SRI International

Technical Note 329 • June 1984

Sublanguage and Knowledge

Prepared by:

Jerry R. Hobbs
Senior Computer Scientist
Artificial Intelligence Center
Computing and Engineering Sciences Division

This paper will appear in *Proceedings of the Workshop on Sublanguage Description and Processing*, Ralph Grishman and Richard Kittredge, editors.

This research was supported by NIH Grant LM03611 from the National Library of Medicine, by Grant IST-8209346 from the National Science Foundation, and by a gift from the System Development Foundation.

333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

Sublanguage and Knowledge

Jerry R. Hobbs

SRI International

Menlo Park, California

1. Introduction

Over the last year and a half we have been building a fairly large knowledge base for a natural language processing system. This effort may be viewed as an alternative to, or perhaps an alternate perspective on, the sublanguage approach to text processing. Instead of specifying constraints particular to a sublanguage, one axiomatizes in a formal language some of the knowledge in the underlying domain. In this paper I first set the context by describing the system we ultimately hope to construct. A knowledge base is an important component of this system. The methodology used in building the knowledge base is then discussed -- first the method of selecting the facts to be encoded, followed by several observations on the internal organization of the knowledge base. Examples are then given of two ways in which the knowledge base will be used in discourse processing. Finally, this approach is compared to, and primarily contrasted with, the sublanguage approach.

The work described in this paper has been carried out as part of a project to build a system for natural language access to a computerized medical textbook on hepatitis, the HKB (Bernstein et al., 1980). The intent is that the user will ask a question in English, and rather than attempting to answer it, the system will return the passages in the text relevant to the question. As illustrated in Figure 1, the English query is translated into a logical form by the DIALOGIC system, a syntactic and semantic translation component (Grosz et al., 1982). The textbook is represented by a "text structure", consisting, among other things, of summaries of the contents of individual passages, expressed in a logical language. Inference procedures, making use of a knowledge base, seek to match the logical form of the query with some part of the text structure. In addition, they attempt to solve various pragmatics problems posed by the query, including the resolution of coreference, metonymy, and the implicit predicates in compound nominals.

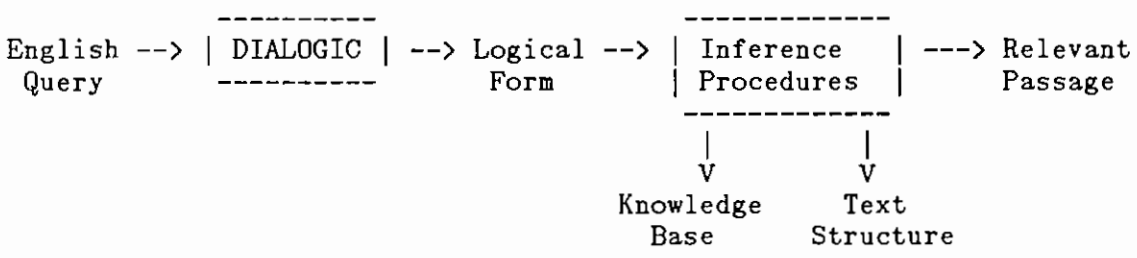

Figure 1. Structure of the Text Access System

Several examples will demonstrate just how rich the knowledge base must be for this application. In one of the dialogues we have collected, the user asks the question,

Is saliva infective?

It happens that the relevant passage is one that is about the transmission of hepatitis B virus by saliva. The knowledge base must therefore have an axiom relating "infective" and "transmission".

Similarly, the question

Can a patient with mild hepatitis go on a strenuous rock climb?

is answered by a passage whose heading is "Management: Requirements for Bed Rest". The rather indirect relation between "strenuous" and "bed rest" must be encoded in the knowledge base.

In addition to matching the query with the text structure, the system is intended to solve various discourse problems, such as the problem of coreference resolution. Consider the two successive queries

Could this episode be a reactivated form of his prior disease?

How can I confirm this possibility?

The definite noun phrase "this possibility" refers to the operand of "could" in the first sentence. To resolve it, we must have encoded the relation between "possibility" and "could".

We see that a large knowledge base is necessary. But how large? Unfortunately, there is no limit. It is well known that discourse can require arbitrarily detailed world knowledge for its interpretation. If you know that John's maternal grandmother's maiden name is "Davis", you will be able to understand the compound nominal in the sentence

John has a Davis smile.

It is of course impossible to encode this much knowledge. What we would like is a knowledge base that is richer than the simple sort hierarchy that most natural language processing systems have, but does not include facts as detailed as the maiden name of John's maternal grandmother. More particularly, we would like to develop a principled methodology for building intermediate-size knowledge bases for natural language applications. This paper describes such a methodology, developed in the course of constructing a knowledge base of about one thousand "facts" for the text access system.

One way to build the knowledge base would have been to analyze the queries in the dialogues we collected, to determine what facts they seem to require, and to put just these facts into our knowledge base. However, we are interested in discovering general principles for the selection and structuring of such intermediate-sized knowledge bases, principles that would give us reason to believe our knowledge base would be useful for unanticipated queries.

Thus, we have developed a three-stage methodology:

1. Selection of the facts that should be in the knowledge base, by determining what facts are linguistically presupposed by the medical textbook. This gives us a very good indication of what knowledge of the domain the user is expected to bring to the textbook and would presumably bring to the text access system.

2. Organization of the facts into clusters and within each cluster, according to the logical dependencies among the concepts they involve.

3. Encoding the facts as predicate calculus axioms, regularizing the concepts, or predicates, as necessary.

In this paper only the first two stages are discussed, since they are most relevant to the study of sublanguages.

Before we continue, three examples of facts and their corresponding axioms might help orient the reader. The knowledge base includes commonsense knowledge, which everyone knows, and medical knowledge, which is usually known only by experts. An example of the former is the "lexical decomposition" inference for the predicate "produce":

If x produces y, then x causes y to exist.

(A x,y)(E e) produce(x,y) --> cause(x,e) & exist'(e,y)*

The following is a medical fact that everyone knows:

Food can be the medium for transmission of an (etiologic) agent.

(A x,y)(E e,z,w) food(x) & agent(y) --> possible(e) & transmit'(e,x,y,z,w)

---

* The mildly nonstandard aspects of the logical notation used in this paper are described in Hobbs (1983, 1984). In this formula, e is an event or "self" argument, standing for y's existence. In general, if p(x) means that p is true of x, then p'(e,x) means that e is the condition of p's being true of x.

The knowledge base also includes mildly arcane knowledge about medicine, such as

The lobules of the liver move bilirubin from the blood stream into the bile.

(1) (A x,y)(E z,w) lobule(x) & blood-stream(y) --> bilirubin(z) & bile(w) & move(x,z,y,w)

It should be emphasized that the knowledge we seek to encode in the knowledge base is not the knowledge in the textbook itself, but the knowledge the reader is already expected to have when he opens the textbook.

2. Selecting the Facts

To be useful, a natural language system must have a large vocabulary. Moreover, when one sets out to axiomatize a domain, unless one has a rich set of predicates and facts to be responsible for, a sense of coherence in the axiomatization is hard to achieve. One's efforts seem ad hoc. So the first step in building the knowledge base is to make up an extensive list of words, or predicates, or concepts (the three terms will be used interchangeably here), and an extensive list of relevant facts about these predicates. The words or predicates we chose were the predicates from the text structure (about 250) and the additional content words (about 80) that occurred in our target dialogues. Since the knowledge base is intended to link the dialogues and the text structure, this seemed to be the bare minimum.

Because there are dozens of facts one could state about any one of these predicates, we were faced with the problem of determining those facts that would be most pertinent for natural language understanding in this application. Our principal tool at this stage was a full-sentence concordance of the HKB, displaying the contexts in which the words were used.

Our method was to examine these contexts and to ask what facts about each concept were required to justify each of the occurrences of the words. Morphologically related (and not so related) words were considered as evidence for a single underlying predicate. Thus, "vary", "varied", "varying", "variable", "various", "variation" and "variety" were all seen in this application as decomposable into expressions involving the predicate "vary", and "oral" was viewed as related to the predicate "mouth".

The three principal linguistic phenomena we looked at were predicate-argument relations, compound nominals, and conjoined phrases. As an example of the first, consider two uses of the word "data". The phrase "extensive data on histocompatibility antigens" seems to presuppose that the reader knows the fact about data that it is a set (justifying "extensive") of particular facts about some subject (justifying the "on" argument). The phrase "the data do not consistently show ..." points to the fact that data is assembled to support some conclusion. To arrive at the facts, we ask questions like "What is data that it can be extensive or that it can show something?"

For compound nominals we ask, "What general facts about the two nouns underlie the implicit relation?" So for "casual contact circumstances" we posit that contact is a concomitant of activities; the phrase "contact mode of transmission" leads us to the fact that contact possibly results in transmission of an agent. Conjoined noun phrases indicate the existence of a superordinate in a sort hierarchy covering all the conjoined concepts. Thus, the phrase "epidemiology, clinical aspects, pathology, diagnosis, and management" tells us to encode the facts that all of these are aspects of a disease; "renal dialysis units and other high-risk institutional settings" tells us that a renal dialysis unit is a high-risk setting.

As an illustration of the method, let us examine various uses of the word "disease" to see what facts it suggests:

"destructive liver disease": A disease has a harmful effect on one or more body parts.

"hepatitis A virus plays a role in chronic liver disease": A disease may be caused by an agent.

"the clinical manifestations of a disease": A disease is detectable by signs and symptoms.

"the course of a disease": A disease goes through several stages.

"infectious disease": A disease can be transmitted.

"a notifiable disease": A disease has patterns in the population that can be traced by the medical community.

We emphasize that this is not a mechanical procedure but a method of discovery that relies on our informed intuitions. Since it is largely background knowledge we are after, we cannot expect to get it directly by interviewing experts. Our method is a way of extracting it from the presuppositions behind linguistic use.

The first thing our method gives us is a great deal of selectivity in the facts we encode. Consider the word "animal". We know hundreds of facts about animals. In this domain, however, there are only two facts we need. Animals are used in experiments, as seen in the compound nominal "laboratory animal", and animals can have a disease, and thus transmit it, as seen in the phrase "animals implicated in hepatitis". Similarly, the only relevant fact about "water" is that it may be a medium for the transmission of disease ("water-borne hepatitis"). Similarly, all that is relevant about alcohol is that it is consumed by and may be harmful to people, as indicated by the phrase "prescription regarding alcohol intake".

By using this method, we often find differences between the meaning of a word in general English and its meaning in the sublanguage of this application. For example, the word "history" in this application does not refer to a record of past events, but only to a record of past episodes of disease in a patient. The word "vertical" in this application does not mean "perpendicular to the horizon", but refers rather to the transmission of disease from a mother to a fetus. In addition, concepts must be categorized in different ways. For example,

we do not want to say that persons are animals, for almost none of the inferences are the same for the two. A person cannot be a laboratory animal, and an animal cannot be a patient (in this application). One of the few relevant properties they share is that both can be sources of an infection.

When we see a number of uses that seem to fall within the same class, the method leads us toward generalizations we might otherwise miss. For example, the uses of the word "laboratory" seem to be of two kinds:

1. "Laboratory animals", "laboratory spores", "laboratory contamination", "laboratory methods".

2. "A study by a research laboratory", "laboratory testing", "laboratory abnormalities", "laboratory characteristics of hepatitis A", "laboratory picture".

The first of these rests on the fact that experiments involving certain events and entities take place in laboratories. The second rests on the fact that information is acquired there.

A classical issue in lexical semantics that arises at this stage is the problem of polysemy, and the concordance method suggests solutions. Should we consider a word, or predicate, as ambiguous, or should we try to find a very general characterization of its meaning that abstracts away from its use in various contexts? The rule of thumb we have followed is this: if the uses fall into two or three distinct, large classes, the word is treated as having separate senses, whereas if the

uses seem to be spread all over the map, we try to find a general characterization that covers them all. The word "derive" is an example of the first case. A derivation is either of information from an investigative activity, as in "epidemiologic patterns derived from historical studies", or of chemicals from body parts, as in "enzymes derived from intestinal mucosa". By contrast, the word "produce" (and the word "product") can be used in a variety of ways: a disease can produce a condition, a virus can produce a disease or a viral particle, something can produce a virus ("the amount of virus produced in the carrier state"), intestinal flora can produce compounds, and something can produce chemicals from blood ("blood products"). All of this suggests that we want to encode only the fact that if x produces y, then x causes y to come into existence. The word "distribute" is an intermediate case. We could view it as having three senses: an abnormal condition is distributed through tissue, a disease is distributed through a population, and various degrees of severity are distributed through a set of cases. Or we could characterize it in the abstract terms of a condition's or property's being distributed among the elements of a system. My predilection is for a single, more general predicate, where possible.

The polysemy problem is complicated by another issue -- metonymy. Consider an example. We can talk about both "managing a pathological condition" and "managing a patient". There are three ways we could view this situation. (1) "Manage" is an polysemous word. There is a sense

"manage1", which means to manage a pathological condition, and a sense "manage2", which means to manage a patient. (2) We could say that there is only one word "manage" but that it has disjunctive selectional constraints. If x manages y, then either y must be a pathological condition or y must be a patient. A difficulty with this approach is that different inferences will be drawn from "x manages y", depending on which disjunct is true of y. (3) We could say that we have an example of metonymy, or referring to an entity by referring to something functionally related to it, and that the apparent referent must be coerced into the intended referent. Here it could go either way. We could say that one can only manage a pathological condition, and when we encounter "manage a patient", we must coerce it into "manage a pathological condition in a patient". Or we could say that one can only manage a patient, and when we encounter "manage a pathological condition" we must coerce it into "manage a patient with a pathological condition". I have chosen to view "manage" in the last of these ways, and in general I favor the metonymy approach over the ambiguity approach.

Similar choices face us with the pair "transmit an agent" and "transmit a disease", and with the pair "exposed to an agent" and "exposed to a disease". In both cases the etiologic agent has been chosen as the canonical argument of the predicate.

3. Organizing the Knowledge Base

The second stage in building a knowledge base is to sort the facts into domains and to organize them within each domain. The aim of this step is to discover gaps and logical dependencies in the knowledge base. A general principle is that we should not use a predicate in an axiom unless that predicate is characterized, though not necessarily defined, in a set of axioms elsewhere in the knowledge base.

Sorting the facts into natural domains, or "clusters" (see Hayes, 1984), is for the most part fairly straightforward. For example, the fact "If x produces y, then x causes y to exist" is a fact about causality. The fact "The replication of a virus requires components of a cell of an organism" is a fact about viruses. The fact "A household is an environment with a high rate of intimate contact, thus a high risk of transmission" is in the cluster of facts about people and their activities. The fact "If bilirubin is not secreted by the liver, this may indicate injury to the liver tissues" is in the medical-practice domain.

Problems sometimes arise, however. Concepts can span two domains, like the word "strain". Viruses come in types, such as the hepatitis A virus and the hepatitis B virus, generally defined by the disease they cause. But within each type there may be several strains, definable by properties involving laboratory tests. Only viruses come in strains in

this application, but the concept of "strain" rests on an understanding of the nature of classification systems. Thus, it is unclear whether we should consider "strain" to be a concept in the "viruses" domain or in the "medical science" domain. Since the division into domains is primarily a heuristic aimed toward achieving elegant axiomatizations, a reasonable rule of thumb is to choose the domain in which the concept would exert the most influence on the formal characterizations of other facts.

It is useful to distinguish between commonsense domains, which would be useful for most natural language applications, and medical knowledge. Among the commonsense domains are space, time, belief, modality, normality, and goal-directed behavior, but at the very foundation of the knowledge base is a domain that might be called "naive topology". The medical knowledge includes clusters of facts about viruses, immunology, physiology, disease, and medical practice and science. The cluster of facts about people and their activities lies somewhere between these two categories.

We are taking a rather novel approach to the axiomatization of commonsense knowledge. Much of our knowledge and our language seems to be based on an underlying "topology", which is then instantiated in many other areas, such as space, time, belief, biological systems, social organizations, and so on. We have begun by axiomatizing this fundamental topology. At its base is set theory, axiomatized along traditional lines around the predicates "member" and "subset". Some

common ways of talking about "predication" are then axiomatized, in a way that is abstracted away from content. These include the expressions "having a property" and "being in a state". This is required for such concepts as "related to", "associated with", "aspect", "sharing a property" and "differ". Next a theory of granularity is encoded, in which the key concept is "x is indistinguishable from y with respect to granularity g". The progressive tense and the contrasting uses of "at" and "in", among other things, require such a theory. Next a "scale" is defined as a partial ordering with an associated granularity. A theory of scales is developed to the point that such words as "minimal", "limit" and "range" can be characterized.

Next there is an axiomatization centered around the notion of a "system", which is defined as a set of entities and a set of relations among them. In the "system" cluster we provide an interrelated set of predicates enabling one to characterize the "structure" of a system, producer-consumer relations among the components, the "function" of a component of a system as a relation between the component's behavior and the behavior of the system as a whole, and distributions of properties among the elements of a system. The notion of "system" is broadly applicable; among the entities that can be viewed as systems are viruses, organs, activities, populations, and scientific disciplines.

Among the other "naive topological" concepts that are given (perhaps overly simple) axiomatizations are the idea of the location of an entity with respect to scale or system; change of state, as required

for such words as "acquire", "appear" and "continue"; and following up a suggestion by Hayes (1984), a cluster of facts centered upon the notions of "enclosure" and "causality", building toward such words as "produce", "resist" and "penetrate".

Other general commonsense knowledge is built on top of this naive topology. The domain of time is seen as a particular kind of scale defined by change of state, and the axiomatization builds toward such predicates as "regular" and "persist". The domain of belief has three principal subclusters in this application: learning, which includes such predicates as "find", "test" and "manifest"; reasoning, explicating predicates such as "leads-to" and "consistent"; and classifying, with such predicates as "distinguish", "differentiate" and "identify". The domain of modalities explicates such concepts as necessity, possibility, and likelihood. The domain of normality deals with such concepts as "abnormal", "disturbance" and "recover". Finally, in the domain of goal-directed behavior, we characterize such predicates as "help", "care" and "risk".

The lowest level of the knowledge base that has explicit medical content includes the domains of "viruses", "immunology", "physiology", and "people and their activities". Within "physiology" there are sublevels for "biochemistry", "cells, tissues and body fluids", and "organs and bodily systems". The facts about biochemistry are primarily of a classificatory nature. Those about cells and tissues concern their interactions with one another, with etiologic agents, with the

immunological system, and with body fluids. The facts about organs and bodily systems concern primarily their structure and functions, such as axiom (1) in Section 1.

"People and their activities" is of course a huge domain. For this application, however, the only facts we need to know about people are that they have bodies, they are aware of certain abnormal conditions in their bodies, their activities bring about contact with other people and hence possibly the transmission of disease, and their activities consume bodily resources that could be used for recovery instead. We also need certain notions of instrumentality, for words such as "instrument" and "equipment". An activity, such as eating, is defined as a system of actions. An environment, such as an institution, household or school, is defined as a system of people and associated activities.

The "disease" domain is defined here primarily in terms of a temporal schema: A virus or other agent enters a body; the immunological system responds; there is an incubation period during which there are no apparent effects; with onset the infection becomes apparent and can be described at the level of tissues and other body parts or at the level of the person as a whole and his or her awareness; and one of several possible outcomes results.

Above the level of disease is the domain of "medical practice" or medical intervention in the natural course of the disease. It can be axiomatized as a plan, in the artificial intelligence sense, for

maintaining or achieving a state of health in the patient. The different branches of the plan correspond to where in the temporal schema for disease the physician intervenes and to the mode of intervention. Thus, prevention and treatment are intervention before and after onset, respectively. To prevent a disease, one can block its transmission by avoiding certain kinds of contacts, or one can by immunization prevent the infection from becoming established. Diagnosis is the acquisition of the knowledge that is a prerequisite for treatment.

The domain of "medical science, takes all of the foregoing as its subject matter. It can be viewed as the discovery and organization of shared knowledge about this subject matter. Discovery is of two types -- data collection, such as experiments and surveys, and inducing general principles. Organization involves classification and standardization of such things as nomenclature.

It is not clear whether there are general principles for axiomatizing these domains, but certain strategies prove useful. In each of these domains we begin by specifying its ontology -- the different sorts of entities and classes of entities in the domain -- and the inclusion relations among the classes. Next we ask what relations and interactions occur among the entities, such as the "react with" relation between antigens and antibodies in the "immunology" domain; for these predicates, we encode the constraints they impose on their arguments. The entities in one domain are frequently systems whose

components are entities in another domain, as organs are systems of tissues. Where this is the case, articulation or bridging axioms are written to express the relations between the properties and behavior of the components and the properties and behavior of the system as a whole. We then often want to specify temporal schemas involving interactions of entities from several domains, goal-directed intervention in the natural course of these schemas, and efforts at learning and systematizing all this subject matter.

The concordance method of the second stage is quite useful in ferreting out the relevant facts, but it leaves some lacunae, or gaps, that become apparent when we look at the knowledge base as a whole. The gaps are especially frequent in commonsense knowledge. The general principle we follow in encoding this lowest level of the knowledge base is to aim for a vocabulary of predicates that is minimally adequate for expressing the higher-level, medical facts and to encode the obvious connections among them. One heuristic has proved useful: If the axioms in higher-level domains are especially complicated to express, this indicates that some underlying domain has not been sufficiently explicated and axiomatized. For example, this consideration has led to a fuller elaboration of the "systems" domain. Another example concerns the predicates "parenteral", "needle" and "bite", appearing in the domain of "disease transmission". Initial attempts to axiomatize them indicated the need for axioms, in the "naive topology" domain, about membranes and the penetration of membranes allowing substances to move from one side of the membrane to the other.

Another example of a gap concerns the word "bilirubin". The concordance method yields only the fact that bilirubin is a chemical. Yet its elimination from the bloodstream is one of the principal functions of the liver; jaundice, which results from the liver's failure to eliminate bilirubin, is an important symptom of hepatitis. This leads us to construct a fuller explication of the structure and function of the liver.

Within each domain, concepts and facts seem to fall into small groups that need to be defined together. For example, the predicates "clean" and "contaminate" need to be defined in tandem. There is a larger example in the "disease transmission" domain. The predicate "transmit" is fundamental; once it has been characterized as the motion of an infectious agent from a person or animal to a person via some medium, the predicates "source", "route", "mechanism", "mode", "vehicle" and "expose" can be defined in terms of its schema. In addition, relevant facts about body fluids, food, water, contamination, needles, bites, propagation, and epidemiology rest on an understanding of "transmit". In each domain there tends to be a core of central predicates whose nature must be explicated with some care and thoroughness. A large number of other predicates can then be characterized fairly easily in terms of these.

4. Using the Knowledge Base

The ways in which a natural language processing system would use such a knowledge base are described more fully elsewhere (Walker and Hobbs, 1981). Here we consider just two examples: matching with the text structure, since that is the primary aim of the text access system; and resolving metonymy, since that illustrates an advantage of the approach presented here over the sublanguage approach.

Suppose the user asks the question, "Can a patient with mild hepatitis engage in strenuous exercise?" The relevant passage in the textbook is labeled "Management of the Patient: Requirements for Bed Rest". The inference procedures must show that this heading is relevant to the question by drawing the appropriate inferences from the knowledge base. Thus, the knowledge base must contain the facts that rest is an activity that consumes little energy, that exercise is an activity, and that if something is strenuous it consumes much energy. In addition, the knowledge base must contain axioms that relate the concepts "can" and "require" via the concept of possibility.

The logical form of the query is the following: (It is assumed that certain problems of resolving implicit arguments have been solved.)

(2) can(P,E) & engage-in'(E,P,A) & patient(P) & with(P,H)
& hepatitis(H) & mild(H) & exercise(A,P) & strenuous(A)

The text structure representation of the content of the passage in the HKB is as follows:

(3) require(p,r) & patient(p) & with(p,h) & hepatitis(h) & rest(r)

The following set of axioms allows us to deduce almost a complete match of (2) and (3) by forward-chaining and looking for unifiable predications.

```
If p requires r, then r is necessary for p.
(A p,r) require(p,r) --> necessary(r,p)

If r is necessary for p, then it is not the case that n is
   possible for p, where n is the nonoccurrence of r.
(A r,p)(E n) necessary(r,p) --> ~ possible(n,p) & not'(n,r)

If p can do r, then r is possible for p.
(A r,p) can(p,r) --> possible(r,p)

If p engages in an activity, then it is p's activity.
(A a,p,x) engage-in(p,a) & activity(a,x) --> p = x

If e is exercise by p, then e is an activity by p.
(A a,p) exercise(a,p) --> activity(a,p)

If a is strenuous, then a uses much energy e.
(A a)(E e) strenuous(a) --> use(a,e) & energy(e) & much(e)

Much is not little.
(A e) much(e) --> ~ little(e)

Rest r by p is an activity by p that uses little energy e.
(A r,p)(E e) rest(r,p) --> use(r,e) & energy(e) & little(e)
```

The second example of the use of the knowledge base by inference processes involves determining what congruence there is between a predicate and its arguments.

Let us suppose that there are certain axioms in the knowledge base that are written as follows:

```
(A x,y) p(x,y) : q(x) & r(y)
```

This means the same thing as

(A x,y) p(x,y) --> q(x) & r(y),

except that the inference is drawn obligatorily and the text is modified in whatever way necessary to make this inference true. This amounts to treating q and r as conditions that must be satisfied by the arguments x and y of p. satisfy. Thus, our knowledge base may contain the rule

(4) (A x,y) exposed-to(x,y) : person(x) & agent(y)

We may state the metonymy resolution operation in something like the following form:

Given "p(A)" in the text,
if "(A x) p(x) : q(x)" is in the knowledge base,
then infer "q(f(A))" from the properties of A in the text
and transform "p(A)" into "p(f(A))" in the text.

If the coercion function f turns out to be the identity function, we have simply checked that the selectional constraint q(x) is obeyed. For example, suppose the text refers to

exposure to hepatitis A virus

Then the selectional constraints on the second argument of "exposed-to", expressed in (4), can be verified by forward-chaining through the following two axioms about the hepatitis A virus, or HAV:

(5) (A y) HAV(y) --> virus(y)
(A y) virus(y) --> agent(y)

If we allow more general functions f, then we can resolve cases of metonymy, and the operation becomes one of coercion. Supposed the text had referred to

(6) exposure to type A hepatitis

Then we could perform the coercion by using the following axiom

(A z)(E y) type-A-hepatitis(z) --> HAV(y) & cause(y,z)

together with axioms (5). Then the expression in the logical form for the sentence would be expanded from

... & exposed-to(X,H) & ...

to

... & exposed-to(X,V) & HAV(V) & cause(V,H) & ...

We have thus expanded the metonymic (6) to

exposure to the hepatitis A virus that causes type A hepatitis.

5. Comparison of Approaches

The sublanguage approach to text processing, briefly and too simply, is this. One determines the various subclasses of nouns in the language used in a particular application, and one determines for verbs

and other operators what subclasses their arguments must belong to. One's parser can then use this semantic information about the domain for syntactic disambiguation and other interpretation processes.

The approach described in this paper, which might be called the "axiomatization approach", is a generalization of the sublanguage approach. If in a sublanguage we specify that the operator "exposed-to" requires as its subject a noun of subclass PERSON and as its object a noun of subclass (etiologic) AGENT, then we are stating axiom (4) in slightly different terms. Similarly, when we state that the noun HAV is in noun subclass AGENT, we are stating axioms (5) in different terms.

But sublanguage constraints encode only one kind of fact about a domain. The axiomatization approach can capture a richer set of facts and constraints, such as the bulk of the information contained in the knowledge base. For example, we can state conditions on arguments that are much more complex than simple sort constraints. Consider the predicate "range". If x ranges from y to z, then x must be a set whose members are located on a scale on which y and z are points, where y is less than z on the scale. This may be stated as follows:

```
(A x,y,z) range(x,y,z) : (E s) set(x) & scale(s)
            & on(y,s) & on(z,s) & y<z
            & (A w)(member(w,x) --> (E u)(on(u,s) & at(w,u)))
```

A rule like this will give us a chance of interpreting such examples as

severity can range from mild to fulminant,

where we must expand "severity" into "the severity of a set of possible cases of infection". It is not clear how one could express such constraints in the sublanguage approach.

Finally, there is the problem of metonymy. Natural language discourse is riddled with examples of metonymy, and I see no obvious way of handling metonymy in the sublanguage approach. In the approach proposed in this paper, it can be handled by a simple extension of the mechanism for checking selectional constraints, as described in Section 4.

But the sublanguage approach was developed with a view to computational efficiency. It is a way of transforming an important part of a difficult semantic problem into a problem in which well-understood syntactic methods apply. What can be said about the axiomatization approach in this regard? Is it computationally tractable?

I of course view this work as long-term research. The approach will not be computationally tractable in the near future. But is there any reason to believe it will not be a computational disaster in the long run?

Three possible answers occur to me. The first is that parallel architecture will come to our rescue. But this answer is irresponsible unless we ourselves are helping to design the machines we expect to rescue us, and I am not.

A more responsible answer is that it is a viable research strategy first to discover what classes of inferences are used most frequently in sophisticated natural language processing, and only then to work on the optimization of these classes. This contrasts with a more common strategy of devising some class of inferences that looks useful in a few examples and beginning immediately to optimize this class. If we are to adopt the former strategy, the essential first step is to build a large knowledge base for use in a natural language system, so that we can acquire the necessary experience.

The third answer is that very little is known, beyond anecdotes, about the relation between the complexity properties of inference processes and the structure of the knowledge base they run on. It is generally assumed that the more axioms one has in the knowledge base, the less efficient the inference processes will be. But a simple example shows that this is not necessarily true. Suppose we add one thousand axioms to our knowledge base of the form

$$P_i \rightarrow Q, \text{ for } i = 1, \ldots, 1000,$$

where all the Pi's are different. If our inference processes are entirely backward-chaining, we have introduced a computational disaster. On the other hand, if they are entirely forward-chaining, we have not made the inference processes less efficient at all. In order even to begin investigating this relationship, we need much more experience with a large knowledge base like the one whose construction is described in this paper.

Acknowledgments

I am indebted to Bob Amsler and Don Walker for discussions concerning this work. This research was supported by NIH Grant LM03611 from the National Library of Medicine, by Grant IST-8209346 from the National Science Foundation, and by a gift from the System Development Foundation.

REFERENCES

Bernstein, L., E. Siegel, and C. Goldstein, 1980. "The Hepatitis Knowledge Base," Annals of International Medicine, Vol. 93, pp. 165-222.

Grosz, B., N. Haas, G. Hendrix, J. Hobbs, P. Martin, R. Moore, J. Robinson, and S. Rosenschein, 1982. "DIALOGIC: A Core Natural Language Processing System," Proceedings of the Ninth International Conference on Computational Linguistics, (Prague, Czechoslovakia), pp. 95-100.

Hayes, P., 1984. "The Second Naive Physics Manifesto," To appear in J. Hobbs and R. Moore, eds., Formal Theories of the Commonsense World (Ablex Publishing Company, Norwood, New Jersey).

Hobbs, J., 1983. "An Improper Treatment of Quantification in Ordinary English," Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, (Cambridge, Massachusetts), pp. 57-63.

Hobbs, J. 1984. "Ontological Promiscuity," Unpublished manuscript.

Walker, D. and J. Hobbs, 1981. "Natural Language Access to Medical Text," SRI International Technical Note 240. March 1981.