ni hashiridasu - "launch into a tie up"
teikei ni noridasu - "embark on a tie up"

teikei o kyooka suru - "strengthen a tie up"
teikei o fukumeru - "strengthen a tie up"
teikei o sekkyokka suru - "beef up a tie up"

Because descriptions of companies or corporations are central to the JV corpus, another thrust of our analysis was to determine what types of properties were needed to adequately represent this information in TMRs. After researching various sources, including the Japan Company Handbook and the Standard Industrial Classification Manual (see Tokyo Keizai, 1990, and Executive Office of the President, 1987, respectively), and running a keyword in context search on *kaisha* ("company") a number of slots were defined to account for company attributes. The examples below illustrate some of the more extensively used slots:

"Auto supply vendor Autovax Seven (headquarters, Osaka) ... "

```
%company_1
      name              $"Autovax Seven"
      headquarters      $Japan (country), $Osaka (city)
      activity          %sales_1
      product           %supply_1

%supply_1
      supply-type       %automotive_1
```

"The Seibu Sezon Group's hotel chain, Intercontinental Hotels (IHC; headquartered in New Jersey in the United States; chairman, Tsutsumi Yuji) ... "

```
%company_1
     name            $"Intercontinental Hotels"
     headquarters    $"United States"(country),
                     $"New Jersey" (province 1)
     alias           $IHC
     chairman        $"Tsutsumi Yuji"
     owned-by        %company_3

%company_3
     name            $"Seibu Sezon Group"
```

To date, around 80 company property slots have been identified, along with candidate fillers for those slots. These will continue to be expanded and modified with further analysis of the corpus.

Future Directions

Field work in the Mikrokosmos project will continue. We will investigate how to best represent the textual meaning of a variety of phenomena, including causal relations, speech acts, attitudes, scalar attributes for adjectives of comparison, and modifying roles for states and events. This will, in turn, help in the task of the acquisition of lexicon and ontology entries.

Acknowledgements

The authors would like to thank the following members of the Mikrokosmos team: Ralf Brown, Jonathan K. Davis, Donalee Hughes Attardo, Sergei Nirenburg, Boyan Onyshkevych and Jerry Reno.

References

Brown, R. 1994. FRAMEPAC. Center for Machine Translation. Carnegie Mellon University. CMU-CMT-MEMO.

Carlson, L. and S. Nirenburg. 1990. World Modeling for NLP. Center for Machine Translation, Carnegie Mellon University, Tech Report CMU-CMT-90-121.

Dahlgren, K. 1988. *Naive Semantics for Natural Language Understanding*. Boston, MA: Kluwer Academic Press.

Executive Office of the President. Office of Management and Budget. *Standard Industrial Classification Manual, 1987*. Springfield, Virginia.

Goodman, K. and S. Nirenburg (eds.). 1991. *The KBMT Project: A Case Study in Knowledge-Based Machine Translation*. San Mateo, CA: Morgan Kaufmann.

Hirst, G. 1989. Ontological Assumptions in Knowledge Representation. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, Toronto, Canada, 157-169.

Hovy, E. 1988. Generating Natural Language Under Pragmatic Constraints. Ph.D. Dissertation. Yale University.

Hovy, E. and E. Maier. 1994. Parsimonious or Profligate: How Many and Which Discourse Structure Relations? *Discourse Processes* (to appear).

Lenat, G. and R. Guha. 1990. *Building Large Knowledge-Based Systems*. Reading, MA: Addison-Wesley

Meyer, I., B. Onyshkevych and L. Carlson. 1990. Lexicographic Principles and Design for Knowledge-Based Machine Translation. Technical Report (CMU-CMT-90-118), Center for Machine Translation. Carnegie Mellon University, Pittsburgh, PA.

Miller, G. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4).

Nirenburg, S., J. Carbonell, M. Tomita and K. Goodman. 1992. *Machine Translation: A Knowledge-Based Approach*. San Mateo, CA: Morgan Kaufmann.

Nirenburg, S. and C. Defrise. 1993. Lexical and Conceptual Structure for Knowledge-Based Machine Translation. In J. Pustejovsky (ed.), *Semantics and the Lexicon*. Dordrecht: Kluwer.

Nirenburg, S. and L. Levin. 1991. Syntax-Driven and Ontology-Driven Lexical Semantics. In *Lexical Semantics and Knowledge Representation: Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics*, Berkeley, CA, 9-19.

Nirenburg, S. and J. Pustejovsky. 1988. Processing Aspectual Semantics, *Proceedings of the Tenth Annual Meeting of the Cognitive Science Society,* Montreal, Canada.

Nyberg, E. 1988. The FRAMEKIT User's Guide, Version 2.0. Center for Machine Translation, Carnegie Mellon University. CMU-CMT-MEMO.

Onyshkevych, B. and S. Nirenburg. 1991. Lexicon, Ontology and Text Meaning. In *Lexical Semantics and Knowledge Representation: Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics,* Berkeley, CA, 238-249.

Tipster Text Program. 1993. *Phase I: Proceedings of a Workshop held at Fredericksburg, Virginia*, Morgan Kaufmann.

Tokyo Keizai Inc. 1990. *Japan Company Handbook*. Tokyo, Japan.