A Broad-coverage Natural Language Analysis System

Karen Jensen
IBM
April, 1989

0. Introduction

This paper discusses the components of our broad-coverage natural language analysis system, as they appear at this time.

A broad-coverage goal requires a robust and flexible natural language processing base, one that is adaptable to linguistic needs and also to the exigencies of computation. The Programming Language for Natural Language Processing (PLNLP: Heidorn 1972) is well suited for this task. PLNLP provides a general programming capability, including a rule-writing formalism and algorithms for both parsing ("decoding") and generation ("encoding"). Although linguistic scholarship and linguistic intuitions motivate our system strongly, we have chosen not to commit our computational formalism to any of the reigning linguistic theories. To quote Ron Kaplan:

> the problem is that, at least in the current state of the art, (linguists) don't know which generalizations and restrictions are really going to be true and correct, and which are either accidental, uninteresting or false. The data just isn't in... (Kaplan 1985, p. 5)

So our work is experimental, descriptive, and data-driven. This does not mean that it has no theoretical implications. Any functioning unit of this size is an embodiment of some theory. The theory behind this program of grammar development just hasn't been thoroughly articulated yet.

The system that is emerging has, so far, three components:

1. The PLNLP English Grammar (PEG) makes an initial syntactic analysis for each input sentence (Jensen 1986).

2. The reattachment component takes syntactically consistent, but semantically inaccurate, parses, and then reattaches constituents, when necessary, based on information gained from a rich semantic data base (Jensen and Binot 1987).

3. The paragraph modelling component receives sentence parses and, for connected text, builds them into logically consistent and coherent models of the chunks of discourse that are typically called paragraphs (Zadrozny and Jensen 1989).

Hand-in-hand with each of these components goes a separate dictionary access.

1. The first dictionary access (for PEG) is to a lexicon that is essentially just a glorified word list. However, it is a word list that, when coupled with morphological rules and a default strategy provided by the access mechanism, aims at supplying an entry for every word of the language, including neologisms. We started with the full online *Webster's Seventh New Collegiate Dictionary* (W7). We have modified this word list somewhat, but only to enlarge it -- never to reduce its scope. Although the word coverage is great, the amount of information per word

is small. Only reduced, streamlined feature information is available in each entry; subcategorization, or valency, information is not distinguished by word senses.

2. The second dictionary access (for reattachment) consults a far richer source than before. For English, we make central use of online dictionary entries -- both their definitions and their example sentences. W7 and the *Longman Dictionary of Contemporary English* (LDOCE) are available to us. We can parse the definitions and examples with PEG, and use the syntactic information that PEG provides in order to bootstrap our way into semantics. The amount of information per word obtainable during this second access is huge -- much greater than what is typically described, even for lexicalist systems.

3. The third access (for paragraph modeling) again includes full natural language text. Since this component is only at a very early stage, there is not much to be said about it. We envision a NL knowledge base that contains information from every available source, from word lists to dictionaries and beyond, to encyclopedias.

It is interesting that the purposes of the separate components divide so neatly along linguistic levels: syntax, semantics, discourse. We do not mean to insist that the ultimate version of this system would need to have its components so cleanly divided. Neither has separation of the components been done for reasons of theoretical elegance or symmetry, but simply because the necessities of broad-coverage NLP have brought it about.

1. A syntactic sketch: PEG

PEG is an augmented phrase structure grammar which has been useful in a number of different settings -- text critiquing and machine translation, to name two. PEG's significant characteristics include:

- binary rules, in most cases (Jensen 1987);
- a wealth of conditions on the operation of the rules -- conditions that range from those that are strongly general, and express real grammatical patterns of the language, to those that are quite specific, and are intended to filter out certain semantically anomalous parses;
- a "relaxed" or "textual" approach to parsing, which means that we consistently avoid the use of selectional ("semantic") information to condition the parse, and that we also try, in so far as possible, to avoid, or at least to soften, the use of subcategorization (valency) information for that purpose. We assume, for example, that almost any verb can have a sense which will fit almost any frame; and that almost any noun might be used as an argument to almost any verb; and that the job of a computational parsing grammar is not to separate grammatical and ungrammatical sentences, but to provide the most reasonable analysis for any input string. The system is certainly able to distinguish grammatical from ungrammatical input, but this can be done by commenting on, rather than by failing to accept, an ungrammatical string.

The lexicon that supports this initial syntactic parse started out, in 1981, as a list of all the main entries in W7 -- minus, of course, morphological variants that could be productively described by rules. W7 claims to have 130,000 entries; after morphological variants were subtracted, the list contained 63,850 entries. That number has been increased from time to time; it now stands at roughly 70,000. As stated earlier, the goal of this lexicon is to supply useful syntactic information for every word of the language, including neologisms.

Because it contains so many entries, this lexicon provides very broad coverage. However, for each entry it contains only very limited information. The information is for parts of speech, morphology (tense, number, etc.), and word class features (transitive, ditransitive, factive, etc.). The features are mostly binary (present or absent), but include some lists, such as lists of verbal particles.

Word class features are valency features -- granted. But both the presentation and the use of these features are different from what is described for most other parsing systems. First, no attempt is made to specify the nature of the valency arguments. Second, although different parts of speech for a single word are listed and marked separately, all other sense distinctions, within each part of speech, are collapsed. One lexical item might have many, often contradictory, feature markings. The word "go," for example, appears in the lexicon as follows:

go(NOUN SING)
go(VERB COPL INF PLUR PRES TRAN)

The first definition of "go," as a SINGular NOUN, collapses two different noun entries for "go" in W7. One is the Japanese game; the other has seven subsenses, including "the act or manner of going"; "the height of fashion"; etc. The definition of "go" as a VERB collapses 19 intransitive or COPLulative senses (e.g., "to go crazy"), and six TRANsitive senses (e.g., "to go his way," "to go bail for").

The word "know" also has two entries:

know(NOUN SING)
know(VERB INF NPTOV PLUR PRES THATCOMP TRAN WHCOMP)

This means that "know" can be a singular noun ("in the know") or a verb. If it is a verb, besides being INFinitive, PLURal, and PRESent, it might be expected, with fair frequency, to have one of the following complementation types:

NPTOV: We know him to be a good man.
THATCOMP: We know that he is here.
TRAN: We know him.
WHCOMP: We know what he wants.

The great advantage to this collapsing strategy (affectionately known as "smooshing") is that it helps to avoid multiple parses in a simple, straightforward way. And this is no trivial accomplishment: a broad-coverage, bottom-up parallel parser can easily strangle on proliferating parses. With simple lexical information, however, we can expect a manageable number of parses, even in the worst case. We aim for a single parse that carries forward all of the necessary data. We like to think of this as a syntactic sketch; we have also called it an "approximate parse." The techniques for writing this kind of grammar are varied, and use all sorts of syntactic and morphological hooks. We can exploit the presence of valency features, but we try to blunt their force, using them to favor one situation over another, rather than as strict necessary conditions for the success of a certain rule.

The result of the operation of PEG's augmented phrase structure rules, coupled with the streamlined lexicon just described, is an attribute-value data structure (in PLNLP terms, a "record structure"). Here is a somewhat pared-down example of the top-level record produced from the simple input sentence, "Geometry is a very old science":

```
SEGTYPE    'SENT'
SEGTYP2    'DECL'
STR        " geometry is a very old science"
RULES      4000 4080 5080 7200
BASE       'BE'
POS        VERB
INDIC      SING PRES COPL PERS3
PRMODS     NP1 "geometry"
HEAD       VERB1 "is"
PSMODS     NP2 "a very old science"
PSMODS     PUNC1 "."
SUBJECT    NP1 "geometry"
PREDNOM    NP2 "a very old science"
TOPIC      NP1 "geometry"
```

Figure 1. PLNLP record for "Geometry is a very old science"

Attribute names are in the left-hand column; their values are to the right. The attributes SEGTYPE and SEGTYP2 refer to different labelings of the topmost node; STR has as its value the character string covered by this node; and RULES contains a list of rule numbers, a derivational history for the parse at this level. POS indicates the possible parts of speech of the BASE; the INDICator features are fairly self-explanatory. Most of the values in Fig. 1 are actually pointers to other records. For example, the value of the PRMODS attribute is a pointer to the noun phrase (NP1) which covers the noun "geometry."

All of the analysis information is carried in the record structure. For ease of recognition, however, we also display a variant of the standard parse tree:

```
-----------------------------------------------------------------------
DECL1   NP1     NOUN1*  "geometry"
        VERB1*  "is"
        NP2     DETP1   ADJ1*   "a"
                AJP1    AVP1    ADV1*   "very"
                        ADJ2*   "old"
                NOUN2*  "science"
        PUNC1   "."
-----------------------------------------------------------------------
```

Figure 2. Parse tree for the same sentence

Note that the start node presents the value of the SEGTYP2 attribute from Fig. 1, plus a number (each node is numbered for easy reference). The other, fairly standard, node names are the values of the SEGTYP2 attributes in their corresponding records. Trees are produced by a routine that uses just five attributes from the record structure: PRMODS, HEAD, PSMODS, SEGTYP2, and STR. Since such a tree is conventionally said to depict phrase- or constituent-structure, it might be said that these five attributes make up the *constituent structure* for the parse.

More than constituent structure is contained in the records, however. During the operation of the grammar rules, attributes are assigned that point to subject, object, indirect object, predicate nominative, etc. In other parlance, these might be assigned by "...a function that goes from the nodes of a tree into f-structure space" (Kaplan 1985, p. 11). Figure 1 shows two examples, SUBJECT and PREDNOM. Such attributes, and their values, could be said to present the *functional structure*. The TOPIC of the sentence is also computed, based on some exploratory work done in Davison 1984. Other attributes will be added during further processing, and these attributes will define higher levels of analysis. Progress in the analysis seems not to involve jumping between levels, but rather a smooth accumulation (and sometimes an erasing) of attributes and values.

Now, some people might object that the same analysis could be obtained by using subcategorization frames (together, perhaps, with selectional features on NPs), either as conditions on the rules or, within a lexicalist framework, as statements within the dictionary, to be honored by the rules. According to this way of thinking, we would control multiple parses by exercising valency information, not by ignoring it. From experience, we have found this to be a dangerous path, for several reasons. The most forceful reason is that real text (at least, real English text) just does not behave in the well-disciplined fashion that such specifications would require. If we really want to do broad-coverage parsing, then we have to be prepared for many imaginative uses of words to occur; and strict subcategorization does not allow for that.

Strict subcategorization expects, for example, that verbs will occur in well-defined contexts. "Give" should be either transitive or ditransitive, surely not intransitive. But what about the sentence "I gave at the office"? It's no good saying that there is an "understood" NP; if the computational grammar depends on the presence of at least one object in context, then this sentence will fail to parse. And even though there are subcategorizational differences between "go" and "know" (by our own earlier definitions), it is possible to use "go" with a *that*-complement, as in:

> I said, no. And then he goes, "See you later."

or with a *wh*-complement, as in:

> We'll go whatever amount (i.e., bail) is necessary.

These real-life facts of language tend in one direction: stated in extreme form, any word can, and might, be used in any context. But to mark every verb in the lexicon with every possible subcategorization frame would be absurd, of course. And to add some sort of 'recovery' procedures into the grammar would be costly. The most sensible way to regard subcategorization (valency frames) is as codified frequency information. A verb that is marked transitive is quite frequently used in its transitive sense -- that's all.

This does not mean that we ignore the semantic implications of valencies. On the contrary, what we do is postpone the differentiation of word senses until after the initial syntactic sketch is completed. This strategy allows us to get our hands on any input string, assign it some (reasonable, we hope) structure, and then interpret the input, whatever it might be. Before making the interpretation, however, the parse may have to pass through the reattachment component.

2. Semantic readjustment

No matter how clever the grammarian's exploitation of word order, word class, and morphological hooks is, there are many analyses in English that just will not yield a correct analysis from syntax alone. Among these are the correct attachment of prepositional phrases and of relative and other embedded clauses; the optimal structure of complex noun phrases; and the degree of structural ambiguity exhibited by coordinated elements (Langendoen, p.c.). There are no markers, in English, that serve to disambiguate these constructions; the plain fact is that semantic (or even broader, contextual) information is required.

Consider the following parse, summarized in Fig. 3 by its tree structure. Where the correct structure cannot be determined by syntax, attachment is arbitrarily made to the closest available node, encouraging right branching.

```
-----------------------------------------------------------
DECL2  NP6     DETP7   ADJ1*   "this"
               NOUN9*  "re-measuring"
               PP1     PP2     PREP1*  "of"
                       DETP2   ADJ2*   "the"
                       NOUN1*  "land"
       VERB2*  "was"
       AJP1    ADJ3*   "necessary"
               PP3     PP4     PREP2*  "due to"
                       DETP3   ADJ4*   "the"
                       AJP2    ADJ5*   "annual"
                       NOUN2*  "overflow"
                       PP5     PP6     PREP3*  "of"
                               NP2     DETP4   ADJ6*   "the"
                                       NP3     NOUN3*  "river"
                                       NOUN4*  "Nile"
                        ?      CONJ1*  "and"
                               NP4     DETP5   ADJ7*   "the"
                                       AJP3    ADJ8*   "consequent"
                                       NOUN5*  "destroying"
                                       PP7     PP8     PREP4*  "of"
                                               DETP6   ADJ9*   "the"
                                               NOUN6*  "boundaries"
                                               PP9     PP10    PREP5*  "of"
                                                       NP5     NOUN7*  "farm"
                                                       NOUN8*  "lands"
       PUNC1   "."
-----------------------------------------------------------
```

Figure 3. Parse tree for a sentence with structural ambiguity

The question mark indicates doubt about the acceptability of the coordinate NP inside PP5: "the river Nile and the consequent destroying of the boundaries of farm lands." Should NP4, "the consequent destroying...," be *and*-ed with NP2, "the river Nile," or with the NP in PP3, "the annual overflow..."?

Question marks are placed at various points in the parse tree by a routine that is sensitive to problematic constructions in English. We could have produced two separate analyses; but, given the large number of such attachment situations, this approach would have led straight to the fatal trap of proliferating parses. The question marks, in effect, collapse different possible parses, and allow for efficient handling of ambiguities (Jensen 1986, pp. 22-23).

Human readers of the sentence will not hesitate to say that the NP attachment shown in PP3 of Figure 3 is not the intended one; the attachment indicated by the question mark is what we want. Our problem is how to enable the computer to determine that.

The sort of information that enables the right decision to be made, in this and similar cases, generally falls under the rubric of "background" or "commonsense" knowledge. The usual method for making such knowledge available to a computer program has been to hand-code the relevant concepts, in whatever format. Although some hand-coding will undoubtedly be necessary and valuable, we approach the problem from another angle.

Written text is itself a rich source of information. It can be viewed as a knowledge base; the language that it is written in, even though this is a natural language, is a knowledge representation language. In particular, reference works like dictionaries actually contain a storehouse of commonsense knowledge. We can parse the entries in an online dictionary with a syntactic grammar, and retrieve a surprising amount of the information that is necessary to resolve syntactic ambiguities, like the one displayed in Fig. 3 (Binot and Jensen 1987, Jensen and Binot 1988).

The problem presented in Fig. 3 reduces to a question: which of the following pairs is more likely?

- *overflow* and *destroying*
- *Nile* and *destroying*

Bearing in mind the old adage that "likes conjoin," we will consider that pair more likely whose terms can be more easily related through dictionary entries -- including both definitions and example sentences. (Das Gupta 1987 also uses dictionary entries for interpreting conjoined words.)

Decisions on where to start these search procedures will ultimately be important, but here we avoid them. Assume that we start with the first pair, first word. The noun definition for "overflow" in W7 begins:

overflow...n 1: a flowing over: INUNDATION

Here "inundation" is asserted to be a synonym for "overflow." The noun "inundation" has no definition of its own, but is merely listed under the verb "inundate":

inundate...vt...: to cover with a flood: OVERFLOW

The circularity of the synonym definitions is no problem, because now we can infer something new about "overflow": it involves the act of covering by means of a flood. The definition of "flood" in W7 is not much help, but in LDOCE, the first example sentence quoted in the entry for the noun "flood," when analyzed by PEG, takes us right where we want to go:

flood..n...1...The town was destroyed by the floods after the storm.

Focusing on only the relevant information, these dictionary entries present a small part of a conceptual network:

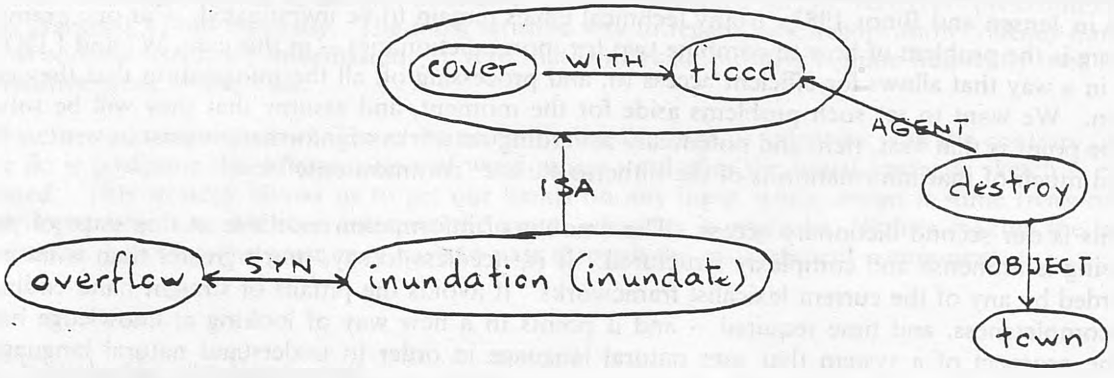

Figure 4. Network connecting "overflow" to "destroying"

and the path from "overflow" to "destroying" is clear in three steps.

Any attempt to connect "Nile" with "destroying" is bound to take longer. We can link "Nile" with "river" (this link is actually present in W7, in the Pronouncing Gazetteer); but we still have to get from "river" to "water," and then from "water" to "flood," and from "flood" to "destroy" (a total of four steps). The link between "water" and "flood" is also likely to incur a penalty, since moving from "water" to "flood" is difficult (i.e., "flood" does not appear in the definition of "water"), although moving in the reverse direction is easy ("water" does appear in the definition of "flood"). On this basis, we can revise the analysis of the sentence in Fig. 3 to reflect the more likely coordinate structure:

```
--------------------------------------------------------------
DECL2  NP6     DETP7   ADJ1*   "this"
               NOUN9*  "re-measuring"
               PP1     PP2     PREP1*  "of"
                       DETP2   ADJ2*   "the"
                       NOUN1*  "land"
       VERB2*  "was"
       AJP1    ADJ3*   "necessary"
               PP3     PP4     PREP2*  "due to"
                       NP2     DETP3   ADJ4*   "the"
                               AJP2    ADJ5*   "annual"
                               NOUN2*  "overflow"
                               PP5     PP6     PREP3*  "of"
                                       NP3     DETP4   ADJ6*   "the"
                                               NP4     NOUN3*  "river"
                                               NOUN4*  "Nile"
                       CONJ1*  "and"
                       NP5     DETP5   ADJ7*   "the"
                               AJP3    ADJ8*   "consequent"
                               NOUN5*  "destroying"
                               PP7     PP8     PREP4*  "of"
                                       DETP6   ADJ9*   "the"
                                       NOUN6*  "boundaries"
                                       PP9     PP10    PREP5*  "of"
                                               NP6     NOUN7*  "farm"
                                               NOUN8*  "lands"
       PUNC1   "."
--------------------------------------------------------------
```

Figure 5. Readjusted parse for sentence in Figure 3

We have not yet implemented this particular disambiguation, although it is similar to work reported on in Jensen and Binot 1987. Many technical issues remain to be investigated. For one example, there is the problem of how to combine two (or more) dictionaries -- in this case, W7 and LDOCE -- in a way that allows for efficient access to, and processing of, all the information that they contain. We want to set such problems aside for the moment, and assume that they will be solved. The point is that vast, rich, and potentially rewarding networks of information exist in written text, and much of that information is of the hitherto elusive "commonsense" sort.

This is our second dictionary access. The amount of information available at this stage of processing is immense and complexly structured. It is, needless to say, much greater than what is afforded by any of the current lexicalist frameworks. It avoids the pitfalls of straight hand-coding -- incompleteness, and time required -- and it points to a new way of looking at knowledge bases. The prospect of a system that uses natural language in order to understand natural language is pleasingly recursive. Words may yet prove to be the most adequate knowledge representation tools.

3. The paragraph as a discourse unit

Beyond the semantic readjustment component lies the whole world of connected text processing. This area is generally referred to as "discourse." We take the paragraph (loosely defined) to be the first formal unit of discourse. It is the smallest reasonable domain of anaphora resolution, and the smallest domain in which topic and coherence can be reliably defined (Zadrozny and Jensen 1989, p. 1, pp. 4ff).

The sentences in Figures 2 and 3 are actually part of a paragraph taken from a reading comprehension exercise in a well-known series used by countless prospective college students who want to prepare for the standard Scholastic Aptitude Test (Brownstein et al. 1987, pp. 144-5). Here is the complete text:

> Geometry is a very old science. We are told by Herodotus, a Greek historian, that geometry had its origin in Egypt along the banks of the river Nile. The first record we have of its study is found in a manuscript written by Ahmes, an Egyptian scholar, about 1550 B.C. This manuscript is believed to be a copy of a treatise which dated back probably more than a thousand years, and describes the use of geometry at that time in a very crude form of surveying or measurement. In fact, geometry, which means "earth measurement," received its name in this manner. This re-measuring of the land was necessary due to the annual overflow of the river Nile and the consequent destroying of the boundaries of farm lands. This early geometry was very largely a list of rules or formulas for finding the areas of plane figures. Many of these rules were inaccurate, but, in the main, they were fairly satisfactory.

Figure 6. Paragraph from Barron's,*How to prepare for the SAT*

PEG parse trees for the paragraph in Fig. 6, sentence by sentence, are presented in Appendix A.

If we are going to make discourse sense of this text, however, we need something more than a linear concatenation of syntactic sentence parses -- just as, in order to make syntactic sense out of a sentence, we need something more than a linear concatenation of words. A popular and effective way of modeling this non-linear set of sentence relationships is as a network with nodes connected by arcs (e.g., Sowa 1984). We can label the nodes with content words and the arcs with function (or relation) names, for a simple beginning. For now, we use a fairly intuitive set of relation names, rather than take the time to explain precisely how each arc gets labeled.

The basic network for one sentence derives not directly from the surface syntactic structure, but from the underlying predicate-argument structure, which itself is derived from the surface structure, after all necessary readjustments have been made (Jensen forthcoming). Here is a network representation, or model, for the first sentence in the geometry paragraph:

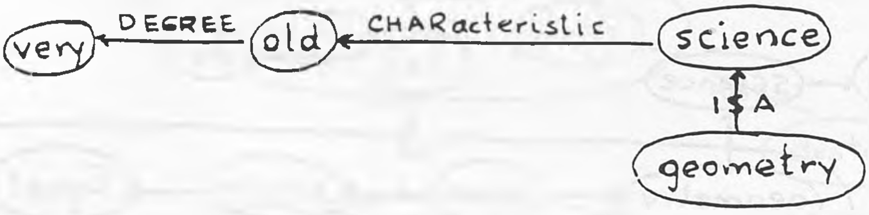

Figure 7. A network representation for "Geometry is a very old science"

To build a model for an entire paragraph (a P-model), the trick now is to map the network for each consecutive sentence onto the network for the preceding sentence or sentences, joining nodes whenever possible. Stated simply, nodes can be joined when they "mean" the same thing. To a first approximation, sameness of meaning can be defined by:

1. use of the same word;
2. use of a synonym or paraphrase;
3. use of a pronoun reference;
4. use of zero anaphora (e.g., ellipsis in coordination).

Identification of "same word" is easy enough, and syntax will suffice to determine the referents for most cases of zero anaphora, and for many pronouns. However, there are also many pronoun referents that cannot be syntactically resolved, and *nothing* in syntax will identify synonyms and paraphrases. This fact has prevented the development of a formal discourse model (Bond and Hayes 1983, p. 16).

For a solution to the problems of pronoun reference and synonym identification, we turn again to reference works written in natural language. Dictionaries and thesauri are full of such information.

Here is part of the model that can be built for the paragraph in Fig. 6. It includes information from only the first, second, fifth, and sixth sentences in that paragraph. Even so, many details have been left out:

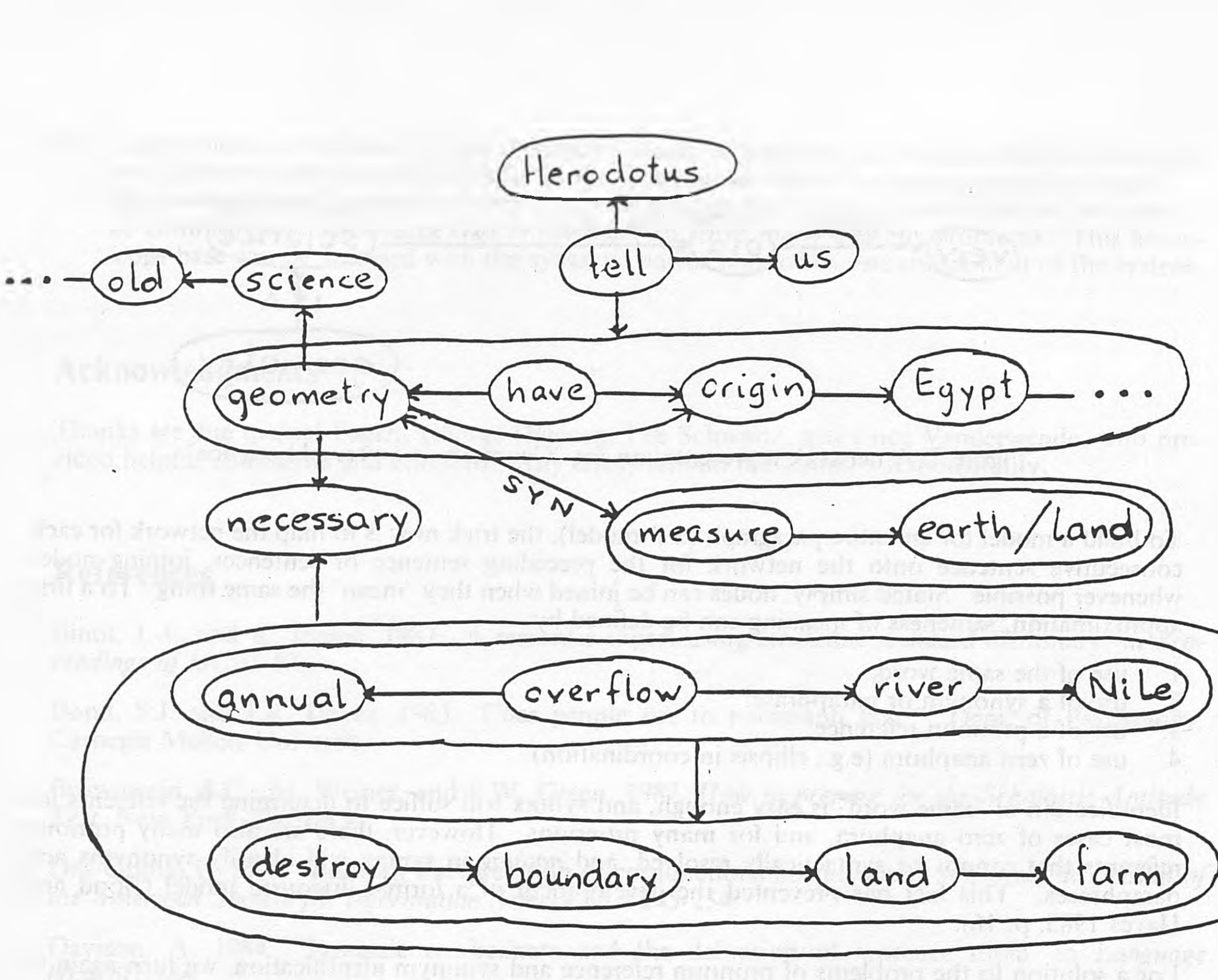

Figure 8. Partial P-model for the text in Figure 6

In order to build the link between "necessary" and "geometry," we have to know that "re-measuring of the land" is a paraphrase for "geometry." We are told that "earth measurement" is a synonym for "geometry" in the fifth sentence. Syntax allows us to say that "NOUN measurement" and "measurement of NOUN" are possible equals. If we can establish that "earth measurement" and "land re-measuring" are equals, then the problem is solved. "Measurement" and "re-measuring" are transparently related, so the problem reduces to finding a link between "earth" and "land."

This, of course, is quite easy to find in dictionaries and thesauri. In LDOCE, one definition of "earth" contains "land" as a synonym, and vice versa (actually, the first four definitions for "land" contain the word "earth" in a critical position in the parse). Similar conditions exist in W7. *Roget's Thesaurus* (RT) lists "land" as a synonym for "earth" and "earth" as a synonym for "land." Q.E.D.

The intended purpose for paragraphs like the one we have been playing with, of course, is to test a reader's comprehension ability by requiring sensible answers to questions based on the information in the paragraph. In Brownstein et al., the first test concerning our paragraph is

(1) The title below that best expresses the ideas of this passage is

and the possible solutions are

(A) Plane Figures
(B) Beginnings of Geometry
(C) Manuscript of Ahmes
(D) Surveying in Egypt
(E) Importance of the Study of Geometry

It is tempting to ask whether a program that is able to build and manipulate the P-model in Fig. 8 could also answer (1) successfully.

Without going into any formal explanation of topic definition, let's assume that we can identify the node labeled "geometry" as the main idea, or topic, of the paragraph. (Note that it occupies a central position in the network.) So we discard all possible answers to (1) except for those that contain the word "geometry." This leaves us with two candidates, (B) and (E). We then search the graph around the "geometry" node, looking for related nodes that express either "beginnings" or "importance of the study of." The latter alternative is not easy to find. But the "origin" node can be immediately identified with "beginnings." In W7, the entry for "beginning" has "origin" as a synonym, and the second sense definition for "origin" is "rise, beginning, or derivation from a source..." Furthermore, "origin" and "beginning" are mutual synonyms in RT.

Resolving the referent for the possessive pronoun "its" in the second sentence of our test paragraph allowed us to draw the arc between the "geometry" and "origin" nodes in Fig. 8, which we now label:

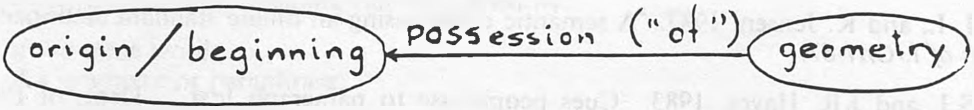

Figure 9. Network for the answer to (1)

In this subgraph, the preferred answer to question (1) is clear: the title that best expresses the ideas in the test passage is (B), "Beginnings of Geometry."

Obviously a tremendous amount of important detail has been left out in order to produce this blueprint for a formal model of a discourse unit. The challenges of implementation lie ahead. But the general structure seems promising, and most promising of all is the possibility of finding a repository of background knowledge, already coded for us, in online natural language sources.

Here is another comprehension question on the same paragraph:

(2) It can be inferred that one of the most important factors in the development of geometry as a science was

An answer must be picked from the following alternatives:

(A) Ahmes' treatise
(B) the inaccuracy of the early rules and formulas
(C) the annual flooding of the Nile Valley
(D) the destruction of farm crops by the Nile
(E) an ancient manuscript copied by Ahmes

We suggest that the preferred answer to (2) can also be found by using the P-model in Fig. 8, in conjunction with a good dictionary and thesaurus; and we leave this as an exercise for the interested reader.

4. Conclusion

This paper contains an overview of our broad-coverage NL analysis system, including components that already exist, that are currently being worked on, and that are projected for the future. Some aspects of our system that differentiate it from other NL analysis systems are

- It is not modeled along the lines of any currently accepted linguistic theory; rather it is highly experimental and data-driven.
- Separate components are emerging from this experimental process; they coincide roughly with the accepted linguistic levels: syntax, semantics, discourse.

- Each component makes its own dictionary access or accesses, and the dictionaries associated with different components will differ in the type and amount of information they contain.
- The written text of standard reference works is used as a repository for much of the background or commonsense knowledge that is necessary to solve many analysis problems. This knowledge base can be accessed with the syntactic parser that forms one component of the system.

Acknowledgments

Thanks are due to Joel Fagan, George Heidorn, Lee Schwartz, and Lucy Vanderwende, who provided helpful comments and criticism. Any errors remain the author's responsibility.

References

Binot, J.-L. and K. Jensen. 1987. "A semantic expert using an online standard dictionary" in *Proceedings of IJCAI-87.*

Bond, S.J. and J.R. Hayes. 1983. "Cues people use to paragraph text." Dept. of Psychology, Carnegie Mellon University.

Brownstein, S.C., M. Weiner, and S.W. Green. 1987. *How to prepare for the Scholastic Aptitude Test.* New York, Barron's.

Das-Gupta, P. 1987. "Boolean interpretation of conjunctions for document retrieval." in *Journal of the American Society for Information Science* 38.4.245-254.

Davison, A. 1984. "Syntactic markedness and the definition of sentence topic" in *Language* 60.4.797-846.

Heidorn, G.E. 1972. "Natural Language Inputs to a Simulation Programming System." Ph.D. dissertation, Yale University.

Jensen, K. 1986. "PEG 1986: A broad-coverage computational syntax of English." Unpublished paper.

Jensen, K. 1987. "Binary rules and non-binary trees" in A. Manaster-Ramer (ed.), *Mathematics of Language.* Amsterdam, John Benjamins, pp. 65-86.

Jensen, K. forthcoming. "PEGASUS: deriving predicate-argument structures from a syntactic parse."

Jensen, K. and J.-L. Binot. 1987. "Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions." in CL 13.3-4.251-60.

Kaplan, R. 1985. "Three seductions of computational psycholinguistics" in P. Whitelock, M.M. Woods, H.L. Somers, R. Johnson, P. Bennett (eds.), *Linguistic Theory and Computer Applications.* London, Academic Press, 1987, pp. 149-81.

Longman Dictionary of Contemporary English. 1978. Harlow and London, Longman Group Limited.

Roget's Thesaurus of English Words and Phrases. 1962. New York, St. Martin's Press.

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

Webster's Seventh New Collegiate Dictionary. 1967. Springfield, Mass., G. & C. Merriam Co.

Zadrozny, W. and K. Jensen. 1989. "Semantics of paragraphs." Unpublished paper.

Appendix A

```
Sentence 1:
-------------------------------------------------------------------
DECL1    NP1      NOUN1*   "geometry"
         VERB1*   "is"
         NP2      DETP1    ADJ1*    "a"
                  AJP1     AVP1     ADV1*    "very"
                           ADJ2*    "old"
                  NOUN2*   "science"
         PUNC1    "."
-------------------------------------------------------------------

Sentence 2:
-------------------------------------------------------------------
DECL1    NP1      PRON1*   "we"
         AUXP1    VERB1*   "are"
         VERB2*   "told"
         PP1      PP2      PREP1*   "by"
                  NOUN1*   "Herodotus"
                  PUNC1    ","
                  NAPPOS1  DETP1    ADJ1*    "a"
                           NP2      NOUN2*   "Greek"
                           NOUN3*   "historian"
                           PUNC2    ","
         VP1      COMPL1   "that"
                  NP3      NOUN4*   "geometry"
                  VERB3*   "had"
                  NP4      DETP2    ADJ2*    "its"
                           NOUN5*   "origin"
                   ?       PP3      PP4      PREP2*   "in"
                                    NOUN6*   "Egypt"
                   ?        ?       PP5      PP6      PREP3*   "along"
                                             DETP3    ADJ3*    "the"
                                             NOUN7*   "banks"
                                             PP7      PP8      PREP4*   "of"
                                                      DETP4    ADJ4*    "the"
                                                      NP5      NOUN8*   "river"
                                                      NOUN9*   "Nile"
         PUNC3    "."
-------------------------------------------------------------------
```

Sentence 3:

```
------------------------------------------------------------
DECL1  NP1     DETP1   ADJ1*   "the"
               AJP1    ADJ2*   "first"
               NOUN1*  "record"
               RELCL1  NP2     PRON1*  "we"
                       VERB1*  "have"
                       PP1     PP2     PREP1*  "of"
                               DETP2   ADJ3*   "its"
                               NOUN2*  "study"
       AUXP1   VERB2*  "is"
       VERB3*  "found"
       PP3     PP4     PREP2*  "in"
               DETP3   ADJ4*   "a"
               NOUN3*  "manuscript"
               PTPRTCL1VERB4*  "written"
        ?       ?      PP5     PP6     PREP3*  "by"
                               NOUN4*  "Ahmes"
                               PUNC1   ","
                               NAPPOS1 DETP4   ADJ5*   "an"
                                       NP3     NOUN5*  "Egyptian"
                                       NOUN6*  "scholar"
                                       PUNC2   ","
        ?       ?       ?       ?      PP7     PP8     PREP4*  "about"
                                               YEAR1*  "1550"
                                               LABEL1  NOUN7*  "B.C."
       PUNC3   "."
------------------------------------------------------------
```

Sentence 4:

```
--------------------------------------------------------------
DECL1  NP1     DETP1   ADJ1*   "this"
               NOUN1*  "manuscript"
       VP1     AUXP1   VERB1*  "is"
               VERB2*  "believed"
               INFCL1  INFTO1  "to"
                       VERB3*  "be"
                       NP2     DETP2   ADJ2*   "a"
                               NOUN2*  "copy"
                               PP1     PP2     PREP1*  "of"
                                       DETP3   ADJ3*   "a"
                                       NOUN3*  "treatise"
                                ?      RELCL1  NP3     PRON1*  "which"
                                               VERB4*  "dated"
                                               AVP1    ADV1*   "back"
                                               AVP2    AVP3    ADV2*
          "probably"

                                                       ADV3*   "more"
                                                       PP3     PP4     PREP2*
          "than"

                                                               QUANP1  ADJ4*

          "a thousand"

                                                               NOUN4*  "years"
                                                               PUNC1   ","
       CONJ1*  "and"
       VP2     VERB5*  "describes"
               NP4     DETP4   ADJ5*   "the"
                       NOUN5*  "use"
                       PP5     PP6     PREP3*  "of"
                               NOUN6*  "geometry"
                ?       ?      PP7     AVP4    ADV4*   "at that time"
                ?       ?       ?      PP8     PREP4*  "in"
                                       DETP5   ADJ6*   "a"
                                       AJP1    AVP5    ADV5*   "very"
                                               ADJ7*   "crude"
                                       NOUN7*  "form"
                                       PP9     PP10    PREP5*  "of"
                                               NP5     NOUN8*  "surveying"
                ?               ?       ?      CONJ2*  "or"
                                               NP6     NOUN9*  "measurement"
       PUNC2   "."
--------------------------------------------------------------
```

```
Sentence 5:
---------------------------------------------------------------------------
DECL1     PP1       PP2       PREP1*    "in"
                    NOUN1*    "fact"
                    PUNC1     ","
          NP1       NOUN2*    "geometry"
                    PUNC2     ","
            ?       RELCL1    NP2       PRON1*    "which"
                              VERB1*    "means"
                              NP3       PUNC3     """"
                                        NP4       NOUN3*    "earth"
                                        NOUN4*    "measurement"
                                        PUNC4     """ ,"
          VERB2*    "received"
          NP5       DETP1     ADJ1*     "its"
                    NOUN5*    "name"
            ?       PP3       PP4       PREP2*    "in"
                              DETP2     ADJ2*     "this"
                              NOUN6*    "manner"
          PUNC5     "."
---------------------------------------------------------------------------

Sentence 6:
---------------------------------------------------------------------------
DECL1     NP1       DETP1     ADJ1*     "this"
                    NOUN1*    "re-measuring"
                    PP1       PP2       PREP1*    "of"
                              DETP2     ADJ2*     "the"
                              NOUN2*    "land"
          VERB1*    "was"
          AJP1      ADJ3*     "necessary"
                    PP3       PP4       PREP2*    "due to"
                              DETP3     ADJ4*     "the"
                              AJP2      ADJ5*     "annual"
                              NOUN3*    "overflow"
                              PP5       PP6       PREP3*    "of"
                                        NP2       DETP4     ADJ6*     "the"
                                                  NP3       NOUN4*    "river"
                                                  NOUN5*    "Nile"
                                ?       CONJ1*    "and"
                                        NP4       DETP5     ADJ7*     "the"
                                                  AJP3      ADJ8*     "consequent"
                                                  NOUN6*    "destroying"
                                                  PP7       PP8       PREP4*    "of"
                                                            DETP6     ADJ9*     "the"
                                                            NOUN7*    "boundaries"
                                                            PP9       PP10      PREP5*    "of"
                                                                      NP5       NOUN8*    "farm"
                                                                      NOUN9*    "lands"
          PUNC1     "."
---------------------------------------------------------------------------
```

```
Sentence 7:
-------------------------------------------------------------
DECL1   NP1     DETP1   ADJ1*   "this"
                AJP1    ADJ2*   "early"
                NOUN1*  "geometry"
        VERB1*  "was"
        AVP1    AVP2    ADV1*   "very"
                ADV2*   "largely"
        NP2     DETP2   ADJ3*   "a"
                NOUN2*  "list"
                PP1     PP2     PREP1*  "of"
                        NP3     NOUN3*  "rules"
         ?              CONJ1*  "or"
                        NP4     NOUN4*  "formulas"
                 ?              PP3     PREP2   "for"
                                        VERB2*  "finding"
                                        NP5     DETP3   ADJ4*   "the"
                                                NOUN5*  "areas"
                                                PP4     PP5     PREP3*  "of"
                                                        AJP2    ADJ5*   "plane"
                                                        NOUN6*  "figures"
        PUNC1   "."
-------------------------------------------------------------

Sentence 8:
--- ---------------------------------------------------------
CMPD1   DECL1   NP1     QUANP1  ADJ1*   "many of"
                        DETP1   ADJ2*   "these"
                        NOUN1*  "rules"
                VERB1*  "were"
                AJP1    ADJ3*   "inaccurate"
                        PUNC1   ","
        CONJ1*  CONJ2*  "but"
                PUNC2   ","
        DECL2   PP1     PP2     PREP1*  "in"
                        DETP2   ADJ4*   "the"
                        NOUN2*  "main"
                        PUNC3   ","
         ?      NP2     PRON1*  "they"
                VERB2*  "were"
                AJP2    AVP1    ADV1*   "fairly"
                        ADJ5*   "satisfactory"
        PUNC4   "."
-------------------------------------------------------------
```