

To appear in Proc. 6th ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting.

October 8, 1995, Chicago IL.

Semantic Classification for Practical Natural Language Processing*

Kavi Mahesh and Sergei Nirenburg

Computing Research Laboratory
Box 30001, Dept. 3CRL
New Mexico State University
Las Cruces, NM 88003-8001 USA
(505) 646-5466 FAX: (505) 646-6218
mahesh@crl.nmsu.edu, sergei@crl.nmsu.edu

Abstract

In the field of natural language processing (NLP) there is now a consensus that all NLP systems that seek to represent and manipulate meanings of texts need an ontology, that is a taxonomic classification of concepts in the world to be used as semantic primitives. In our continued efforts to build a multilingual knowledge-based machine translation (KBMT) system using an interlingual meaning representation, we have developed an ontology to facilitate natural language interpretation and generation. The central goal of the Mikrokosmos project is to develop a computer system that produces a comprehensive Text Meaning Representation (TMR) for an input text in any of a set of source languages. Knowledge that supports this process is stored both in language-specific knowledge sources (such as a lexicon) and in an independently motivated, language-neutral ontology of concepts in the world.

Keywords: Semantic classification, ontology, concept, knowledge representation, natural language processing, machine translation, world model, meaning representation.

*Research reported in this paper was supported in part by Contract MDA904-92-C-5189 from the U.S. Department of Defense.

1 Ontologies for NLP: Introduction

The field of natural language processing (NLP) is concerned with the construction of computer software systems that can process texts written in English and other natural languages with interpretation and generation capabilities much like those of human beings. Unlike text processing or word processing which processes texts at more or less superficial levels, many NLP systems try to extract the meanings contained in a sentence or an entire text. Meanings thus extracted may be used for various tasks such as performing robot actions, retrieving information from a database, or, in the case of machine translation systems, producing an equivalent translation in a different natural language.

In order to extract meanings from a text and to process the meanings for performing various tasks, the NLP system must be able to represent meanings in a form suitable for manipulation by the computer. A first step towards representing meaning is selecting a set of symbols as the primitive elements of which more complex meaning representations are constructed. For example, if the only meaning we wanted to represent was the gender of a person, we could have selected the symbols **M** and **F** to represent it. Since natural language texts contain a wide range of complex meanings, the set of symbols selected for representing meaning tends to be much larger. A traditional dictionary describes meanings of words using other words in the same or another language. This is not a good choice for computer processing of meanings for a variety of reasons. For example, words in most languages are highly ambiguous; words have synonyms and do not map to meanings uniquely; and so on.

We call the symbols used to represent meanings *concepts* to distinguish them from words in a language. For NLP purposes, we not only need to select a set of concepts but also to tell the computer how a concept is related to some or all of the other concepts known to the system. Such knowledge of conceptual relationships is invaluable in resolving ambiguities in the meaning of a text. One part of specifying conceptual relationships involves organizing them in a hierarchy or a taxonomic classification. The other part is adding “cross” links between different branches of the classification to represent relationships between concepts other than taxonomic relations. Such a classification system results in a richly interconnected network of concepts in the world (or a particular domain of focus). We call the network an *ontology*.

In the field of natural language processing (NLP) there is now a consensus that all NLP systems that seek to represent and manipulate meanings of texts need an ontology as a source of semantic primitives (Bateman, 1993; Carlson and Nirenburg, 1990). An ontology for NLP purposes is a body of knowledge about the world (or a domain) that: a) is a repository of primitive symbols used in meaning representation; b) organizes these symbols called concepts in a tangled subsumption hierarchy; and c) further interconnects these symbols using a rich system of semantic and pragmatic relations defined among the concepts. In order for such an ontology to become a computational resource for solving problems such as ambiguity and reference resolution, it must be actually constructed, not merely defined formally, as is the practice in the field of formal semantics. The ontology must also be put into well-defined relations with other knowledge sources in the system such as a lexicon.

1.1 The Context: Mikrokosmos

Mikrokosmos (μK) is a knowledge-based machine translation (KBMT) system under development at the computing research laboratory (CRL) of New Mexico State University (Onyshkevych and Nirenburg, 1994; Mahesh and Nirenburg, 1995; Beale, Nirenburg, and Mahesh, 1995).¹ Unlike previous research in interlingual machine translation (MT), this project is building a large-scale, practical MT system. μK already has several thousand Spanish words in its lexicon as well as several thousand concepts in its ontology (or world knowledge base). By the end of the year, a lexicon of approximately 7000 Spanish words supported by an ontology of over 5000 concepts will be in place. High-quality meaning representations of up to 10 article-length Spanish texts from the domain of company mergers and acquisitions will have been produced by the μK system. In the coming years, μK will be expanded into other languages such as Arabic, Japanese, Russian, and Thai.

A comprehensive study of the computational treatment of texts is a multifaceted endeavor covering a wide range of linguistic and pragmatic phenomena. Because the various facets of this knowledge are complex in their own right, study of any individual phenomenon is often conducted in relative isolation from the study of other related phenomena. However, in a KBMT application, knowledge about a large number of interrelated linguistic and language use phenomena is required. A natural way of combining the diverse knowledge required of such a system into a unified whole is for the various phenomena to be treated by separate computational linguistic “microtheories” united through a system’s control architecture and knowledge representation conventions.²

In the Mikrokosmos project, a comprehensive study of a variety of microtheories central to the support of KBMT systems is being carried out with the ultimate objective of defining a methodology for representing the meaning of natural language texts in a language-neutral interlingual format called a text meaning representation (TMR). The TMR represents the result of analysis of a given input text and serves as input to the target language generator. The meaning of the input text is represented in the TMR as instantiated elements 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. The ontology and the lexicon are the two main knowledge sources in the μK system. Information about the nonpropositional components of text meaning such as speech acts, speaker attitudes and intentions, relations among text units, coreferences, etc. is also derived from the lexicon with inputs from other microtheories and becomes part of the TMR. Figure 1 illustrates the μK architecture for analyzing input texts. The workings of this architecture are illustrated below through an example.

¹A more complete overview of the μK project is also available at the CRL World Wide Web home page at <http://crl.nmsu.edu/index.html>

²The name “Mikrokosmos” refers to a society of microtheories that act as meaning specialists each contributing to the construction of a comprehensive meaning representation (TMR) for an input text.

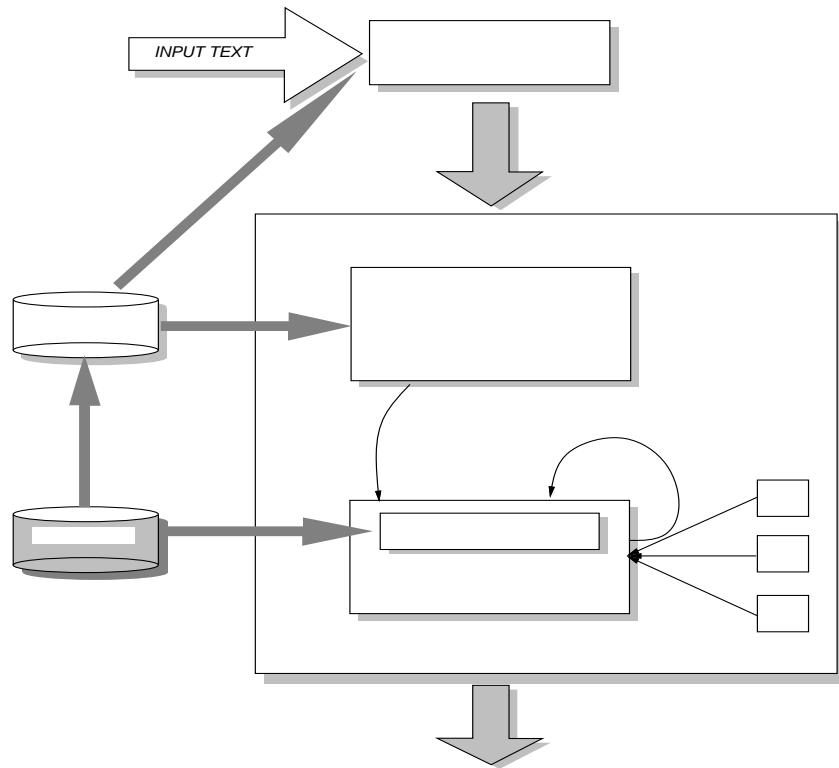


Figure 1: The Mikrokosmos NLP architecture.

1.2 An Example

Let us consider as an example a Spanish sentence and the meaning representation (TMR) produced by μK for that sentence. The following sentence is taken from a news article on the acquisition of a pharmaceutical company by the Roche group:³

(1) El grupo Roche, a través de su compañía en España, adquirió Doctor Andreu, se informó hoy aquí.

When translated into English, the sentence reads:

The Roche group, through its company in Spain, acquired Doctor Andreu, it was announced here today.

³Although the design of μK and its meaning representations includes coreference and other discourse relations, the current implementation of μK processes a text sentence by sentence.

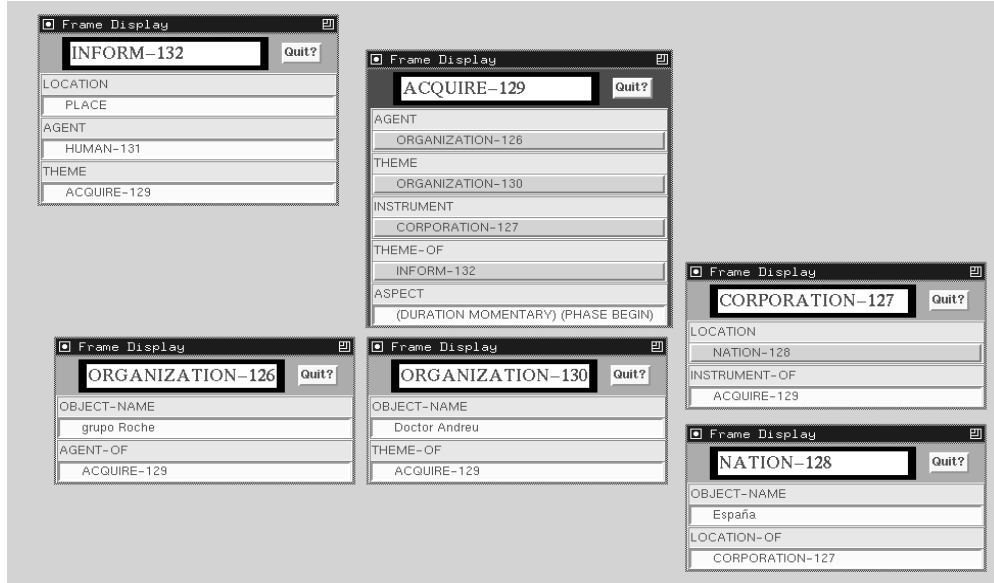


Figure 2: Partial text meaning representation (TMR) of sentence (1).

Given this input, the TMR produced as output by the μ K system is shown in figure 2. The meaning of sentence (1) is comprised of an `INFORM-132`⁴ event carried out by an unknown `HUMAN-131` agent. The theme of the `INFORM-132` event is an `ACQUIRE-129` event where the agent, `ORGANIZATION-126`, named “grupo Roche” acquired the theme, `ORGANIZATION-130`, named “Doctor Andreu.” This acquisition was done through the instrument, `CORPORATION-127`, which is located in the `NATION-128` named “España.” In other words, the meaning of the sentence as represented in the TMR is:⁵

Someone informed that the organization named “grupo Roche” acquired the organization named “Doctor Andreu” through the corporation located in the nation named Spain.

⁴Instances of concepts are given names by appending an arbitrary but unique integer to the end of a concept name. For example, `INFORM-132` is an instance of the concept `INFORM`. A TMR is essentially a network of such instances.

⁵It may be noted that what is shown in Figure 2 does not capture the full meaning of the sentence. For example, it does not say that `CORPORATION-127` is owned by `ORGANIZATION-126` (“its company”) or specify the time and location of the `INFORM-132` event (“here today”). We are currently building various microtheories and integrating them with the core semantic analyzer to enhance the TMR in various ways.

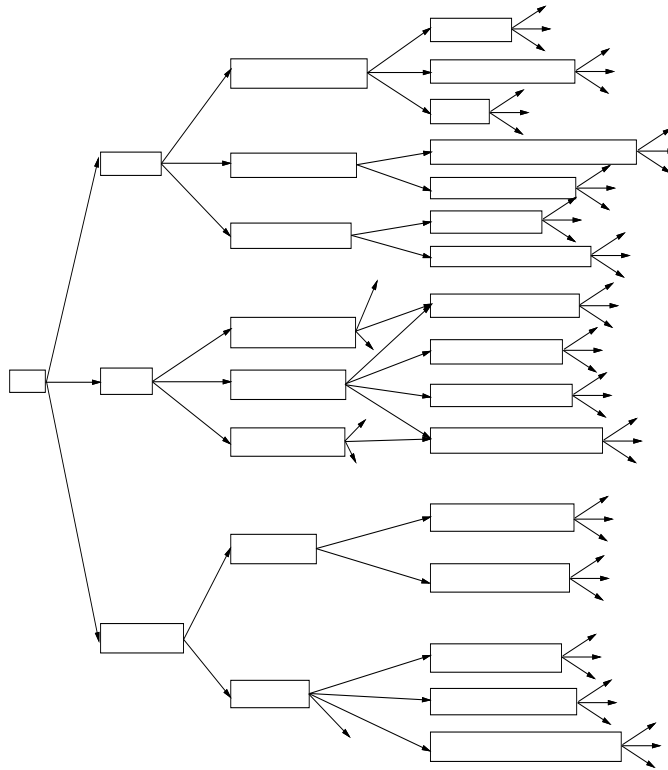


Figure 3: Top-level hierarchy of the Mikrokosmos ontology showing the first three levels of the object, event, and property taxonomies.

It may be noted that symbols such as `INFORM`, `ACQUIRE`, `AGENT`, `THEME`, and so on are, in fact, the concepts that are classified in the ontology. The TMR is produced by instantiating these concepts in the ontology and linking them together according to the relationships between the meanings contained in the input text. We will return to this example later in this article to illustrate how the meaning representation was produced by the μ K system.

2 The Mikrokosmos Ontology

The Mikrokosmos project is focusing on processing texts about mergers and acquisitions of companies. However, since the input language is unrestricted, the ontology must, in fact, cover a wide range of concepts outside this particular domain.

All entities in the μ K ontology are classified into *free-standing entities*⁶ and *properties*. Free-

⁶*Free-standing entity* is not an actual frame in the ontology; we use the term when describing the top-level organization of the ontology in order to distinguish between objects and events on the one hand and properties on

standing entities are in turn classified into *objects* and *events*. Figure 3 shows the top-level hierarchy in the ontology. Objects, events, and properties constitute the *concepts* in the ontology which are represented as *frames*. Each frame is a collection of slots with one or more facets and fillers. The slots (including inherited ones) collectively define the concept by specifying how the concept is related to other concepts in the ontology (through *relations*) or to literal or numerical constants (through *attributes*). Lexicon entries represent word or phrase meanings by mapping them to concepts in the ontology. A number of concepts in the domain of mergers and acquisitions are located under the ORGANIZATION subtree under SOCIAL-OBJECTS and the BUSINESS-ACTIVITY subtree under SOCIAL-EVENTS (see Figure 3).

Each concept is represented by a frame that has a name⁷ and the following slots: a *definition* that is an English string used solely for human browsing purposes, a *time-stamp* for bookkeeping, taxonomic links (*is-a* and *subclasses* for concepts and *instances* and *instance-of* for instances), and other slots (see Figure 6 for an example). Other slots can be any property defined under the property hierarchy of the ontology. The properties, though they are defined as concepts, are not instantiated as stand alone TMR frames; they are present in TMRs only in the form of slots in objects or events.

Unlike many other classifications with a narrow focus (e.g., Casati and Varzi, 1993; Hayes, 1985; Mars, 1993), our ontology must cover a wide variety of concepts in the world. In particular, our ontology cannot stop at organizing terminological nouns into a taxonomy of objects and their properties; it must also represent a taxonomy of (possibly complex) events and include many interconnections between objects and events to support a variety of disambiguation tasks. As such, concepts in the μ K ontology are far from being atomic symbols; they have a rich internal structure to them.

Just as there is no single grammar that is the “true” grammar of a natural language, it is reasonable to argue that there is no unique ontology for any domain. The μ K ontology is one possible classification of concepts in its domain constructed manually according to a well-developed set of guidelines. Its utility in NLP can only be evaluated by the quality of the translations produced by the overall system or through some other evaluation of the overall NLP system (such as in an information extraction or retrieval test). This is not to say that the ontology is randomly constructed. It is not. Its construction has been constrained throughout by the guidelines as well as by the requirements of lexical semantics and their acquisition.

In NLP work, the term “ontology” is sometimes also used to refer to a different kind of knowledge base which is essentially a strict hierarchical organization of a set of symbols with little or no internal structure to each node in the hierarchy (e.g., Farwell, et al. 1993; Knight and Luk, 1994). Frames in the μ K ontology, however, have a rich internal structure through which are represented various types of relationships between concepts and the constraints, defaults, and values upon these relationships. It is from this rich structure and connectivity that one can derive most of the power

the other.

⁷Although the ontology is not specific to any particular language, we use English words in naming concepts purely for convenience. A name such as “concept-423” would make the ontology unreadable and unusable for the human reader. We have developed a set of guidelines for naming concepts (Mahesh, 1995; Mahesh and Nirenburg, 1995).

of the ontology in producing a TMR from an input text. Mere subsumption relations between nearly atomic symbols do not afford the variety of ways listed above in which the μ K ontology aids lexicon acquisition and disambiguation in language processing.

The above distinction between highly structured concepts and nearly atomic concepts can be traced to a difference in the *grain size* of decomposing meanings. Grain size is a scale that denotes the extent to which a complex meaning is decomposed into more primitive concepts and relationships between them as opposed to representing it by a single concept with little internal structure. For example, the meaning of “to teach” can be represented either by a single concept named TEACH or decomposed into several subevents such as lecturing, question answering, and evaluating, each of which involves several participants such as the teacher, students, a class room, a lesson, and so on. A highly decompositional (or compositional) meaning representation relies on a very limited set of primitives (i.e., concept names). As a result, the representation of many basic concepts becomes too complex and convoluted. The other extreme is to map each word sense in a language to an atomic concept. As a result, the nature of interconnection among these concepts becomes unclear, to say nothing about the explanatory power of the system (cf. the argument about the size of the set of conceptual primitives in Hayes, 1979). Though, presumably, any piece of world knowledge could be useful for NLP, in μ K we take a hybrid approach and strive to contain the proliferation of concepts for a variety of methodological reasons, such as tradeoffs between the parsimony of ontological representation and that of lexical representation and the need for language independent meaning representations. Control over proliferation of concepts is achieved by situated development and a set of guidelines that tell the ontology acquirer when not to introduce a new concept (see Mahesh 1995; Mahesh and Nirenburg, 1995). The μ K ontology is not limited to its domain but is more developed in the chosen domain.

The μ K ontology also makes a clear distinction between conceptual and episodic knowledge and includes only conceptual knowledge. Instances and episodes are acquired in a separate knowledge base called the *onomasticon*. The methodology for acquiring the onomasticon includes a significant amount of automation and is very different from ontology acquisition, which is done manually via continual interactions with lexicographers.

2.1 Ontology Size and Quality: Some Numbers

We are currently in the process of a massive acquisition of objects, events, and their properties related to the domain of company mergers and acquisitions. Over the period of about three months, the μ K ontology has acquired over 2000 concepts organized in a tangled hierarchy with ample interconnection across the branches. Figure 4 shows the rate of growth of the ontology over the last eight months. This graph shows our initial acquisition phase starting from an older ontology developed at Carnegie Mellon University (Carlson and Nirenburg, 1990), an intermediate clean up phase when we deleted hundreds of questionable and unrelated concepts and the current phase of massive acquisition.

The ontology emphasizes depth in organizing concepts and reaches depth 10 or more along a number of paths. The quality of the semantic classification in the ontology is measured by several

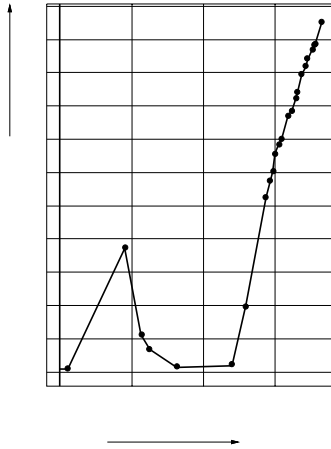


Figure 4: Rate of growth of the μ K ontology.

parameters which are monitored continually with the help of computer programs. For example, the branching factor is kept less than 5 at most points. Each concept has, on average, 5 to 10 slots linking it to other concepts or literal constants. The top levels of the hierarchy have proved very stable as we are continuing to acquire new concepts at the lower levels.

In parallel, we have built a Spanish lexicon of over 3000 words (where each entry is at least as elaborate as the entry shown in Figure 5) each of which maps to one or more of the over 4000 concepts in the ontological world model. These concepts cover a wide variety of categories (but with particular emphasis on the domain of mergers and acquisitions of companies). Each concept has links to 15 other concepts on an average, making the ontology a richly connected network of the kind ideally suited to the search algorithm we employ for checking constraints.

μ K is able to process an unedited Spanish news article and produce TMRs of reasonably good quality as judged by native Spanish speakers and expert Spanish linguists. The TMRs produced by μ K are evaluated by comparing them against “golden” TMRs for the same texts produced by hand by an independent team of linguistic semanticists.⁸ We have so far tested the system thoroughly on three texts and produced TMRs for all the sentences in the texts. A second, large scale testing and TMR production effort has just been started. By the end of the year 1995, we expect to have acquired over 7000 entries in the Spanish lexicon supported by about 5000 concepts in the μ K ontology and to have completed testing the system on up to 10 article-length texts. The above sizes of the lexicon and the ontology are sufficient to support the processing of over 400 Spanish texts on mergers and acquisitions that we have in our corpus.

⁸Several metrics, such as coverage and correctness of the TMR produced by μ K with respect to the “golden” TMR, are being used for the ongoing evaluation. Results from such evaluations for at least three texts will be presented at the Workshop.

3 Ontology in Use: Aiding NLP

Returning to the example described earlier, the first step in processing the input sentence is to recognize word boundaries and analyze the sentence morphologically and syntactically. This is done in μK using the Panglyzer Spanish analyzer developed in the Pangloss project (Pangloss, 1994). The output of such analyses is a syntactic structure of this fairly complex sentence. Panglyzer retrieves entries from a Spanish lexicon for the words in the sentence and uses syntactic information therein to build the syntactic structure of the sentence.

In order to produce the meaning representation given the syntactic structure, μK uses both semantic knowledge represented in the Spanish lexicon and world knowledge represented in a language-independent ontology. The lexicon represents meanings of words by mapping them to concepts in the ontology. In addition, it also specifies syntax-semantics mappings by binding syntactic arguments to fillers of semantic roles in the slots of the ontological concept. A text meaning representation (TMR) is the result of instantiating concepts from the ontology that correspond to the chosen senses of words in a text and linking them together according to the constraints in the concepts as well as the syntax-semantics mappings represented in the lexicon entries. Skeletal TMRs thus constructed are also enhanced by various microtheories which are specialized experts carrying different types of knowledge of the language such as microtheories of space, time, aspect, speaker attitudes, and so on.

Figure 5 shows one of the lexical entries for the Spanish word “adquirir,” the root form of “adquirió” in sentence (1). This entry, in its *lex-map* zone, maps to the concept named ACQUIRE in the ontology and binds the syntactic arguments of the verb “adquirir” ($\$var1$ and $\$var2$ in the *syn-struct* zone of the entry) to the agent and theme roles of the ACQUIRE event. The ontological concept for the ACQUIRE event is shown in Figure 6 and has constraints on the fillers of agent, theme, and other slots represented using other concepts in the ontology. For example, the agent must be filled by a HUMAN.

There is a second entry for “adquirir” in the Spanish lexicon corresponding to a different sense of the word that maps to the LEARN concept in the ontology. It is one of the jobs of the μK semantic analyzer to select the right sense of ambiguous words such as “adquirir.” In this example, the analyzer picked ACQUIRE, which is the appropriate sense in sentence (1), using ontological information as explained below. Other examples of ambiguous words in this sentence can be found in “compañía” and the preposition “a-través-de.” “Compañía” means either a CORPORATION or an INTERACT-SOCIALLY event. Similarly, “a-través-de” has a spatial location meaning and an instrument meaning. Similarly, “en” and “Doctor Andreu” are also ambiguous.

μK is able to choose the appropriate meaning of a word by combining information from its linguistic and world knowledge sources. For example, in the case of “adquirir,” the analyzer instantiates both the ACQUIRE and LEARN concepts and sets up constraints on their slot fillers. These constraints come from both the *lex-map* zone of the lexicon entry for the word “adquirir” and the slots in the ACQUIRE and LEARN concepts themselves. After identifying possible fillers per the syntax-semantic variable mappings specified in the lexicon, the analyzer checks each constraint and assigns a score to it. A constraint is checked by determining the proximity of potential fillers to

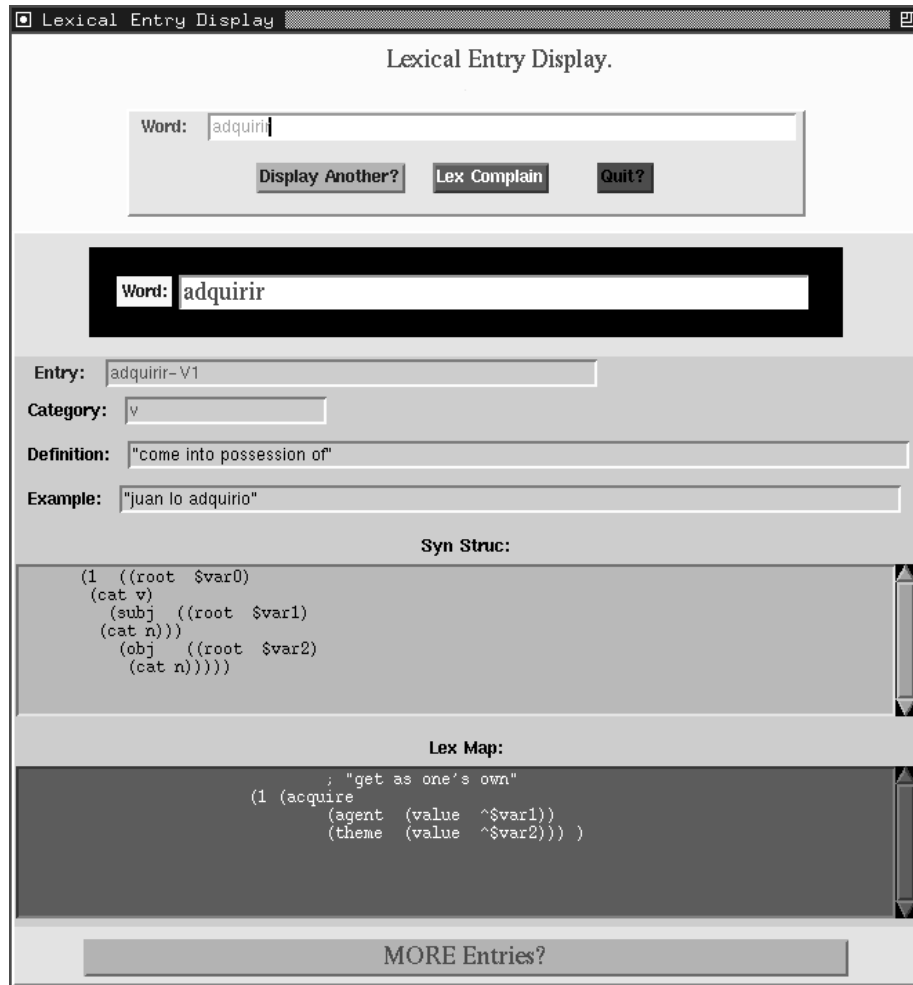


Figure 5: A lexical entry for the Spanish verb “adquirir” with its semantic mappings to the ACQUIRE event.

the specified constraint within the ontological network.

The theme of ACQUIRE must be an OBJECT other than HUMAN while the theme of LEARN must be INFORMATION. These constraints set up search tasks for the closeness of an ORGANIZATION (the potential filler being the Doctor Andreu organization) and each of PHYSICAL-OBJECT and INFORMATION. It turns out that the former is “closer” to ORGANIZATION than the latter and hence gets a higher score. Using this score and combining it with scores from all the other constraints on the meanings in a sentence, the μK analyzer selects the ACQUIRE meaning of “adquirir” in the TMR for sentence (1). The analyzer does this search in the space of all the constraints in an efficient best-first manner using dependency analysis (Beale, Nirenburg, and Mahesh, 1995). The ontological search method to determine the “closeness” of a pair of concepts is also valuable in figuring out the meaning of a complex nominal (such as a compound noun) and in processing metonymies and other nonliteral expressions.

Ontology Concept Display

Ontology Concept Display.

Enter Concept Name or Keyword:

Concept Name:

DEFINITION		
VALUE	The transfer of possession event where the agent transfers an object to its possession.	
TIME-STAMP		
VALUE	created by mahesh at 17:36:28 on 03/13/95 lori at 09:43:24 on 05/09/95 lori at 11:44:27 on 05/16/	
IS-A		
VALUE	TRANSFER-POSSESSION	
SUBCLASSES		
VALUE	INHERIT	
AGENT		
SEM	HUMAN	
PRECONDITION-OF		
SEM	OWN	
SOURCE		
SEM	HUMAN	
PURPOSE-OF		
SEM	BID WIN	
INSTRUMENT		
SEM	PHYSICAL-OBJECT EVENT	
THEME		
SEM	OBJECT (NOT HUMAN)	
LOCATION		
SEM	PLACE	
DESTINATION		
SEM	PLACE ANIMAL	
HAS-PARTS		
VALUE	TRANSFER-OBJECT	

Figure 6: Frame representation for the concept ACQUIRE.

4 Ontology Acquisition: Methodology

A situated ontology such as the μ K ontology is best developed incrementally, relying on continuous interactions with other knowledge sources. In practice, this translates into the concurrent development of both the ontology and the lexicon through a continual negotiation. This negotiation to meet the constraints on both a lexical entry and a concept in the ontology leads to the best choice in most cases. It also ensures that every entry in each knowledge base is consistent, compatible with its counterparts, and has a purpose towards the ultimate objective of producing quality TMRs. Though there is no algorithm for acquiring concepts, sets of guidelines have been developed in the Mikrokosmos project for deciding (a) what concepts to acquire, (b) where to place a concept in the hierarchies, and (c) what to name a concept (Mahesh, 1995; Mahesh and Wilson, in preparation). A few guidelines for deciding what concepts to add to the ontology are shown in Figure 7.

-
1. Do not add instances as concepts in the ontology. Rules of thumb for distinguishing an instance from a concept are:
 - **Instance-Rule1:** See if the entity can have its own instance. Instances do not have their own instances; concepts do.
 - **Instance-Rule2:** See if the thing has a fixed position in time and/or space in the world. If yes, it is an instance. If not, it is a concept. For example, SUNDAY is a concept, not an instance, because it is not a fixed position in time (“last Sunday,” “the first Sunday of the month,” etc.).
 2. Do not decompose concepts further into other concepts merely because you can. It is important to focus on building those parts of the ontology that are needed immediately for the μ K task. For example, though EVENTS like BUY or MARKETING can be decomposed to a great extent, unless there is an indication that detailed decompositions are needed for the task, do not decompose such EVENTS.
 3. Do not add a concept if there is already one “close” to it or slightly more general than the one being considered. Consider the expressiveness of the representation provided by gradations (i.e., attribute values) before adding separate concepts. For example, we do not need separate concepts for “suggest,” “urge,” and “order.” They are all gradations of the same concept, a DIRECTIVE-ACT, with various degrees of force which can be captured in an appropriate attribute.
 4. Do not add specialized EVENTS with particular arguments as new concepts. For example, we do not need separate concepts for “walk to airport terminal” and “walk to parking lot.”
 5. Certain elements of text meaning such as aspect, temporal relations, attitudes, and so on, that are instance-specific belong only in the TMRs. For example, BREAKFAST is probably a concept in the ontology (and a subclass of MEAL, say) but a meal that happened at 3 O’clock on a particular day is not a separate concept in the ontology.
 6. One must also remember that ontologies are supposed to be language independent. As such, if any part of a meaning representation is specific to a particular natural language that part does not belong in the ontology.
 7. Mikrokosmos representations have a very expressive **set** and **subset** notation. Hence, there is no need to create ontological concepts for collections of different types of things in the world.
-

Figure 7: Guidelines for deciding what concepts to add.

5 Ontology Acquisition: Technology

In order to aid ontology acquisition and maintenance, to check its consistency, and to support interactions with lexicographers, a variety of semi-automated tools have been developed and deployed in the Mikrokosmos project. These are state of the art software tools with easy-to-use graphical interfaces and are particularly useful for visualizing a large, intricate classification system while editing or browsing it. Some of the tools were used to produce the displays captured in the figures in this article. Tools have been developed for:

- browsing the hierarchies and the internals of concepts in the ontology;
- editing graph structures;⁹
- translating between two different representations (e.g., the object oriented representation suitable for computational purposes and the plain text representation that is more suitable for manual search and maintenance purposes);
- checking for consistency both within the ontology and between the ontology and the lexicon and for conformance with the guidelines; and
- supporting interactions with lexicon acquirers through the use of a graphical interface for submitting requests for changes or additions.

6 Conclusion

In our continued efforts to build the Mikrokosmos machine translation system, we have developed an elaborate semantic classification of concepts, or ontology, to facilitate natural language interpretation. The μK ontology is a large-scale classification of a wide variety of meaning elements (called concepts) with a rich interconnection of conceptual relations. We have shown that the semantic classification we developed is of great value in solving ambiguities and other related problems in natural language processing. In addition, we have also developed a methodology for developing the classification as well as a variety of state of the art software tools for editing, browsing, and accessing large classification systems. We are continuing to expand the coverage of the ontology by acquiring more concepts. In addition, we are also looking at new applications for the μK ontology.

Acknowledgements

The authors would like to thank Lori Wilson for doing the bulk of the work in ontology acquisition and other members of the Mikrokosmos team, including Stephen Beale, Evelyne Viegas, Victor Raskin, and Boyan Onyshkevych, for their valuable contributions to the work reported in this article.

⁹The “Mikrokarat” tool Developed by Ralf Brown at the Center for Machine Translation, Carnegie Mellon University, supports complete functionality for editing graph structures in an ontology.

References

- Bateman, J. A. (1993). Ontology construction and natural language. In Proc. International Workshop on Formal Ontology. Padua, Italy, pp. 83-93.
- Beale, S., Nirenburg, S., and Mahesh, K. (1995). Semantic Analysis in the Mikrokosmos Machine Translation Project. To appear in the Proceedings of the Second Symposium on Natural Language Processing (SNLP-95), August 2-4, 1995, Bangkok, Thailand.
- Carlson, L. and Nirenburg, S. (1990). World Modeling for NLP. Technical Report CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.
- Casati, R. and Varzi, A. (1993). An Ontology for Superficial Entities I: Holes. In Proc. International Workshop on Formal Ontology. Padua, Italy, pp. 127-148.
- Farwell, D., Guthrie, L., and Wilks, Y. (1993). Automatically creating lexical entries for ULTRA, a multilingual MT system, *Machine Translation*, vol. 8:3, pp. 127-146.
- Hayes, P. J. (1979). Naive physics manifesto. *Expert Systems in the Microelectronic Age*. Edinburgh: Edinburgh University Press.
- Hayes, P. J. (1985). Naive Physics I: Ontology for liquids. In *Formal theories of the common sense world*, ed. J. Hobbs and R. C. Moore, pp. 71-107. Norwood, NJ: Ablex.
- Knight, K. and Luk, S. K. (1994). Building a Large-Scale Knowledge Base for Machine Translation. In *Proc. Twelfth National Conf. on Artificial Intelligence*, (AAAI-94).
- Mahesh, K. (1995). Ontology Development: Ideology and Methodology. Technical Report (under preparation), Computing Research Laboratory, New Mexico State University.
- Mahesh, K., and Nirenburg, S. (1995). A Situated Ontology for Practical NLP. To appear in Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Aug. 19-20, 1995. Montreal, Canada.
- Mahesh, K. and Wilson, L. (in preparation). Ontology Acquisition: Guidelines and Technology. Technical Report (in preparation). Computing Research Laboratory, New Mexico State University.
- Mars, N. (1993). An ontology of measurement units. In Proc. International Workshop on Formal Ontology. Padua, Italy, pp. 297-303.
- Onyshkevych, B. and Nirenburg, S. (1994). The lexicon in the scheme of KBMT things. Technical Report MCCS-94-277, Computing Research Laboratory, New Mexico State University.
- Pangloss. (1994). The PANGLOSS Mark III Machine Translation System. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT, Jan. 1994.