Acquiring Intersentential Explanatory Connections in Expository Texts

July 1995

Accepted with revisions in *International Journal of Human-Computer Studies*

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

Abstract

A taxonomy [1] of explanatory links connecting sentences in expository texts is presented. It is also shown that there are two types of knowledge, which we have called analytical and empirical knowledge in analogy to the distinction between analytical and empirical sentences, that allow us to find and learn the explanatory connections. The role that the notion of analyticity plays in learning explanatory connections from expository texts is also emphasized. A program that embodies the ideas and that finds and learns explanatory connections in expository texts is also briefly explained.

[1]Sections 2 and 3 in this paper are modified and extended versions of a paper that appeared under the title "A Model of Comprehension of Elementary Scientific Texts," in the *Proceedings of the Workshop on Theoretical Approaches to Natural Language Understanding*, Halifax, Nova Scotia, 1985.

1 Introduction

This paper presents an analysis of intersentential explanatory connections in expository texts. It analyzes what kinds of explanatory links there are and the types of knowledge necessary to find and learn the explanatory connections. In descriptive texts such as those found in science books, encyclopedias, etc., some sentences may relate to previous sentences by elaborating or expanding some of the concepts already introduced; while other sentences may relate to previous ones by explaining some relations that were introduced in the latter. The following short discourse is an example of the first situation:

> *An alveolar air sac looks like a bunch of balloons. The walls of each sac are very thin and moist. Alveoli are surrounded by a rich supply of capillaries.*

The second sentence in this text is expanding or elaborating the concept *alveolar air sac* by saying its walls are thin and moist. The same thing can be said of the third sentence that is predicating a new conceptual relation of *alveoli.* However, consider the following passages:

1. Antibiotics work against infections. They kill the germs that cause them.
2. Lions are very powerful. Few animals dare to bother them.
3. Plants are dying. It is 20 degrees F outside.

In (1), the second sentence is an explanation of how antibiotics work against infections. In (2), the second sentence explains a consequence of lions being powerful. In (3), the second sentence is an explanation of why plants are dying. The understanding of these sentences requires the learning of the explanatory links connecting them. Without the acquisition of these links, one would not be able to answer the questions: *How do antibiotics work against infections? Why do few animals dare to bother lions?* or *Why are plants dying?* In these situations, subsequent sentences are an answer to a question suggested by previous sentences.

What exactly we mean by "a question suggested by previous sentences" is explained in the next section, where a taxonomy of explanatory links or connections between sentences is presented. In section 3, an analysis of the types of knowledge necessary to find out the explanatory links in expository texts is described. This analysis reveals that there are two distinct kinds of knowledge that permit the finding of explanatory links. We have termed these two types of knowledge analytical and empirical knowledge. This distinction is clearly based on the well-known philosophical distinction between analytical sentences and empirical sentences (Carnap, 1956). In section 4, the knowledge representation structures necessary to integrate the techniques presented in previous sections with an ongoing system for understanding expository texts and acquiring knowledge from them are presented. Sections 5 and 6 contain a brief discussion of the implementation; section 7 briefly explains the testing of the system; and sections 8 and 9 describe related research and our conclusions, respectively.

2 Explanatory Links between Sentences

In this section, we present some of the explanatory connections or links between sentences. There is no claim of completion for the taxonomy presented here; although we think that it may be about 80% complete. This taxonomy has been developed mostly from science books and encyclopedias. The explanatory links are organized in eight categories: completion of thematic roles (13 links), causal relation (2 links), effect or consequence (6 links), enablement (1 link), reaction of animate beings to other animate beings (1 link), properties (2 links), specification of fuzzy adverbs or adjectives (5 links), and explanation of the locations of things (1 link).

2.1 Completing Thematic Roles

In all these examples, the second sentence is filling one of the thematic roles left unspecified by the first sentence. The question being answered by the second sentence is "What is the *actor*, *theme*, etc. of the first sentence?" We have been able to find sensible examples for most thematic roles. It seems that the complement structure of the verb is realized in most sentences. As a consequence, temporal and locative adjuncts and also *instrument*, which is considered a quasi-argument by some Government and Binding researchers Pritchett (1992), are the most frequent examples of intersentential links. Because *actor* and *theme* are syntactically realized in most sentences by the subject and the direct object respectively, it is unlikely that one may find explanatory links filling these thematic roles. However, there are interesting exceptions such as those examples listed in (1). In the first sentence of (1a) it is unclear whether "Peter" or "Kelly" is the *actor* of *drive*. That ambiguity is clarified in the subsequent sentence. This subsection points out what could be a potentially interesting relation between the argument structure of the verb and explanatory connections.

1. Actor/Agent
   (a) Paul drove with Kelly to Tampa. Kelly drove her new car.
   (b) The house caught fire. It was struck by lightning.
2. Theme
   Bill donates to charity every year. He gives his used clothes.
3. Recipient/To-Poss
   Mary donates a lot of money every year. She gives it to conservation groups.
4. From-Poss
   Jennifer has borrowed a lot this year. The bank refuses to lend her more money.
5. At-Loc
   Peter married Mary August 20th. The ceremony took place in England.
6. From-Loc
   Peter arrived in Tampa at 5 p.m. He came from Orlando by train.
7. To-Loc
   Peter left Tampa at 3 p.m. by train. He arrived at Orlando at 6 p.m.

8. AT-TIME

   Peter married Mary in England. The ceremony took place August 20th.

9. START-TIME

   Peter went from Orlando to Tampa by train. He left Tampa at 5 p.m.

10. END-TIME

    Peter left for New York. He arrived there at 5 p.m.

11. DURATION

    Peter studied four articles on Kafka. He studied all night.

12. TIME FREQUENCY

    Kelly runs a lot. She runs 3 miles every day.

13. INSTRUMENT

    (a) Brutus killed Caesar. He stabbed him.
    (b) Antibiotics work against infections. They kill the germs that cause them.

2.2 A Causal Relation

This type of link expresses a causal relation between the actions and states expressed by two sentences. The causal relation may explain why actors perform certain actions, and why animate beings are in certain states.

14. WHAT CAUSES AN ACTOR TO DO SOMETHING?

    (a) In winter, birds fly south. They need food to eat.
    (b) Peter started drinking. He was depressed.
    (c) Mary treated Peter to ice cream. Recently he had been retired.

The second sentence in each example is an explanation of the action performed by the animate actor in the first sentence. In (14c), the knowledge structures that mediate between the two sentences are such that answering the question *Why did Mary treat Peter to ice cream* by saying *He had been retired* is begging the question. The work on narratives (see

Wilensky (1983), Norvig (1989)), on dialogues (see Allen and Perrault (1980), Litman and Allen (1987), Cohen and Lesveque (1985), Pollack (1986), Lambert and Carberry (1991)), and on natural language generation (see McKeown (1988), Paris (1988), Cawsey (1992)) has focussed on the knowledge structures needed to understand human plans and motivation as reflected in this type of sentence.

15. WHAT CAUSES AN ANIMATE BEING TO BE IN A STATE?
    (a) Fish are dying. The river is polluted.
    (b) Peter hated John. He stole his car.
    (c) The plants are dying. It is 20 degrees outside.

In (15b), the ambiguity of "he" and "his" occurs because one can understand the second sentence as the *cause* of Peter's state resulting in binding "he" to John and "his" to Peter, or as an *effect* of Peter's state resulting in binding "he" to Peter and "his" to John.

2.3 Effect or Consequence

These links are symmetrical to those explaining a causal relation. There are six connections. The first one explains the effect or consequence of being in a state. The other five explain the effect or consequence of actions on the *actor*, *theme*, *instrument*, *destination* and *recipient* of actions.

16. WHAT IS THE EFFECT OF SOMETHING BEING IN A STATE?
    (a) Peter was sick. He died.
    (b) Peter was depressed. He started drinking.
    (c) It is 22 degrees outside. The plants are dying.

17. WHAT IS THE EFFECT OF ACTIONS ON THE ACTORS?
    (a) Peter has read many books. He knows a lot.

(b) Peter ate poisoned food. He is sick.

18. What is the effect of actions on the themes?

    (a) Peter pushed John. He fell.

    (b) The bottle fell. It broke.

19. What is the effect of actions on the instruments?

    Peter hit the metal table with the cup. It broke.

20. What is the effect of actions on the destination/goal of those actions?

    (a) Put water in the bread. Microbes need moisture to live.

    (b) John loaded the little truck with salt. It corroded.

21. What is the effect of actions on the recipient of those actions?

    Mary's uncle gave her two million dollars. She is rich.

2.4 Enablement

Enablement is something that an animate being has, allowing or enabling it to perform certain actions, where "has" may stand for *animate-body-part*, *psychological-feature* or *possession*.

22. What allows an agent to perform certain actions?

    (a) Frogs have large hind legs. They allow them to jump up to 20 feet.

    (b) John has a lot of money. He can buy anything he likes.

    (c) Barbara is very intelligent. She got a perfect score on the test.

2.5 Reactions of Animate Beings to the Actions Performed by Other Animate Beings Towards Them

This link relates any two thematic roles that are realized by two animate beings.

23. HOW DO ANIMATE BEINGS REACT TO THE ACTIONS OF OTHER ANIMATE BEINGS TOWARDS THEM?
    (a) Kelly asked Jennifer to go to school with her. She refused.
    (b) Germs enter the human body. The body attacks them.

In (23a), the second sentence explains the reaction of the *actor*, "Kelly," to the *theme*, "Jennifer." Humans react to other humans in a variety of ways. Fortunately, the way in which lower animate beings react to other animates is not as diverse. In order to implement this type of link, we have classified these reactions as positive, negative or neutral reactions. A negative reaction takes place when animate beings attack or harm each other. For instance, the way in which termites react to intruders, or the the way the body reacts to germs. A positive reaction occurs when the action performed by one of the animate beings benefits the other. A neutral reaction happens when the action is neither positive or negative.

2.6 Links Based on Properties

Two links are based on the concept of *property*. These links answer the questions "What are the consequences of an entity having property B?" and "Why does an entity have property B?"

24. WHAT ARE THE CONSEQUENCES OF HAVING CERTAIN PROPERTIES?
    (a) Transistor radios are cheap. Many people can afford it.
    (b) Some caterpillars are poisonous. Most birds avoid them.
    (c) Lions are powerful. Few animals dare to bother them.

The chain of inferences is as follows: something that is cheap means that it costs little; and, if something costs little, many people can afford them. Likewise, if something is poisonous, it implies that it may kill animate beings; and if something may kill animate beings, it implies that animate beings avoid it. The meaning of "imply" in these sentences is explained in the next section.

25. WHY DOES CONCEPT A HAVE PROPERTY B?

   (a) Life is meaningless. People are exploiting each other all the time.
   (b) Germs are bad. They cause diseases.
   (c) Transistor radios are cheap. They are mass-produced.

The way in which the first sentence is connected to the second sentence in this type of link depends on the kind of property. For instance, if the property has a negative connotation, such as "bad" or "dangerous" or "meaningless", and that property is attributed to an animate being, then the second sentence is connected to the first one by expressing something that animate beings do or cause which also has a negative connotation (25a and 25b). In other cases, establishing the connection requires a more specialized knowledge about the things to which the property is attributed (25c).

2.7 Specifications of Fuzzy Adverbs and Adjectives

The specification of fuzzy adjectives and adverbs, e.g., tiny, almost, many, few, most, etc., affects mainly the following categories. Most of these fuzzy terms are also context sensitive. For instance, if a three year-old ate three apples, somebody could refer to this fact by saying that he/she ate many apples. However, if the human in question were a 200 pound weight-lifting adult, few people would use the noun phrase "many apples." This problem is not handled by the program described in later sections.

26. QUANTIFICATION

   (a) Few people conserve energy. Only one of every five practices some conservation.
   (b) Peter bought many books. He bought 23 novels and 6 textbooks.

   The second sentence explains exactly the value of the fuzzy quantifier in the first sentence.

27. DIMENSION

   (a) Plankton are tiny animals. Most plankton are one-thousandth of an inch in length.

(b) Peter is very tall. He is 6.2 feet tall.

The second sentence specifies unambiguously the dimension of "tiny," "tall," etc.

28. Degree

The water is very hot. It is 200 C.

29. Classification

(a) Almost all birds lay their eggs in nests. Penguins are the exception.

(b) Not all creatures that live in the sea are fish. Whales and dolphins are mammals.

The second sentence completes the classification of some concepts left underspecified in the first sentence. In (29a), the first sentence mentions, without providing a name, some classes of birds that do not lay their eggs in nests. The second sentence provides the name of a such class of birds. Likewise, in the second example, the second sentence is explaining "not all" by providing examples of classes of sea animals that are not fish. These examples can also be considered as an example of the discourse relation *contrast* in rhetorical structure theory (RST) (Mann and Thompson (1987)), that is discussed in the section on related research.

These links specify fuzzy adverbs modifying actions.

30. Modifiers of Action

(a) Tracy almost got elected to the senate. She lost by only 100 votes.

(b) Mary ate her breakfast very fast. She took two minutes.

2.8 Explanations of the Locations of Things

These links explain the locations of certain things. Specifically, they explain the *at-loc* role in the logical forms of sentences whose main verb is a form of "be." Example (31b) illustrates that what qualifies as an explanation changes with history, and that some knowledge which is considered today to be common-sense knowledge was once scientific knowledge.

31. Why do certain things happen to be at certain places?

    (a) There is a bolt in the cereal. The packer must have been fixing his/her car.
    (b) There are worms in the meat. They grew spontaneously.

3 Acquiring the Connections: Analyticity, Empiricity and Learning

This section introduces the notion of analytic and empirical knowledge based on the philosophical distinction between analytical and empirical sentences. There are two types of knowledge that allow the system to acquire the links connecting sentences, which we have called analytical and empirical knowledge. These types of knowledge are analogous to the notions of empirical and analytical sentences. An analytical sentence is one whose truth value can be established by only having knowledge about the language. An analytically true sentence, also called a tautology, is one that holds true in every possible world, an idea that goes back to Leibnitz. Classical examples of analytical sentences are *All bachelors are unmarried* and *All white horses are horses.* In contrast, the determination of the truth values of empirical sentences requires knowledge about the world. For instance, the sentence *Plants need light to live* requires knowledge about plants and light to establish its truth. We have used two types of rules, which we have called analytical and empirical rules, to encapsulate analytical knowledge and empirical knowledge. In his classical paper, *Two dogmas of the empiricism*, Quine (1953) argued that there is not a dichotomous distinction between analytical and empirical sentences, but that the distinction is only one of degrees. Piaget (1959) also showed that people's intuitions of analyticity vary. Only in very extreme cases, such as *All white dogs are white*, did the subjects studied by Piaget agree about the analyticity of the sentences. Although we agree with Quine that there is not a dichotomous distinction between analytical and empirical sentences, we do think that the distinction is a relevant one

(Putnam (1962)). This relevance becomes more manifest when one considers the distinction from the point of view of learning or knowledge acquisition. If one is told that all white dogs are white, and one is asked "What did you learn from that statement"? The answer is nothing because the predicate "white" is already contained in the subject "white dog." This is the explanation given by Kant in his *Kritik der reimen Vermunft.* However, if one is told that all American wild horses came from Europe, the situation is radically different because the predicate "came from Europe" cannot be derived from "American wild horse." Hence, it seems that null learning occurs in extremely analytical sentences and that learning is at its fullest in extremely empirical sentences, where "extremely analytical sentences" is a behaviorist concept meaning that at least 90% of people agree on their analyticity. However, we will show that analyticity does play an important role in learning, and that the analytical and empirical distinction provides an excellent means for organizing the knowledge necessary to acquire new knowledge from texts. Hence, we take the notion of analyticity in a Quinean sense, namely, as those sentences that require no knowledge or very little knowledge of the world to establish their truth value. This position, we think, is an agreement with Quine's view, whose major position is that there is not a distinction in principle between analytical and empirical sentences. Moreover, he was not interested in learning, but in rejecting the notion of synonymy that could be reconstructed if a dichotomous distinction between analytical and empirical sentences is held (see Carnap (1956), Katz (1990)).

3.1 Analytical Rules

However, instead of using the term "analyticity" as Quine does, we use the term "analytical knowledge," because our goal is to investigate which aspects of learning depend on analytical knowledge and which aspects depend on empirical knowledge. Thus by "analytical knowledge" we mean that knowledge that derives from understanding the concepts underlying language and its interrelations. In our system, analytical rules are anchored in each verbal concept, or predicate. Some examples of analytical rules are:

```
R1 if ingest(?x,?y) then enter(?y,?x)
R2 if ingest(?x,?y) and at-loc(?z,?y) then ingest(?x,?z)
```

The negation of rule R1 will result in a contradiction, one in which the meaning of "ingest" will be changed. Regarding R2, it is hard to imagine a world in which it does not hold. Let us consider the relation between analyticity and learning. Every time that an analytical rule fires, connecting two sentences, a new piece of knowledge is learned. Consider the passage:

Antibiotics work against infections. They kill the germs that cause infections.

Even if this passage is read with no knowledge of "antibiotics," "infections" and "germs," one still can learn why antibiotics work against infections. The learning of this fact is based on an analytical rule that says:

```
if destroy(?x,?y) and cause(?y,?z) then work-against(?x,?z)
```

(where ?x is instantiated to antibiotics, ?y to germs, and ?z to infections). Note that the antecedent or premise of the rule is the explanation of how antibiotics work against infections. Thus, the analytical rules are not only making understanding possible but also learning. In the above example, a new empirical fact has been acquired, namely, that antibiotics are effective against infections because they destroy the germs that cause them. The analytical rule used is of a general nature and, consequently, can be applied in many other different situations.

The structure and content of the analytical rules depend not only on the ontology of action verbs and descriptive verbs (contain, include, consist of) but also on the ontology of properties and entities. (Descriptive verbs are those in which the meaning of the verb is neither an action nor an experience. Some examples are "consist of," "contain," "include," (in one sense), "border," "intersect," (in one sense), "enclose," "meet," (in one sense), etc.) Consider the three examples below, followed by their respective rules:

Cars are bad for people. They produce pollution.

```
if produce(?x,?y) and bad-for(?y,?z) then bad-for(?x,?z)
```

Some germs are bad for humans. They cause infections in humans.

```
if cause(?x,?y,?z) and bad-for(?y,?x) then bad-for(?x,?z)
```

Some germs are good for humans. They attack other germs that cause infections in humans.

```
if attack(?x,?y) and cause(?y,?z,?w) and
                     bad-for(?z,?w) then good-for(?x,?w)
```

In the above rules, the predicates and actions subsume many other predicates and actions belonging to the same semantic category. For instance, the predicate *good-for*, subsumes the following predicates and actions: "beneficial," "positive," "benefit," "help," etc.

3.2 Empirical Rules

Now, we turn our attention to the empirical rules. Consider the passage:

> *In 1988 there was a hard freeze in Florida. Most orange groves in Central Florida perished.*

In order to establish the connection between the two sentences, one needs to have knowledge about the world. One still may know what "freeze" is, but that may not be sufficient to establish the connection unless one knows that freezing temperatures kill certain plants, including orange trees. In contrast to this, in order to establish an analytical connection between two sentences, knowledge about the language is sufficient. If one cannot establish a connection between the sentences *Peter went to Tampa* and *He arrived there at 5 p.m.*, it implies that one does not know what "went" and "arrive" mean. But, one may know exactly what freezing means and still be unable to establish a connection between the dying of orange trees and freezing temperatures. Cue phrases, e.g., "because," "as result," etc., help to establish this connection. However, in many cases in which the connection can be established by using common sense knowledge, the authors do not use cue words. For instance, consider the following two discourse taken from the *World Book Encyclopedia* (World Book, Inc., Chicago, 1987.)

> *People are probably the worst enemies of the house mouse. They set traps and place poisons where mice can easily find them. Bats of the tropics do not hibernate. The temperature and food supply there remain suitable all year.*

In other cases, the cue word does not reveal its deep meaning, like the word "where," meaning also "because" in the passage:

> *Most kinds make their home in the tropics co where they can find food the year around.*

In still other cases, the link is not between two sentences but between a matrix clause and its subordinate clause, which can stand for a temporal, instrumental, causal link (see section 2.2), etc., like the subordinate clause in

> *Bats perform a valuable service for people by eating large numbers of insects.*

Compare it to

> *Bats avoid many enemies on the ground by hanging from high places.*

All these examples are taken from the article on bats in the aforementioned encyclopedia. Empirical knowledge ranges from very general laws of nature to episodic knowledge. Consider the sentences:

1. The moon orbits the earth. The mass of the earth is greater than that of the moon.
2. The microbes in the bread culture died. Peter had not put water on the culture for days.
3. Plants are dying. It has not rained in 20 days.
4. During the San Francisco earthquake most houses burned. They were made of wood.
5. Peter was depressed. He failed the test.
6. Peter was depressed. He grabbed a knife.

One can see that the connection between the sentences in the first four examples can be more naturally grasped in the form of a general rule expressing some facts about the world. However, when one considers the last two examples it seems that rules are not adequate enough to grasp the connection between the sentences. The reason for this is that knowledge establishing the connection is of an episodic nature, having to do with our past personal experiences (see Schank (1982) and Kolodner (1983)). Moreover, for most people, the knowledge allowing them to establish the connection in the examples (3), (4), (5) and (6) is of episodic nature. We all have experienced or seen the dying of plants because of droughts, the burning of wood, etc. It seems that these facts are represented/recorded in our episodic memory as concrete experiences, not as general facts expressed in the form of rules. Thus, in many situations, one does not use general empirical rules to learn the connections between sentences, but concrete episodic facts. However, empirical knowledge in the form of rules does also play a role in learning the connections between sentences, especially in expository texts. Suppose that a program, or a human for that matter, learns by being told that *All living beings need oxygen to live.* This general fact can be expressed in the following rule:

```
if animate(?x) and at-loc(?x,?z) and lack(?z,oxygen) then die(?x)
```

Now, let us consider the passage:

The fish in the lake are dying. The algae in the lake have consumed the oxygen in it.

Even if a human has not experienced this situation before, she/he will be able to draw the connection between the two sentences. Besides the empirical rule above, the human and the computer will need the following analytical rule for *consume*:

```
if consume(?x,?y) and at-loc(?y,?z) then lack(?z,?y)
```

As another example of the need for general empirical rules, consider the statement *People eat what they like.* A statement like this needs a rule with a variable to represent "what." In all these examples, the empirical knowledge that connects the sentences can be expressed in a few rules. However, in other cases an entire scientific theory may mediate between the two sentences, as in the examples below:

There are worms in the meat. They grew spontaneously.
Cars are cheap. They are mass produced.

The first example invokes the theory of spontaneous generation to explain the appearance of worms in the meat. In the second example, a theory of mass production is needed for a thorough explanation of the link connecting the sentences. These examples are not covered by our present implementation.

3.3 Summary

In conclusion, there are two types of knowledge connecting intersentential explanatory links: analytical knowledge and empirical knowledge. Analytical knowledge is very close to logic inferences and lends itself to being grasped naturally in the form of rules. Empirical knowledge manifests itself in two dimensions: as general empirical knowledge expressed in the form of rules, and as episodic knowledge. Our present implementation includes a small set of empirical rules dealing with plants and animate beings in general. These rules are stored in the concepts in the hierarchy of concepts so that they can be inherited by child concepts. In the remainder of the paper, we explain how these ideas are integrated in a model, embodied in a program called SNOWY, for understanding and acquiring knowledge from expository texts (see Gomez and Segami (1989)). SNOWY is a model intended to fully automate the acquisition of knowledge from texts, in contrast to semi-automated knowledge acquisition systems (see Szpakowicz (1990), Moulin and Rousseau (1994), Delisle (1993)).

4 Brief Overview of the Representation of Concepts

This section describes very briefly the knowledge representation language used by our model, so that the reader may follow the discussion of the examples in the next sections. A detailed discussion and a comparison to KL-ONE (Brachman and Schmolze (1985)) may be found in Gomez and Segami (1991) and Gomez and Segami (1989). Concepts are represented in LTM (long-term memory) in frame-type structures, which are linked through *is-a* and *classes-of* slots to yield a hierarchical organization. Two kinds of representation structures may be part of LTM: *object-structures* and *relation-structures.* Two kinds of relations may be distinguished: those that describe objects and those that attribute actions to physical things. Thus, *Centipedes are arthropods* and *All centipedes have many legs* express descriptive relations, while *Antibiotics kill germs* and *Cells take in sugar* express actions. The sentence *All centipedes have many legs* is represented by creating two *object-structures* and one *relation-structure* as indicated below:

```
centipede (has-part (leg ($more (@a8845)))
leg (part-of (centipede ($more (@a8845)))
```

Every argument in a relation is represented in an *object-structure* and indexed by creating the inverse relation pointing to the representation of the relation. The relation in the *object-*

structure contains only one argument (no argument if the relation is intransitive), followed by a pointer to the representation of the relation itself, which is represented separately by means of a *relation-structure.* The relation *has-part* is represented as:

```
@a8845
  (instance-of (description))
  (args (centipede) (leg))
  (pr (has-part))
  (descr-subj (centipede (q (all))))
  (descr-obj (leg (q (many))))
```

The *instance-of* slot indicates the type of relation. The slots *args* and *pr* identify the arguments of the relation and the name of the relation, respectively. The slot *q*, which stands for quantifier, indicates the quantifier for the argument. The scope of quantifiers is from left to right. For instance, the FOPC formula for the relation *@a8845* is:

$\forall(x)(Centipede(x) \Longrightarrow \exists(y)Leg(y)\wedge \textit{Has-Part}(x,y))$

Note that we have represented the quantifier "many" as an existential quantifier. This is clearly a problem with FOPC, because quantifiers such as "few," "most," "many," etc. play an important role in inductive algorithms. The content of the slot *q* may also be a question mark, meaning that the value of the quantifier is unknown. That will be the case if the system reads the sentence *Antibiotics kill germs*, in which the *q* slots for the *actor* and *theme* contain a question mark, because it is unknown if all or some antibiotics kill all or some germs. Our language is not restricted to binary relations like the one illustrated and can represent n-ary relations. Some of the arguments of the relations may be relations themselves, and relations may be connected to other relations, as in the sentence *The human body attacks germs by producing antibodies.* See Gomez and Segami (1991) for a detailed discussion of these representational issues.

Another aspect of the knowledge representation language needed in order to follow the subsequent discussion is the representation of concepts denoted by restrictive relative clauses and, in general, any restrictive qualifier such as "mammals that live in the sea," "birds with long legs," "sea mammal," etc. These concepts are called by Fodor (1981) phrasal concepts and distinguished from lexical concepts, e.g., "whale," "mammal," etc. Consider the passage

> *Mammals that live in the Pacific are endangered. They have been and are still hunted by humans to the verge of extermination.*

This text introduces the concept "mammals that live in the Pacific," for which a name is not given. However, a unique name in memory needs to be created for this concept so that additional knowledge about it can be learned by storing this knowledge under the same node, and so that this node can be linked to the other nodes in the hierarchy. Restrictive qualifiers are represented in LTM via *object-structures* with a dummy name (a gensym), and their representation is characterized by the presence of a slot, called *cf*, whose purpose is to identify the concept by providing necessary and sufficient conditions that define it. Thus, the concept "mammals that live in the Pacific" is represented as:

```
x1
cf (is-a (mammal)) (inhabit (pacific (q (constant))))
```

The concept so created is not identified by its name, *x1*, which is arbitrary (in the program, it is a gensym), but by the content of its *cf* slot. The value "constant" of the quantifier slot means that "Pacific" is a constant. The meaning of this in FOPC is:

$\forall(x)(x1(x) \Longleftrightarrow Mammal(x) \wedge Inhabit(x, pacific))$

Subsequent knowledge about "mammals that live in the Pacific," for instance, that they are endangered, etc., will be stored under x1.

5 Algorithm for firing the rules

Because the system in which these ideas are integrated is reading novel texts and assimilating new concepts and relations, it embodies a knowledge-lean model of comprehension, that is, one which does not involve schemas or frames to be instantiated as sentences are read (see also Kieras (1985), Mayer (1985), Van Dijk and Kintsch (1983)). This differentiates it from knowledge-intensive models of comprehension based on scripts, plans or MOPs (Wilensky (1983) and Dyer (1983)). There are several components in this model.

1. Parsing. The parser uses a lexically-driven syntax-based algorithm based on the notion of the syntactic usage of a word (Gomez (1989)). The parser builds the following syntactic relations: subject, object1, object2, predicate, and prepositional phrases (PPs). Object1 is built for the first post-verbal NP of transitive verbs, and object2 for the second postverbal NP of diatransitive verbs. The parser also recognizes temporal adjuncts of the verb. But, it does not resolve structural ambiguity. PPs are left unattached in the structure built by the parser until the semantic interpreter finds their meaning and attaches them.

2. Semantic Interpretation. If rules for recognizing the *meaning of the verb* are defined, the parser interacts with a semantic interpretation algorithm, which attaches PPs and other modifiers and determines the thematic roles. The parser continues parsing in the same manner as prior to the recognition of the meaning of the verb, but the syntactic relations are analyzed by the interpretation algorithm (Gomez, Segami and Hull (1994)). The output of the semantic interpreter is a logical form, which becomes the input for the *formation* phase.

3. Formation Phase. The output of the *formation* phase is the final knowledge representation structures. Some of those were briefly explained in the previous section.

4. Recognition. The knowledge representation structures built by the *formation* phase are passed to the *recognition* phase, which checks whether the concepts and conceptual relations that have been formed are already in LTM (long-term memory). The *recognition* algorithm makes extensive use of a *classifier* (Brachman and Schmolze (1985)).

5. Integration. Those concepts that the *recognition* phase fails to recognize are passed to the *integration* phase, which integrates them in LTM upon activating the *classifier* algorithm that indicates where and how they must be integrated in LTM (see Gomez and Segami (1989)). As a result of integrating new concepts in LTM, the classifier may reorganize some of the concepts already in LTM.

6. Explanatory Component. Let L1 be the logical form for a clause, and R1 be the final knowledge representation for that logical form. The list **links** is a stack containing the list of links that have been generated for the last three previous sentences and can be viewed as a short-term memory (STM). Once **links** contains three elements, three sublists containing the links for the last three sentences, the pushing of a new element causes the forgetting of the oldest element in **links** by popping it from **links**.

   (a) Generate links for R1 by examining the representation of the verbal concept and the ontology of adjectives and adverbs, where is indicated which links to generate for each conceptual relation. These links are stored in a temporary list of links, called **current-sentence-links**. If the list **links** is empty, skip the next step.

   (b) Fire the rules. (b.1) Retrieve the analytical rules stored under the verbal concept of L1. If there is no analytical rule stored under that verbal concept, the superconcepts of that verbal concept are searched, and any analytical rule found is retrieved. (b.2) Retrieve any empirical rules stored in any of the concepts of R1 (the arguments of the relation), or in any of their superconcepts. (b.3) Fire the rules in a backward fashion; that is, compare the then-part of the rule with the list of links. If any one matches, try to prove the antecedent of the rule by activating the question-answering component, asking it to prove the antecedent. (b.4) if the question-answering component verifies the antecedent of the rule, then the consequent (the explanatory link) is used to fill the appropriate slot that is now added to the relation structure in LTM. (b.5) Remove any link that is explained by either an analytical or empirical rule from **links**.

(c) Push **current-sentence-links** into **links**. Pop the oldest element in **links** if **links** contains four sublists. Process another clause or sentence.

We illustrate these steps by following through two examples. Suppose that the system reads the sentence *The human body recognizes the germs that enter it.* Once the sentence is parsed and interpreted, the *formation* phase builds the *object-structures* and *relation-structures* depicted in Table 1. Two *object-structures* are created for this sentence: *human-body* and @x00008, and two *relation-structures*: @a00009 and @a00010, corresponding to the actions *enter* and *become-aware* respectively. The *object-structure* named @x00008 represents the concept *germs that enter humans.* Note that its *cf* slot defines it as a subconcept of *germ*, characterized by entering humans. Besides the necessary and sufficient conditions of @x00008, the system knows that @x00008 is recognized by (become-awarehuman body, and that the specifics of that action are stored in the *relation-structure* @a00010.

After completing the phases of *parsing, formation and recognition*, the *integration* phase inserts the newly-built *object-structures* and *relation-structures* into LTM. The integration of *object-structures* involves their classification among the concepts in LTM, and, in some cases, may trigger a reclassification of the concepts currently present into LTM. When a *relation-structure* is integrated in LTM, the *integration* phase examines the representation of the verbal concept, adjectives and adverbs in this structure, in order to determine which links need to be generated for the sentence just read. The explanatory links generated for the sentence of our example are:

```
((instrument ?w ( enter (@x00008 human-body) @a00009)))
(from-loc ?w (enter (@x00008 human-body) @a00009))
(instrument ?w (become-aware (human-body @x00008) @a00010))
```

Two links were generated for the clause *Germs enter the human body* and one link for *The human body recognizes @x00008.* As has been explained, the symbol *@x00008* is the name of the concept *germs that enter humans.* The symbols *@a00009* and *@a00010* are the names of the *relation-structures* for which the corresponding link was generated.

These links can be paraphrased as: *How do germs that enter the human body enter it? From where do germs that enter the human body come?*, and *How does the human body recognize the germs that enter the human body?*

Once the *explanatory* phase is activated, its first task is to determine if the sentence just read answers any of the links generated by previous sentences. But because only one sentence has been read, there are no previous links to be explained. Thus, the list of links generated is pushed onto **links**.

This completes the processing of the first sentence. Let us now suppose that the next sentence read is *These germs are located in the food ingested by the human body.* After inserting the new concepts and relations in LTM, the *integration* phase generates the link shown below, and activates the *explanatory* phase. This sentence will also generate links which are not listed because they are not used in this example. The *explanatory* phase now examines the representation structures built by the *formation* phase for the current sentence. In our example, the structures to be examined are depicted in Table 2.

For each relation introduced by the sentence, the analytical rules attached to its verbal concept are retrieved. In this case, the first relation examined is the action *@a00017*, so the rules attached to *ingest* are retrieved. One of these rules is the following:

```
(if ((at-loc (?x ?y)) and (ingest (?z ?y))) then
                         (instrument ?y (enter (?x ?z))))
```

Once the analytical rules are retrieved, each link in **links** is matched against the *then-part* of the rules. At this point, after reading the two sentences above, the list **links** contains:

```
((instrument ?w ( enter (@x00008 human-body) @a00009))
(from-loc ?w (enter (@x00008 human-body) @a00009))
(instrument ?w (become-aware (human-body @x00008) @a00010)))
```

As we can see, the first link in **links** matches the *then-part* of the above rule, by instantiating the variable ?x to *@x00008* and the variable ?z to *human-body.* Because there is a match, the *explanatory* phase tries to verify the *if-part* of the rule which, after instantiating the variables ?x and ?z, has the form:

```
(if ((at-loc (@x00008 ?y)) (ingest (human-body ?y))) then
    (instrument ?y (enter (@x00008 human-body))))
```

To verify the two conditions of the rule, the *explanatory* phase activates the question-answering module. (See Gomez and Segami (1991) for a detailed discussion of the question-answering module.) Each condition is passed to the question-answering module in the form of a question. Thus, the first condition becomes the question *Where are germs that enter the human body located?.* Since the knowledge the system has about germs is what has been specified in our two sentences, the answer to this question will be *Germs that enter the human body are located in the food ingested by the human body* thus instantiating the variable ?y to @x00016, which stands for "food ingested by the human body." The second condition in the *if* part becomes *(ingest (human-body @x00016)),* which is passed to the question-answering module as the question *Does the human body ingest @x00016?.* Since the answer to this question is affirmative, the rule has fired successfully, providing an answer to the explanatory link *(instrument ?w (enter (@x00008 human-body) @a00009)).* The *explanatory* phase then proceeds to fill the *instrument* slot in the *relation-structure* @a00009 with the concept @x00016 and to add @x00016 as a third argument to the list of arguments of @a00009. Because the concept @x00016 has a question mark as the value for the quantifier slot in the relation @00018, it will have also a question mark for the quantifier in the instrument slot of the relation @a00009. Finally, the *explanatory* phase removes the link *(instrument ?w enter (@x00008 human-body) @a00009)* from **links**. The rest of the links are examined in a similar way.

6 Inferring Knowledge as Needed

In all our examples, the rules have been used to acquire new connections between facts stored in LTM. But, there may be situations in which one may ask the system for a piece of knowledge that the system has not inferred. Hence, a search of LTM looking for that piece of knowledge will fail. Consider the sentences:

Antibiotics work against infections. They kill the germs that cause infections.

When the first sentence is read by the system, a *relation-structure*, say *@a1*, will be built to represent the underlying relation, and one of the explanatory links generated will be *(instrument ?y (work-against (antibiotic infection))).* This explanatory link will be answered by the second sentence because one of the analytical rules attached to the concept *destroy* is:

```
R3 (if ( destroy(?x ?y)  cause(?y ?z)) then
       (instrument destroy(?x,?y) (work-against(?x ?z)))
```

which will fire successfully. Thus, the *instrument* slot of the *relation-structure @a1* will be filled with the relation *destroy (antibiotic x1)*, where x1 is the concept *germs that cause* *infections.* If the question *How do antibiotics work against infections?* is now asked, a simple access to the structure *@a1* in LTM will provide the answer.

On the other hand, suppose that the system first reads the sentence:

Antibiotics kill the germs that cause infections.

As before, once the knowledge in this sentence is integrated in LTM, the *explanatory* phase retrieves the analytical rules attached to the verbal concepts in the sentence, and fires them in an attempt to answer previously generated links. One of the rules retrieved is rule R3, above. However, contrary to our previous example, after reading the sentence *Antibiotics kill the germs that cause infections* the *explanatory* phase will not add any new knowledge to LTM. The reason is that, since no sentences have been read before, there are no explanatory links to be answered. Therefore, questions like *How do antibiotics work against infections?* or *Do antibiotics work against infections?* cannot be answered by the system. An obvious solution to this problem would be to fire all empirical and analytical rules in a forward fashion each time that a new sentence is integrated, and to add to LTM each piece of knowledge obtained. The major problem with this is that many irrelevant pieces of knowledge could be added to LTM. The solution we have given to this problem is the following. If the question-answering module does not find in memory the answer to a given question, it activates the *explanatory* phase to try to infer the answer from memory. The *explanatory* phase selects the analytical rules attached to the verbal concept of the question, and also the empirical rules, if any, attached to the main concepts in the question. This is what will happen if the question *Do antibiotics work against infections?* is asked. The *explanatory* phase gets the analytical rules attached to *work-against* and selects those whose *then-part* have *work-against* in them because the question has the verbal concept *work-against.* Then, it tries to verify the truth of the *if-part* of those rules. In our example, since rule R3 above is attached to *work-against*, as well as to *destroy*, the *explanatory* phase will instantiate this rule as:

```
(if destroy(antibiotics,?y) and cause(?y,infections) then
(instrument destroy(antibiotics, ?y)  work-against(antibiotics,infections)
```

In order to verify the truth of the *if-part* of the rule, the *explanatory* phase, in turn, activates the question-answering module. The *explanatory* phase formulates the question *What do antibiotics destroy?* The question-answering will return a list containing the things killed by antibiotics. The *explanatory* phase instantiates the variable *?y* to each of the elements in the list and, again, asks the question-answering if any of those things cause infections. If the antecedent of the rule is found to be true, the answer to the question *Do antibiotics work against infections?* will be *yes.* The method we have explained allows us to obtain an analytical implication only when it is required to answer a question.

A similar method is used for empirical rules. For instance, suppose that the system has the following rule stored under the concept *human*, expressing that people eat what they like:

```
(if ( (like(human,?x)) and (is-a (?x,food)) ) then (ingest(human,?x)))
```

Let us assume that the system is told that *All Americans are humans* and that *All Americans like chocolate*, and then, is asked *Do Americans eat chocolate?*. In this case, there is no rule under the verbal concept *ingest* applicable. The system will search the concept *American* and its superconcepts, looking for an empirical rule whose consequent says *ingest(?x,?y).* That rule will be found in the concept *human*, and the inference techniques explained above are used to verify the rule.

7 Testing

Recently, we tested the system on the acquisition of knowledge from encyclopedic texts. The goal was to acquire knowledge about the diet and habitat of animals by reading unedited articles from the *World Book Encyclopedia.* The description of this aspect and an evaluation of the different components of the system can be found in Gomez *et al.* (1994). A skimmer was used to select sentences relevant to the topic of the diet of animals. Those sentences selected by the skimmer were then passed to the parser and the semantic interpreter, and those clauses for which a logical form could be built were passed to the *formation*, *recognition* and *integration* phases. The system was initially "trained" on ten articles. That is, we built semantic interpretation rules and verbal concepts for those sentences that were relevant to the diet of animals. After some initial testing on an encyclopedic article about birds, the system was tested on three randomly chosen articles, which happened to be articles on bats, monkeys and mice. The system, running with a lexicon of 35,000 words, was able to parse 93%, 73% and 79% of the sentences selected by the skimmer on the bats, monkeys and mice articles, respectively, and it was able to produce a semantic interpretation for 60%, 75%, and 50% of those sentences, respectively. The interpreter was able to produce logical forms for some of the sentences wrongly selected by the skimmer. The knowledge in these sentences was also integrated into LTM. The system acquired 118, 72 and 123 concepts and conceptual relations for each one of the articles on bats, monkeys and mice, respectively. We have also conducted tests on the habitats of animals yielding results comparable to those reported above.

The *explanatory* phase is an important component of this application because it became necessary to learn that, for example, ants are part of the diet of bears from the sentence *A bear uses its claws to dig up roots, ants, and termites*, or that Bogong moths of Australia inhabit lowlands and mountains from the sentence *Bogong moths of Australia migrate from lowland pastures to the mountains during the summer to avoid extreme heat.* Countless other examples could be used to illustrate these inferences.

There were two changes made to the *explanatory* phase for this application. One change was that some of the empirical rules were stored under the verbal concept, rather than under the hierarchy of entities. The algorithm has faster access to the rules when stored under the verbal concepts than when stored under the entity hierarchy, in which it needs to traverse a long path before finding the relevant rules. Another change was that those rules inferring facts relevant to the topics (habitat and diet of animals) were fired in a forward fashion in order to make readily available to the user the relevant knowledge she/he is searching for. These two changes speeded up our processing time considerably. The average time to completely process a sentence on a SPARC Classic Machine (50 MHZ) was 3.1 seconds. This includes the time for file handling, skimming, parsing, semantic interpretation, formation, recognition and integration. The implementation of the *explanatory* phase may still undergo some changes until it sets up into a final state.

8 Related Research

During the same time in which Gomez (1985) formulated his explanatory relations, Mann and his associates (see Mann and Thompson (1987), Mann, Matthiessen, Thompson (1989)) developed a set of text relations under the name of Rhetorical Structure Theory (RST). Unfortunately, this work has been unknown to us until the revision of this paper for publication. In Mann and Thompson (1987), twenty three relations are classified into two classes: *subject matter* that includes: elaboration, circumstance, solutionhood, volitional cause, volitional result, non-volitional cause, non-volitional result, purpose, condition, otherwise, interpretation, evaluation, restatement, summary, sequence, contrast; and *presentational* that includes: motivation, antithesis, background, enablement, evidence, justify, and concession. As in this paper, RST's authors do not claim that these are all the relations. These relations are defined by providing a *nucleus*, the central information, and a *satellite* (the supporting information). For instance, the relation *background* is defined as follows: the reader will not comprehend the nucleus sufficiently before reading the text of satellite. Although these definitions have been criticized for a lack of precision in Sanders, Spooren and Noordman (1990), they have been very influential in natural language generation (see Bateman, Kasper, Moore and Whitney (1990), Bateman (1993), Hovy (1993)).

A major difference between our relations and those of RST is that most RST relations relate greater chunks of text than ours, and that they give greater importance to the presentation of the text and to the writer's intentions and goals. This latter aspect together with attentional aspects plays an even more important role in theory of discourse proposed in Grosz and Sidner (1986) (see also Allen and Perrault (1980), Litman and Allen (1987), Cohen and Lesveque (1985), Pollack (1986), Lambert and Carberry (1991)). In contrast, the relations presented here deal with the factual knowledge mediating between sentences without making any reference to presentational, intentional, or attentional aspects. This, by no means, denies the importance of these aspects that clearly do play a very important role in discourse, as has been pointed out in the philosophical literature for some time now (see Wittgenstein (1953), Grice (1957), Austin (1962) and Searle (1969)). Another major difference between RST and the relations presented here is that RST does not provide any computational model for recognizing these relations and acquiring the knowledge mediating between semantic relations, although there has been an attempt to operationalize RST relations by formulating them as plan operators (see Hovy (1993) for a discussion). Of the relations listed in Mann and Thompson, only two of the subrelations of the *elaboration* relation, namely *set:member* and *abstract:instance*, seem to be identical to our *classification* relation.

Recently, Hovy (1993) and Maier and Hovy (1993) have presented a taxonomy of discourse relations gathered from different researchers. The three upper relations of the taxonomy are *semantic*, or *ideational*, *interpersonal* and *textual*, corresponding to Hallidays' three metafunctions of language (Halliday, 1985). This taxonomy is intended by the authors to be a basis for further analysis and extensions, rather than a final product. Within this taxonomy, the relations explained in this work seem to fall within the *elaboration* relation, which is a child of the relation *semantic*.

There are striking differences between the type of learning described in this paper and explanation-based learning of Mitchell, Keller and Kedar-Cabelli (1986) and DeJong and

Mooney (1986). First of all, the algorithm explained in this paper does not learn concepts, but relations or connections between concepts. Secondly, and more importantly, the relations to be learned are not given to our system through the specification of necessary and sufficient conditions. The requirement that concepts to be learned be specified by defining them using necessary and sufficient conditions, constitutes, in our opinion, one of the most serious limitations in the current learning research. Our approach explores the learning of the connections linking the concepts in a web with no specific layers (Quine (1960)).

In contrast to EBL, there is no training example in our method. Relations between concepts are learned as they are given in the text. Of course, the performance of our algorithm degrades if many sentences mediate between the introduction of a conceptualization and its explanation, something that occurs in poorly written texts in which explanations are zigzagging, causing them to be difficult to understand even for good readers.

9 Conclusions

We have shown how a program can find and learn the explanatory connections existing between sentences. Analytical rules are general in nature and are applicable to diverse domains. In fact, they are an essential component of the understanding of any domain. Empirical rules are domain dependent and do not play as essential a role as analytical rules. That analytical rules, which are devoid of empirical content, make possible the acquisition of new knowledge may seem questionable. Yet we have discovered that, in reading explanatory expository texts, a layman without *a priori* knowledge of the subject matter is left only with the analytical rules for establishing connections, and consequently, for understanding and learning.

In *Word and Object*, Quine (1960) conceived of language as a fabric of interconnected sentences in which the analytical sentences are at the center of the fabric, and the empirical sentences are situated at the periphery. We can draw an analogy between this and Quine's view of language as a fabric of interconnected empirical and analytical sentences and a subject's knowledge. In the absence of knowledge about the subject matter, a reader relies upon his/her analytical knowledge to acquire new pieces of empirical knowledge. The knowledge situated in the periphery, in contrast to the analytical (central) knowledge, changes and undergoes readjustments. This is clearly indicated by noticing that when we read an article about antibiotics, say, our knowledge about the subject is increased and modified, while our analytical knowledge remains unchanged. But it is interesting to note that it is the analytical knowledge which, to a high degree, is making possible the acquisition of world knowledge.

Whether the distinction between analytical and empirical statements has any cognitive relevance has been the object of much debate in the philosophical and psychological literature. The study of Piaget (1959) and his co-workers indicates that although human subjects do not "feel" a strong demarcation between these two types of sentences, they do sense that a relevant distinction is present in most cases. The preliminary feedback we have obtained from our program indicates that analytical and empirical knowledge are necessary elements in finding and learning the explanatory links that connect sentences in expository discourse. From an engineering point of view, this distinction provides a relevant criterion for organizing the knowledge of a program in two distinct, although not dichotomous, categories. In our ongoing research, we are investigating the role that this distinction of analytical and empirical knowledge plays in the acquisition of knowledge by being told.

This research has been funded in part by NASA/KSC Contract NAG10-0120. I would like to acknowledge Carlos Segami, Richard Hull, and Donna Markey and three anonymous referees for insightful comments on a draft of this paper.

References

Allen, J. & Perrault, C. R. (1980). Analyzing intentions in utterances. *Artificial Intelligence*, **15**, pp. 143-178.

Austin, J. L. (1962) *How to do things with words.* Harvard University Press, Cambridge, Mass.

Bateman, J. A. (1989). Upper modeling for machine translation: a level of abstraction for preserving meaning. Technical Report. USC/Information Sciences Institute.

Bateman, J. A., Kasper, R. T., Moore, J. D. & Whitney, R. A. (1990). A general organization of knowledge for natural language processing: the Penman upper model. Technical Report. USC/Information Sciences Institute.

Brachman, R. and Schmolze, J. (1985). An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, **9**, pp. 171-216.

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

Cawsey, A. (1992). *Explanation and Interaction.* MIT Press, Cambridge, Mass.

Cohen, P.R. and Lesveque, H.J. (1985). Speech acts and rationality. Proceedings 23rd Annual Conference of the Association for Computational Linguistics, Chicago, pp. 49-59.

Delisle, S. (1993) Text Processing without A-Priori Knowledge: Semi-Automatic Linguistic Analysis for Incremental Knowledge Acquisition, Ph.D Thesis, Department of Computer Science, University of Ottawa.

DeJong, G. F. & Mooney, R. J. (1986). Explanation-based learning: An alternative view, *Machine Learning*, **1**, pp. 145-176.

Dyer, M. G. (1983). *In-Depth Understanding*, MIT Press, Cambridge, MA.

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations*, MIT Press, Cambridge, MA. pp. 257-316.

Gomez, F. (1985). A Model of Comprehension of Elementary Scientific Texts. In *Proceedings of the Workshop on Theoretical Approaches to Natural Language Understanding*, Halifax, Nova Scotia, pp. 70-81.

Gomez, F. (1989). WUP: A Parser Based on Word Usage. Technical Report UCF-Tech-89-2, Department of Computer Science, University of Central Florida.

Gomez, F. & Segami, C. (1989). The Recognition and classification of concepts in understanding scientific texts, *Journal of Experimental and Theoretical Artificial Intelligence*, **1**, pp. 51-77.

Gomez, F. & Segami, C. (1991). Classification-based reasoning. *IEEE Transactions on Systems, Man and Cybernetics*, **21**, pp. 644-659.

Gomez, F., Hull, R. and Segami, C. (1994). Acquiring Knowledge from Encyclopedic Texts. In *Proceedings of the ACL's 4th Conference on Applied Natural Language Processing*, Stuttgart, Germany, pp. 84-90.

Gomez, F., Segami, H. and Hull, R. (1994) Determining Prepositional Attachment, Prepositional Meaning, Verb Meaning and Thematic Roles. Department of Computer Science, University of Central Florida. Revised version of UCF-Tech-92-1 with similar title.

Grice, H. P. (1957) Meaning. *Philosophical Review*, **66**, pp.377-388.

Grosz, B.J. & Sidner, C. L. (1986) Attention, intentions and the structure of discourse. *Computational Linguistics*, **12**, pp. 175-204.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, **63**, pp. 341-385.

Halliday, M. A. K. (1985) *An introduction to functional grammar.* London, Edward Arnold.

Katz, J. J. (1990). *The metaphysics of meaning.* MIT Press, Cambridge, Mass.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. pp. 89-107 In B. K. Britton and J. B. Black Eds., *Understanding Expository Texts*, pp. 89-107, Lawrence Erlbaum Associates.

Kolodner, J. L. (1983). Reconstructive memory: A computer model, *Cognitive Science*, **7**, 281-238.

Lambert, L. and Carberry, S. (1991). A tripartite plan-based model of dialogue. Proceedings of the 29th Conference of the Association of Computational Linguistics, Berkeley, CA. pp. 47-54.

Litman, D.J. and Allen, J. (1987). A plan recognition model for sub-dialogues in conversations. *Cognitive Science*, **11**, 2. pp. 163-200.

Maier, E. & Hovy, E. (1993). Organizing discourse structure relations using metafunctions. In Horacek, H. & Zock, M. Eds. *New concepts in natural language generation: planning, realization, and systems.* pp. 69-86. Pinter Publishers, London.

Mann, W. C. & Thompson, S. A. (1987). *Rhetorical structure theory: a theory of text organization*, ISI/IRS-87-190, University of Southern California.

Mann, W. C., Matthiessen, C. M. I. M & Thompson, S. A. (1989). Rhetorical structure theory and text analysis. In W. C. Mann & S. A. Thompson, Eds. *Text description: diverse analyses of a fund raising text.* pp. 39-78. John Benjamins, Amsterdam.

Mayer, R. E. (1985). Structural analysis of scientific prose. In B. K. Britton and J. B. Black, Eds. *Understanding Expository Texts*, pp. 65-87, Erlbaum Associates.

McKeown, K.R. (1988). Generating goal-oriented explanations.*International Journal of Expert Systems Research and Applications*, **4**, pp.377-395.

Mitchell, K. M., Keller, R. & Kedar-Cabelli, S. (1986). Explanation-based generalization: a unifying view. *Machine Learning*, **1**, pp. 47-80.

Moulin, B. and Rousseau, D. (1994). SACD: A system for acquiring knowledge from regulatory texts. *Computers Elect. Eng.*, **20**, vol. 2, pp. 131-149.

Norvig, P. (1989). Marker passing as a weak method for text inferencing. *Cognitive Science*, **13**, pp. 569-620.

Paris, C.L. (1988). Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, **14**, pp. 64-78.

Pollack, M.E. (1986). A model of plan inference that distinguishes between the beliefs of actors and observers. *Proceedings of the 24th Conference of the Association of Computational Linguistics*, New York, pp. 207-214.

Piaget, J. (1959). *Les Liaisons Analytiques et Synthetiques dans le Comportement du Sujet, Etudes d'epistemologie genetique.* Presses Universitaires de France, Paris.

Pritchett, B. L. (1992) *Grammatical Competence and Parsing Performance.* The University of Chicago Press.

Putnam, H. (1962). The analytic and the synthetic. In H. Feigl and G. Maxwell, Eds. *Minnesota Studies in the Philosophy of Science*, University of Minnesota Press, Minneapolis, pp. 269-317.

Quine, V. W. (1953). *From a Logical Point of View*, Harvard University Press.

Quine, V. W. (1960). *Word and Object*, MIT Press.

Sanders, T. J. M., Spooren, W. P. M. S. & Noordman, L. G. M. (1990) Towards a taxonomy of coherence relations. *Discourse Processes*, **15**, pp. 1-35.

Schank, R. (1982). *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*, Cambridge University Press.

Searle, J. R. (1969). *Speech Acts*, Cambridge University Press, Cambridge, 1969.

Szpakowicz, S. (1990). Semi-automatic acquisition of conceptual structures from technical texts, *International Journal of Man-Machine Studies*, **33**, pp. 385-397.

Van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension.* Academic Press.

Wilensky, R. (1983). *Planning and Understanding*, Reading: Addison-Wesley.

Wittgenstein, L. (1953). Philosophical Investigations. Basil Blackwell. Oxford.

APPENDIX

10 A Short Sample Session with the System

In this sample session, lines preceded by >>> correspond to the user's input to the system. Other lines, except for the remarks, correspond to the system response.

```
>>> the human body fights infections.
>>> the human body kills the germs that cause infections.
>>> how does the human body fight infections?

human-body work-against infection
   *how* <human-body cause-to-die <germ which cause infection>>
```

REMARKS: After the integration of the second statement above, the *explanatory* phase answers the link *How does the human body fight infections?* generated by the first statement, by firing the analytical rule "if x kills y and y causes z then this explains how x fights z." The explanation found is inserted in the appropriate *relation-structure* in LTM, and the system is then able to answer the question *How does the human body fight infections?* as shown above.

```
>>> insulin controls the level of sugar in the blood.
>>> does insulin control sugar in the blood?

::: yes
insulin control <sugar which location blood>
because
insulin control <level pertaining-to <sugar which location blood>>
```

REMARKS: Here, the question-answering module is unable to answer the question by accessing the current knowledge in LTM. Therefore, it activates the *explanatory* phase, which fires the analytical rules attached to the verbal concept *control.* One of these rules is "if x controls y and y pertains to z, then x controls z." This rule fires successfully, and the system is able to answer the question, as shown above.

```
>>> the human body produces macrophages which kill bacteria.
>>> does the human body kill bacteria?

::: yes
human-body cause-to-die bacteria
because
human-body make <macrophage which cause-to-die bacteria>
```

REMARKS: This situation is similar to the previous one. Not being able to find an answer in LTM, the question-answering module activates the *explanatory* phase, which fires the rule "if x produces y and y kills z, then x kills z."

```
>>> bacteria cause infections to humans.
>>> infections kill humans.
>>> do bacteria kill humans?

::: yes
bacteria cause-to-die human
because
infection cause-to-die human and
     bacteria cause infection *recipient*  human
```

REMARKS: The analytical rule that allowed this question to be answered is "if x causes y to z and y kills z, then x kills z."

```
>>> germs which enter the human body cause diseases.
>>> these germs are found in the food humans eat.
>>> the human body kills these germs by producing antibodies.
```

REMARKS: The second statement provides the answer to the link *How do germs enter the human body*? generated by the first statement, and this new knowledge is inserted in LTM after integrating the second statement. The system is able to answer the question below by accessing the LTM structures.

```
>>> how do germs enter the human body?

germ enter human-body *instrument* <food ingest%by human>


>>> does the human body fight diseases?

::: yes
human-body work-against disease
because
human-body cause-to-die <germ which enter human-body> and
     <germ which enter human-body> cause disease
```

REMARKS: The answer to this question is found by activating the analytical rule "if x destroys y and y causes z, then x work-against z."

```
>>> do antibodies kill germs?

::: yes
human-body cause-to-die germ *how* <human-body make antibody>
```

REMARKS: In this case, the analytical rule applied is “if x kills y by producing z, then z kills y.”

REMARKS: Empirical rules stored under concepts in LTM can be used to answer other questions. Suppose we have an empirical rule which states that: “if x is a human and y is a food and x likes y, then x eats y.” This rule is stored under the concept human in LTM. Now consider the following statements:

```
>>> americans are humans.
>>> humans like chocolate.
>>> chocolate is food.
```

REMARK: The answer to the question, *Do americans eat chocolate*? can be found by accessing the empirical rule located under the concept human, which is a superconcept of the concept “american.”

```
>>> do americans eat chocolate?

::: yes
because
    human like chocolate and chocolate is-a food and american is-a human
```

```
(human-body
   (become-aware (@x00008 ($more (@a00010))))
   (enter-by     (@x00008 ($more (@a00009)))

(@x00008
   (cf (is-a (germ)) (enter (human-body ($more (@a00009))))
   (become-aware%by (human-body ($more (@a00010))))
   (enter (human-body ($more (@a0009))))

(@a00009 (args (@x00008) (human-body))
   (instance-of (action))
   (pr (enter)) (actor (@x00008 (q (all))))
   (to-loc (human-body (q (?)))))

(@a00010 (args (human-body) (@x00008))
   (instance-of (action))
   (pr (become-aware))
   (actor (human-body (q (?))))
   (theme (@x00008 (q (?))))
```

Table 1: Representation structures built by the *formation* phase after reading the sentence *The human body recognizes the germs that enter it.*

```
(@x00016
 (cf (is-a (food)) (ingest%by (human-body ($more (@a00017))))
 (location-of (@x00008)))

(human-body (ingest (@x00016 ($more (@a00017)))))

(@x00008 (at-loc (@x00016 ($more (@a00018))))

(@a00017 (args (human-body) (@x00016)) (instance-of(action))
          (pr (ingest)) (actor (human (q (?)))) (theme (@x00016 (q (all)))))

(@a00018 (args (@x00008) (@x00016))
         (instance-of (description))
         (pr (at-loc))
         (descr-subj (@x00008 (q (?))))
         (descr-obj (@x00016 (q (?)))))
```

Table 2: Representation structures built by the *formation* phase after reading the sentence *These germs are located in the food ingested by the human body.*