Why Base the Knowledge Representation Language on Natural Language?

Fernando Gomez
School of Computer Science
University of Central Florida
Orlando, FL 32816
gomez@cs.ucf.edu

Abstract

It[1] is argued that a knowledge representation language intended to be applied across diverse domains must be based on natural language. It is also indicated that such a representation language will facilitate the acquisition of knowledge from natural language and the interaction with other programs in need of obtaining some knowledge. The main aspects of a knowledge representation language based on these ideas are presented, and the results of such a language in its application to the task of acquiring knowledge from encyclopedic texts are briefly discussed.

Key Terms: Knowledge Representation and Acquisition, Natural Language, Cognitive Science.

1. INTRODUCTION

There are too many facts indicating clearly that NL (natural language) cannot be used as a knowledge representation formalism. Syntactic (lexical and structural) ambiguity and polysemy come first to mind. But, even if these problems did not exist, a problem would

remain, namely that the meaning of linguistic utterances depend on the knowledge that the reader brings to bear. Thus, a text of physics with all its ambiguity removed would mean different things to a beginner student than to a physics professor. Natural language sentences do not have meaning by themselves independent of the reader. But, because readers bring different background knowledge, sentences are going to have different meaning to diverse readers. See (Zadrozny 1994) for a recent formalization of background knowledge using partial theories, and (Iwanska 1996) for a view expressing that natural language is representational language.

Notwithstanding these comments, we do think that any knowledge representation formalism aiming at some generality, e.g., the encoding of a large knowledge base applicable across diverse domains and the use of this knowledge base by different programs, must be grounded on NL. Consequently, KR (knowledge representation) begins with lexical semantics and builds from there. The KRL (knowledge representation language) must be modeled with the goal of being used to understand language, to acquire knowledge from language and to solve problems in natural language.

By saying that the KRL must be grounded on NL, we mean that it must be based on the underlying or deep structures of language, rather than on its surface linguistic forms. The KRL must be based on the mentalese underlying language, or in other words, on the logical forms of the sentences. The hypothesis that deep structures (Wittgenstein 1958; Chomsky 1965) underlie linguistic forms, has not been refuted, seems well established and has been able to explain the underlying meaning of diverse surface constructions such as passive and active voice, the different syntactic realizations of the same thematic role, e.g., "Peter cut Jennifer's arm," vs "Peter cut Jennifer on the arm," the fact that syntactic

alternations of verbs fall into classes of identical or similar meaning (Levin 1993) and many others. Basing the KRL on the mentalese of natural language means that the KRL should represent relations of any arity, and embedded relations. Hence, all KL-ONE (Brachman & Schmolze 1985) family of languages and, also, CYC (Lenat & Guha 1989), are not based on natural language because they represent only binary relations, and they need to use reification for representing relations of arity greater than 2 (see below). Logic is not either a formalism based on the mentalese underlying language. First of all, the ontology of logic says nothing about the ontology of natural language: entity, physical-thing, intangible, location, substance, etc. Thus, logic has no answer to the question of which entities are needed in order to comprehend language and acquire knowledge from it. Secondly, the syntax of logic is not that of the mentalese underlying language. In fact, it is quite removed from it, as Jackendoff has pointed out (Jackendoff 1983). This makes it very hard to build logical forms from the sentences, and acquire knowledge from it. Finally, logic centers its semantics on reference, or denotation, which is only one of the two elements in Frege's equation on meaning, the other one being *sense*. Although reference does play a role in natural language, the reduction of meaning to reference yields formalisms that are removed from the mentalese underlying natural language, and, as a consequence, heuristically inadequate for many natural language tasks (see (Wilks 1985) for a critique of the role of reference in natural language). The above critique applies also to hybrid knowledge representation formalisms that contain a strong logic component. A good example of this is *episodic logic* (Hwang & Schubert 1993). These observations should not be perceived as disparaging remarks against the above formalisms, but just as saying that these formalisms are not based on the mentalese underlying natural language. The reader may want to look into

(van Benthem & ter Meulen Alice 1997), a book containing a good collection of articles arguing for the role of logic in natural language. Most of the papers in this book emphasize the role of truth-condition semantics, and, although some (Partee 1997) discuss other views briefly, most of them do not address the serious problems with a view of natural language semantics that reduces meaning to reference.

One of the strongest arguments in favor of basing the KRL on mentalese is that there is an ontology underlying natural language. This ontology is already present very early in our mental development as the studies of Keil (Keil 1989) and Carey (Carey 1985) have shown. Children recognize sentences in which category violations have taken place, e.g., "The table ate the apple," and react to them with expression such as "silly," but not "wrong." It is also becoming apparent that ontological categories play an essential role in solving many difficult problems in semantic interpretation (Gomez, Segami, & Hull 1997). For instance, if one is presented with the sentence "Peter ate a cake with X," and asked which constituent in the sentence the PP "with X" modifies, the answer depends on the ontological category of X. If X is food, the attachment would be to "a cake," but if X is a utensil, the attachment would be to the verb.

Natural language ontology (see (Miller *et al.* 1993) for a comprehensive ontology of English) contains a rich set of categories to which everyone relates, which allows us to un-

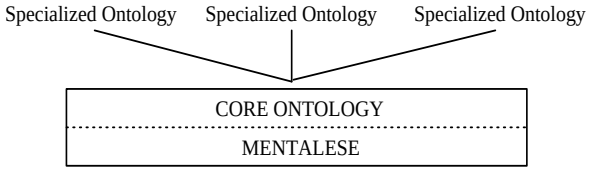

Figure 1: Relation between Ontologies

derstand each other most of the time, to learn a new language, and to translate languages without radically different ontologies (despite Quine's objection (Quine 1960) that translation is impossible) if such languages exist as Whorf claimed (Whorf 1956). A reasonable hypothesis is that more specialized ontologies, say the ontology that a physician needs in order to diagnose a liver disease, are based on this basic and more general ontology, as depicted in figure 1.

It is also reasonable to assume that the reason why a physician can communicate her ontology to a layman or to a beginner student is because they share this basic ontology. The above observations are completely against the view that natural language is just the interface to some deeper representation. If by "natural language" what is meant is syntax, then, yes, the above view is correct. But, if cognitive scientists agree on anything these days, it is that NL is not just syntax. The disagreements begin in assessing the relevance of the roles of syntax and semantics. The view we are proposing is that the answer to basic problems, not only in knowledge representation, but also in problem-solving and knowledge-acquisition, or learning, may lie in the processes and deep structures that allow us to understand natural language. The advantages of a KRL based on the deep structures underlying sentences are multiple. The most obvious one is that the acquisition of knowledge from natural language could be more easily automated, since the building blocks of the KRL would be logical forms. One of the most pervasive problems in KR is that different people provide diverse representations of the knowledge contained in a short paragraph. One of the reasons for this is that we theorize facts in different ways. I have asked my students to represent the knowledge in a short paragraph using CYC (Lenat & Guha 1989), and, in most cases, they provide representations that would be very hard to reconcile, as a result

making it very difficult to share these representations with other users. This problem would disappear if the system were able to automatically represent and integrate the knowledge in the paragraph.

Another immense advantage is that for other programs (or "agents," to use the fashionable expression these days) interfacing with this KRL would be greatly simplified, because it could take place in natural language. These programs would need a very simple English generator, which would generate an English question from their internal representation. The general purpose KRL would build the logical form of the question, and will come with an answer. A taxonomy of answer types corresponding to the type of questions can be made known to both the agents and this KRL. Specialized KR languages could be also developed as refinements of this general purpose KRL.

In the following sections, we briefly present a KRL based on these ideas, and provide some results. The closer KRLs to this language are SNePS (Shapiro & Rapaport 1992), and conceptual graphs (Sowa 1984). The major differences between our KRL and those languages are our distinction between *a-structures* and *object-structures* , quantifiers, and the representation of existentially quantified sentences and restrictive modifiers as classification hierarchies. This language has been under construction for some years now, and is being used by SNOWY, a system designed for the acquisition of knowledge from expository texts.

2. FROM LOGICAL FORMS TO FINAL KNOWLEDGE REPRESENTATION STRUCTURES

Let LF denote the logical form constructed for a sentence by a semantic interpreter. LF

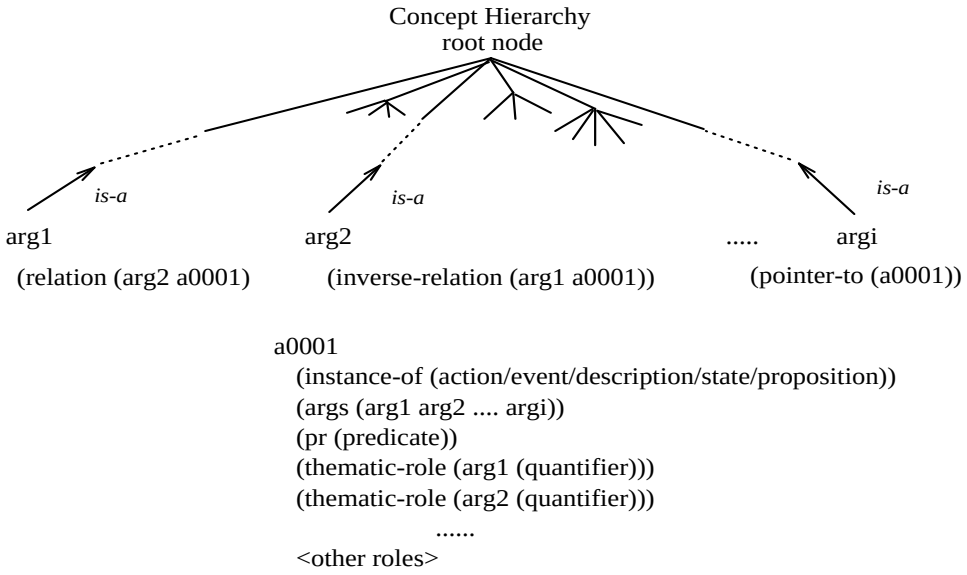

Figure 2: Schema for Representing Relations

consists of a predicate and some thematic roles. By "thematic roles" we refer not only to arguments in Government and Binding theory (Chomsky 1981; Grimshaw 1990; Jackendoff 1990), but also to locative and temporal adjuncts. From a knowledge representation view, we have a relation, the predicate of the LF, and a set of arguments, the thematic roles of the logical form. Every argument of the relation is indexed as an object, which is then linked through *is-a* relations to the other objects in the hierarchy of concepts. Links are also created connecting the objects to the relation itself, which is represented in a separate structure, called an *a-structure*. Thus, a relation R(arg1, arg2, ... argi) is represented as indicated in figure 2.

The second argument of the relation is indexed by using its inverse relation. Other arguments are indexed by just a pointer to the action structure. The node *a00001* has stored in it the representation of the relation itself. The slot *instance-of* contains the type of relation. The slots *args* and *pr* contain the arguments of the relation, and the predicate

of the logical form, respectively. Thematic roles are listed followed by their quantifiers. An example will clarify this. The representation of "Robins eat berries" is shown in figure 3.

The quantifier slots are filled with question marks, because the sentence does not specify them. If the sentence had been "Most robins eat few berries," the content of the quantifier of the *actor* and *theme* would have been *most* and *few*, respectively. *A-structures* allow the representation of a varied class of quantifiers which may be used by inductive algorithms. The scoping of the quantifiers is from left to right. Thus, if the quantifier of the *actor* were "all," and that of the *theme* "some," the meaning of the structure a1 in figure 3 in first-order predicate calculus would be:

$\forall(x)(Robin(x) \Longrightarrow \exists(y)(Berry(y) \ \wedge \ Ingest(x,y)))$

Some of the arguments of an *a-structure* may be relations themselves. That would be the case for the representation of "Everyone wants to own a house," in which the second argument of the relation *want*, the *goal* role, is a relation itself. The fact that the variable of the quantifier of the first argument of *want* is the same as that of the first argument of *own* is captured by using what we have called an "index quantifier." The variable of index quantifiers cannot be instantiated with different individuals in the *a-structures* where they occur, but only with the same individual. The problem that we are discussing is the diverse representation that needs to be given for "Everyone wants to own a house":

$\forall(x)(Human(x) \Longrightarrow \exists(y)(House(y) \ \wedge Want(x, own(x,y))))$

and "Everyone wants everybody to own a house":

$\forall(x)(Human(x) \Longrightarrow \exists(y)(House(y) \wedge \forall(z)(Human(z) \Longrightarrow \ Want(x, own(z,y)))))$

See (Gomez & Segami 1991) for a detailed discussion of this and other problems discussed briefly in this section. *A-structures* can also be connected to other *a-structures* by temporal

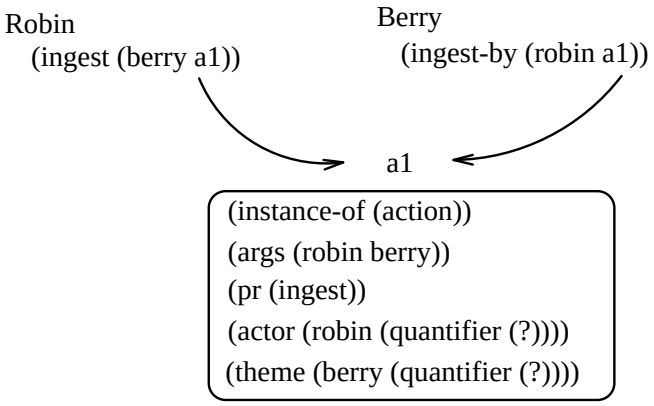

Figure 3: Representation of "Robins eat berries."

and other type of links, as required in the representation of "Bears may attack humans, if they are nursing," or "Warblers help farmers by killing insects that destroy fruits." A great advantage of *a-structures* is that the representation of relations of arity greater than 2 become rather simple, since it reduces to the addition of new arguments to the relation. Thus, the representation of *Robins eat berries in the fall* is achieved by adding a slot *at-time* and "fall" as a third argument to the *a-structure* depicted above. The representation of relations with arity greater than 2 is rather involved in many knowledge representation languages. That is certainly the case of KL-ONE (Brachman & Schmolze 1985) and some of its descendants (Brachman *et al.* 1991; MacGregor 1993) in which one needs to recur to reification to represent relations with arity greater than 2. It is also a problem in CYC, in which it is very awkward to represent predicates of arity greater than 2. The hypothesis expressed in CYC that there are few predicates of arity higher than 2, and that this fact may reflect human cognitive limitations is not supported by our observations of encyclopedic texts. In fact, we have found that predicates of arity higher than 2 abound and they form part of our everyday life. Consider, "A spider stabs an insect with its chelicerae," "The

fishermen caught the cod with illegal nets,' "Peter loaded the hay into the truck with a crane," etc. If one wants to further complicate things, one only needs to add some temporal and locative adjuncts to these examples.

3. REPRESENTING PHRASAL CONCEPTS

We now turn our attention to the representation of complex noun groups and, in general, any restrictive modifiers. The problem is how to represent concepts expressed by "spiders with a cribellum," "mammals that live in the sea," "sea mammal" "eagles that prey on monkeys," "Floridians who live on the West Coast," etc. Fodor (Fodor 1981) has referred to concepts expressed by restrictive modifiers as phrasal concepts and has distinguished them from single-word concepts such as "whale," "mammal," etc. We will use Fodor's term to refer to them. This is the type of concept that students tend to hyphenate, and as a result they forestall all inferences based on the relations connecting the single word concepts that form these complex concepts. Consider the following passage:

> *Spiders that live under water breathe from air bubbles. These spiders are found in Europe and parts of Asia.*

If one observes this passage with an object-centered representation in mind, one has to conclude that the concept "spiders that live under water" needs to be represented as an object in memory so that subsequent information about it can be integrated on that node. An entire paragraph may follow the first sentence describing these spiders and, probably, classifying them further in subclasses. In our representation language, phrasal concepts are represented as classification hierarchies. The complex description of a phrasal concept is represented as a single concept by the following method. A dummy name is created for

the phrasal concept, and, then, the relations that define it are specified in a special slot. For instance, "mammals that live in the Antarctic" is represented by creating the concept, say, *X1*, a gensym in the actual program, which is defined by saying that it is a class of mammals that live in the Antarctic." We have:

```
X1 (cf (is-a (mammal) (a2)))
```

The slot *cf*, which stands for characteristic features, contains the necessary and sufficient conditions that define the concept X1. The relation "live in the Antarctic" predicated of X1 is represented in the *a-structure* a2. The meaning of the representation for X1 in first-order predicate calculus is:

$$\forall(x)(X1(x) \Longleftrightarrow Mammal(x) \wedge Live-in(x, Antarctic))$$

The name given to this concept is arbitrary, and is not used for recognition purposes. The representation of the concept X1 is completed by creating the slot *classes-of*, the inverse of *is-a*, in the concept "mammal" and by filling it with X1. Had the slot *classes-of* existed in the concept "mammal" then X1 would be inserted into it. A recognition algorithm checks the content of the *cf* slot in order to establish if two phrasal concepts are the same concept. This representation permits the collapse of very complex descriptions of concepts into one concept. For instance, the noun group "the gills of fish," can be represented as (we have specified the relation, *body-part-of*, in the *cf* slot, rather than writing the *a-structure* gensym.):

```
Y1 (cf (is-a (gill)) (body-part-of (fish)))
```

Thus, "the membranes in the gills of fish," is represented as:

```
Y2 (cf (is-a (membrane)) (body-part-of (Y1)))
```

Those logical forms whose first argument is existentially quantified are also represented as classification hierarchies in a similar manner as we have done for phrasal concepts. The sentence "Some mammals live in the Antarctic" introduces the concept "mammal that live in the Antarctic." If that sentence is represented by means of an *a-structure*, and the next sentence is "These mammals are endangered," then a node does not exist in memory, corresponding to "these mammals," in which to integrate the fact that they are intelligent. Thus, the representation of "Some mammals live in the Antarctic" is (where a2 is the relation "live in the Antarctic" predicated of X1):

```
X1 (cf (is-a (mammal)) (a2))
```

We have found beautiful texts illustrating the need for representing existentially quantified sentences as classification hierarchies. For instance in the following text, children are asked to write the classification of animals presented in the text. The first sentence in the paragraph introduces the class of unusual birds.

> Some birds are unusual because they cannot fly. The emu, the rea, and the ostrich are such birds. The emu is an Australian bird. The rea lives in South America. The ostrich lives in Africa and runs very fast.

4. ACQUIRING THE KNOWLEDGE AUTOMATICALLY

The representation structures briefly explained permit the automated acquisition of knowledge from natural language. Besides the obvious need for a parser and a semantic interpreter, the following components are needed: a Formation algorithm that constructs the final knowledge representation from the logical forms, and an Integration algorithm which integrates the KR structures in LTM (long-term memory). The Integration algorithm needs

to rely on Recognition and Classification (Brachman & Schmolze 1985) algorithms. The Formation algorithm transforms the logical forms into the final knowledge representation structures. This is a deterministic mechanism, which is rather simple because restrictive modifiers are already collapsed into single concepts by our semantic interpreter. Thus, "All bats that live in the tropics have food all year around," becomes "all X1 have food all year around," where X1 contains the representation of "bats that live in the tropics."

The Recognition algorithm is needed so that concepts and relations already in LTM are not stored again under different nodes. Thus, the Integration algorithm only stores in LTM those concepts and relations that the Recognition algorithm fails to recognize. The recognition of single word concepts is a rather trivial task, but the recognition of phrasal concepts, those represented by using a *cf* slot, is more challenging, requiring the need of the Classifier. This activates an algorithm, called Compare-Classes that returns for two given phrasal concepts, say X1 and X2, whether X1 is a subconcept of X2, X2 is a subconcept of X1, both X1 and X2 are subconcepts of each other, or there is no subsumption relation between X1 and X2. Two concepts are identical if they are subconcepts of each other. Because the KRL allows for the representation of n-ary relations, the Compare-Classes algorithm has a more difficult task than if the relations were only binary. However, we have not noticed any delay in time in the tests we have performed. (See discussion of results below.) The algorithm classifies the concept *Americans that eat fish with forks* as a subconcept of *Humans that eat fish with utensils*, if "American" and "fork" are in LTM as subconcepts of "human" and "utensil," respectively.

Because the KRL allows for a diverse class of quantifiers, relations in superconcepts cannot be automatically inherited by subconcepts. Only those relations whose arguments are

universally quantified, or those that belong to a *cf* slot are inherited by subconcepts. Thus, the relation "eat plankton" in the concept "whale" cannot be inherited by the concept "beluga," a subconcept of "whale." The Classifier treats all quantifiers that are not universal as existential quantifiers. However, some inductive algorithms that are now under construction do use the content of these quantifiers to establish some of the inferences.

5. INFERENCING

There are three levels of inferencing in the KRL. The Classifier, which keeps LTM organized, is the main inference mechanism within the KRL. Because those logical forms whose first argument is existentially quantified are represented as classification hierarchies, the Classifier allows the deduction of complex inferences by just keeping LTM organized. For instance, SNOWY will infer that some parrots are endangered from "All parrots are birds. All birds that live in the rain forest are endangered. Many parrots live in the rain forest." This inference is possible because "Many parrots live in the rain forest" is represented as a phrasal concept, namely *X2(cf(is-a(parrot)) (aj))*, where *aj* is the relation "live in rain forest." The Classifier will insert this concept under X1, "birds that live in the rain forest" and under the concept "parrot," as indicated in figure 4.

As is standard, the Classifier is also used by the Question-Answering component to answer questions about concepts that do not exist in LTM, as would be the case if one asks the question "Are whales that live in the Antarctic endangered?" given "All whales are animals. All animals that live in the Antarctic are endangered," in which the concept "whales that live in the Antarctic" does not exist in LTM.

A second type of deduction is based on inference rules with all their variables universally

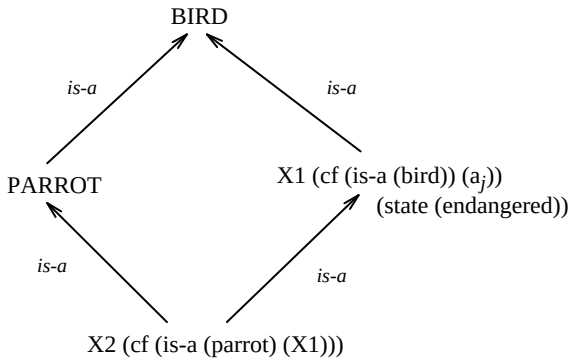

Figure 4: The representation of "All parrots are birds. All birds that live in the rain forest are endangered. Many parrots live in the rain forest."

quantified. The knowledge in sentences such as "Animals eat what they like," or "The diet of raccoons is the same as the diet of bears" require rules to be captured. The KRL has some of these rules stored in the hierarchy of initial nominal concepts and some in the verbal concepts. Thus, the rule

```
if human(x?) and food(y?) and like(x?,y?) then ingest(x?,y?)
```

is stored under the concept "human" and is inherited from all its subconcepts. Other rules are stored on the verbal concepts. The rules below are stored on the *produce*, *ingest*, and *cause* predicates, respectively. These rules are fired in a backward fashion.

```
R1 if produce(?x,?y) and bad-for(?y,?z)
   then bad-for(?x,?z)
R2 if ingest(?x,?y) and at-loc(?z,?y)
   then enter(?x,?z)
R3 if cause(?x,?y,?z) and bad-for(?y,?z)
   then bad-for(?x,?z)
```

These rules are being used not only to establish inferences, but also to establish explanatory connections between sentences. For instance, rules R1, R2 and R3 are used to find out the connections between the first and second sentences in the examples (1), (2) and (3)

below.

(1) Cars are bad for people. They produce pollution.
(2) Many germs enter animals. They are in the food they eat.
(3) Some germs are bad for humans. They cause infections in humans.

Learning the connection linking each pair of sentences is essential if the system is to answer the question "Why are cars bad for people," or "How do germs enter animals." We have categorized these rules into two not dichotomous classes, called analytical and empirical rules, based on the philosophical distinction between analytical and empirical sentences. (See Carnap's *Meaning and Necessity* (Carnap 1956) for a defense of the distinction and Quine's *Two Dogmas* (Quine 1953) for a critique.) In (Gomez 1996), the reader may find a taxonomy of explanatory links connecting sentences and a discussion of the knowledge needed to identify the connections.

Recently, when we applied the system to the acquisition of knowledge from encyclopedic texts, we represented and organized the rules differently in order to speed up the system. In this context, it became essential to infer that bald eagles eat coots from reading the sentence "Bald eagles catch coots," or that bears eat roots from "Bears dig up roots." This is important if the system is reading an encyclopedic article, for building a knowledge-base on the "fly" about the diet of animals, in order to answer a user's question about what eagles or bears eat. In (Gomez, Hull, & Segami 1994) one can find a classification of verbal concepts related to the diet of animals and the inferences that these verbal concepts may trigger. The representation of the inference for the predicate *seize-animal* is given in figure 5. The inference is in the slot *addition-rules*. We have indicated the selectional restrictions for the *subject* and *direct object* of the sentence. The inference only takes place if the actor

```
(seize-animal (is-a (action))
  (subj (animal (actor))
  (obj (animal (theme)))
  (addition-rules
     ((if% (not is-a actor human))
         (infer (actor ingest theme))))
       <other rules>))
```

Figure 5: The representation of the inference for the predicate seize-animal

of the predicate is not a subconcept of "human," because humans do not necessarily eat those animals that they catch. These inference rules are inherited by verbal subconcepts from verbal superconcepts.

6. RESULTS

In April 1994, we applied the model briefly described here to the acquisition of knowledge from encyclopedic texts. The encyclopedia chosen was the *The World Book Encyclopedia* (World Book, Inc., Chicago, 1987.), which is one or two levels less complex than the *Collier's Encyclopedia*. The goal was to build a knowledge base about the diet and habitat of animals by reading an article in the encyclopedia, and then answering questions using the knowledge base constructed. The entire encyclopedia is accessed from a CD-ROM drive, and the system was trained using 10 articles. Then, tests were conducted by selecting articles randomly taken from the encyclopedia. A preprocessor was used to recognize proper names, dates and some punctuation symbols. Then, a skimmer selected sentences related to the topics. Those sentences chosen by the skimmer were then parsed, interpreted, formed, recognized and integrated into LTM. The parser contains a lexicon of 70,614 words as of this writing. The system accesses the data in the lexicon through a client and a server, which are written

in Perl. Each instance of the system starts its own client and captures its input and output streams using a Lisp function. The server stores each word and its file position in a dbm file. When SNOWY needs to look up a word, it first checks its cache. On a miss, Snowy asks its client to get the word from the server. The server gets the word's file position from the dbm file and seeks directly to the word in the lexicon. The parser is a top-down algorithm that uses two stacks to hold the non-terminal symbols and the actions associated with them, respectively. Parsing is driven mostly by the word subcategorizations that are stored in the lexicon. These subcategorizations are pushed onto the stacks as they are encountered in the sentence. The parser resolves all gaps resulting from relative clauses and questions. But, it does not resolve modifier attachment, which is the task of the semantic interpreter. During parsing, the parser uses a built-in tagger to resolve morpho-syntactic ambiguity. The tagger is built as a set of categorical rules that access the content of the stacks and the sentence being parsed to decide about the category of a word. As of this writing, the parser is parsing correctly 82% of the sentences of the encyclopedia chosen randomly from any article. The parser produces a partial parse, meaning a parse in which modifier attachment is not resolved. Modifier attachment is the sole task of the semantic interpreter (Gomez, Segami, & Hull 1997), which deals with lexical ambiguity, attaches modifiers and determines thematic roles. The average time taken to entirely process a sentence, including skimming, parsing, interpretation, formation, recognition and integration, was about 3.2 seconds on a SPARC Classic machine (microSPARC 50MHz CPU) running Allegro Common Lisp. The average length of a sentence in the encyclopedia is 16 words. In recent tests, the average time is 1.2 seconds on a *Sparc 5* machine. A detailed discussion of these results can be found in (Gomez, Hull, & Segami 1994). The system was able to integrate into LTM 299, 236

and 159 concepts and relations related to the diet and habitat of animals from the articles on bats, eagles and monkeys. A list of questions about different encyclopedic articles about animals and the system's answers follows. The system reads these articles without knowing any subclasses of the animals mentioned in its answers. It built these subclasses by reading the articles.

- *What do birds eat?* sapsucker ingest tree-sap; hummingbird ingest nectar; duck ingest plant-matter, grass, seaweed; louisiana-water-thrush ingest water-insect; young-bird ingest earthworm, insect, small animal; .....

- *Which birds eat nectar?* hummingbird ingest nectar

- *What kinds of insect eaters are there?* chickadee, creeper, flycatcher, kinglet, swallow, swift, thrasher, titmice, vireo, warbler, woodpecker, owl.

- *What is gravel?* I don't know, but I know that: bird ingest gravel < related-to > bird *assist* grinding-process

- *What kills birds?* eagle is-a bird, and hunters and trappers kill eagles; osprey is-a bird, and hunters and trappers kill osprey

- *When do most birds search for food?* at-time day

- *Do birds help people?* yes, bird help farmer

- *How do birds help farmers?* bird ingest <insect which ingest crop>; bird ingest weed-seed

- *Where do golden eagles live?* golden-eagle live-in *at-loc* asia, europe, mexico, north-america.

- *Where do most golden eagles live?* most golden-eagle live-in *at-loc* north-america.
- *Where do crowned eagles live?* <crowned-eagle live-in *at-loc* africa > is-a eagle, and all <crowned-eagle live-in *at-loc* africa > inhabit *at-loc* rain-forest.
- *Which bats live in tropical regions?* <bat size-r largest>, fruit-bat live-in *at-loc* most <tropical-region part-of africa>.
- *Do bats live alone?* yes, because hoary-bat is-a bat, and hoary-bat live-in *accompany* alone *at-loc* tree.
- *Do bats eat blood?* yes, some bat eat blood because vampire-bat is-a bat and vampire-bat ingest blood
- *How much blood do vampire bats eat?* vampire-bat ingest blood quantity 1 tablespoon *frequency* day
- *Do vampire bats attack human beings?* yes, vampire bat harm human *frequency* sometimes
- *Do monkeys have enemies?* yes, some monkey has-enemy cheetah hyena jackal leopard lion because <monkeys inhabit at-loc ground> has-enemy cheetah hyena jackal leopard lion
- *Which bears live in alaska?* alaskan-brown-bear, grizzly-bear live-in *at-loc* alaska.
- *How long do beetles live?* adult-beetle live-exist *duration* week (q (few))
- *Which sharks live in fresh water?* bull-shark is-a shark, and bull-shark live-in *at-loc* <water property-r fresh >.

7. CONCLUSIONS

In conclusion, we have argued that a KRL must be based on the representation structures and reasoning mechanisms necessary to understand natural language and acquire knowledge from it. The KR structures described are biased by the task that they intend to serve, namely acquiring knowledge from natural language. We have also briefly described a KRL based on this view. Finally, we have provided some results of this KRL in the application of acquiring knowledge from encyclopedic texts. Other results can be found in (Hull & Gomez 1998) in which the methods are applied to the acquisition of biographic knowledge from the *The World Book Encyclopedia*. Because SNOWY does not have specialized knowledge, we have used to our advantage the fact that the *World Book Encyclopedia* does not presuppose specialized knowledge in order to be understood, and, consequently, to acquire knowledge from it. The same cannot be said of other encyclopedias such as the *Britannica*, which assumes more knowledge on the part of the reader. It would be interesting to investigate if SNOWY could understand an article, say, on spiders in the *Britannica* by using the knowledge first acquired by reading the spider article in the *World Book Encyclopedia*.

Our goal is to acquire knowledge from encyclopedic texts in a way that is usable not only by SNOWY, but also by any other program, or a human, requesting pieces of knowledge. Consequently, the KR structures presented here have, of course, limitations. To name just one, they are inadequate to represent knowledge about processes and systems. For instance, consider the following passage describing fish respiration:

> Fish take water into the mouth. From the mouth, water enters the pharynx. From the pharynx, water passes into the gills. As water flows over the gills, oxygen in the water passes through the membranes in the gills and combines with blood, which releases

carbon dioxide. Carbon dioxide combines with water in the gills. Then, water is expelled out of the gills.

Now, consider the question "Which places does the water pass through before getting to the gills?" This question can hardly be answered unless the events describing fish respiration are grouped and indexed together. Moreover, temporal questions, e.g., "What occurs when water passes over the gills?" and what-if questions, e.g., "What would occur if water does not flow into the gills?" could not be answered at all. We have extended the KRL with temporal event structures based on Allen's interval logic (Allen 1984) to capture this type of knowledge (Gomez 1998). The reasoning mechanisms that the representation permits are, however, based on graph traversal rather than on logic.

Acknowledgments

I am grateful to Richard Hull for comments on an earlier draft of this paper. This research has been funded by NASA-KSC Contract NAG-10-0120.

References

Allen, J. F. 1984. Towards a general theory of action and time. *Artificial Intelligence* 23:132–154.

Brachman, R., and Schmolze, J. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science* 9:171–216.

Brachman, R.; McGuinness, D.; Patel-Schneider, P.; and Resnick, L. 1991. Living with CLASSIC: When and how to use a KL-ONE-like language. In Sowa, J., ed., *Principles of Semantic Networks*. Morgan Kaufmann Publishers, Inc.

Carey, S. 1985. *Conceptual Change in Childhood*. Cambridge, Mass.: MIT Press.

Carnap, R. 1956. *Meaning and Necessity*. Chicago: University of Chicago Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht, The Netherlands: Foris.

Fodor, J. 1981. The present status of the innateness controversy. In *Representations.* Cambridge, Mass.: MIT Press.

Gomez, F., and Segami, C. 1991. Classification-based reasoning. *IEEE Transactions on Systems, Man, and Cybernetics* 21:644–659.

Gomez, F.; Hull, R.; and Segami, C. 1994. Acquiring knowledge from encyclopedic texts. In *Proc. of the ACL's 4th Conference on Applied Natural Language Processing, ANLP94*, 84–90.

Gomez, F.; Segami, C.; and Hull, R. 1997. Determining prepositional attachment, prepositional meaning, verb meaning and thematic roles. *Computational Intelligence* 13(1):1–31.

Gomez, F. 1996. Acquiring intersentential explanatory connections in expository texts. *International Journal of Human-Computer Studies* 4(1):19–44.

Gomez, F. 1998. A representation of complex events and processes for the acquisition of knowledge from texts. *Knowledge-Based Systems* 10 no.2:237–251.

Grimshaw, J. 1990. *Argument Structure.* Cambridge, Mass.: MIT Press.

Hull, R., and Gomez, F. 1998. Automatic acquisition of historical knowledge from encyclopedic text. In *Proc. of the 11th Knowlege-Acquisition Workshop*, 1–18.

Hwang, C. H., and Schubert, L. K. 1993. Meeting the interlocking needs of LF-computation, deindexing, and inference: An organic approach to NLU. In *Proc. of 13th Int'l Joint Conference on Artificial Intelligence*, 1297–1302.

Iwanska, I. 1996. Natural language is representational language. In *Proc. of the Fall AAAI Symposium on Knowledge Representation Systems Based on Natural Language.*

Jackendoff, R. 1983. *Semantics and Cognition.* Cambridge, Mass.: MIT Press.

Jackendoff, R. 1990. *Semantic Structures.* Cambridge, Mass.: MIT Press.

Keil, F. 1989. *Concepts, Kinds, and Cognitive Development.* Cambridge, Mass.: MIT Press.

Lenat, D., and Guha, R. 1989. *Building Large Knowledge-Based Systems: representation and inference in the CYC project.* Reading, Mass.: Addison-Wesley Publishing Company.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

MacGregor, R. 1993. Representing reified relations in Loom. *Journal of Experimental and Theoretical Artificial Intelligence* 5:179–183.

Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, K. 1993. Introduction to WordNet: An on-line lexical database. Technical report, Princeton. CSL Report 43, revised March 1993.

Partee, B.H. with Hendriks, H. 1997. Handbook of logic and language. In van Benthem, J., and ter Meulen Alice., eds., *Handbook of Logic and Language.* Amsterdan and Cambridge: Elsevier and MIT Press.

Quine, V. 1953. *From a Logical Point of View.* Cambridge, Mass.: Harvard University Press.

Quine, V. 1960. *Word and Object.* Cambridge, Mass.: MIT Press.

Shapiro, S., and Rapaport, W. 1992. The SNePS family. *Computers and Mathematics with Applications* 23:243–275.

Sowa, J. 1984. *Conceptual Structures - Information Processing in Mind and Machine.* Reading, Mass.: Addison-Wesley.

van Benthem, J., and ter Meulen Alice. 1997. *Handbook of Logic and Language.* Amsterdan and Cambridge: Elsevier and MIT Press.

Whorf, B. 1956. *Language, thought, and reality: selected writings of Benjamin Lee Whorf.* Cambridge, Mass.: MIT Press. Carroll, J.B. (ed).

Wilks, Y. 1985. Reference and its role in computational models of mental representations. Technical report, CRL, New Mexico State University. MCCS-85-30.

Wittgenstein, L. 1958. *Philosophical Investigations.* Oxford: Blackwell.

Zadrozny, W. 1994. Reasoning with background knowledge - a three-level theory. *Computational Intelligence* 10(2):150–184.