DIAGRAM: A GRAMMAR FOR DIALOGUES

Technical Note 205

February 1980

By: Jane J. Robinson, Senior Research Linguist
Artificial Intelligence Center

This paper appeared in Communications of the ACM,
Vol. 25, No. 1, (January, 1982), pg 27-48.

SRI Projects 5844 and 7910

Preparation of this paper was supported by the National Science Foundation under Grant No. MCS76-22004, and the Defense Advanced Research Projects Agency under Contract N00039-79-C-0118 with the Naval Electronic Systems Command

ABSTRACT

This paper presents an explanatory overview of a large and complex grammar, DIAGRAM, that is used in a computer system for interpreting English dialogue. DIAGRAM analyzes all of the basic kinds of phrases and sentences and many quite complex ones as well. It is not tied to a particular domain of application, and it can be extended to analyze additional constructions, using the formalism in which it is currently written. For every expression it analyzes, DIAGRAM provides an annotated description of the structural relations holding among its constituents. The annotations provide important information for other parts of the system that interpret the expression in the context of a dialogue.

DIAGRAM is an augmented phrase structure grammar. Its rule procedures allow phrases to inherit attributes from their constituents and to acquire attributes from the larger phrases in which they themselves are constituents. Consequently, when these attributes are used to set context-sensitive constraints on the acceptance of an analysis, the contextual constraints can be imposed by conditions on dominance as well as conditions on constituency. Rule procedures can also assign scores to an analysis, rating some applications of a rule as probable or as unlikely. Less likely analyses can be ignored by the procedures that interpret the utterance.

In assigning categories and writing the rule statements and procedures for DIAGRAM, decisions were guided by consideration of the functions that phrases serve in communication as well as by considerations of efficiency in relating syntactic analyses to propositional content. The major decisions are explained and illustrated with examples of the rules and the analyses they provide. Some contrasts with transformational grammars are pointed out and problems that motivate a plan to use redundancy rules in the future are discussed. (Redundancy rules are meta-rules that derive new constituent-structure rules from a set of base rules, thereby achieving generality of syntactic statement without having to perform transformations on syntactic analyses.) Other extensions of both grammar and formalism are projected in the concluding section. Appendices provide details and samples of the lexicon, the rule statements, and the procedures, as well as analyses for several sentences that differ in type and structure.

CONTENTS

I INTRODUCTION

"All grammars leak." Edward Sapir, Language, 1921.

DIAGRAM is a grammar that is used in a computer system for interpreting English dialogue. The system itself is a tool for ongoing research to establish the structures and processes necessary for interpreting utterances in a dialogue. It is a basic premise of this research that dialogue participants interpret one another's expressions by taking into account not only the truth-conditional meanings of what is said but also one another's intentions, goals, plans, beliefs, states of knowledge, and focus of attention, to the extent that these can be inferred from text and context [A. Robinson 1978]. DIAGRAM is a grammar of English--not of the whole of English, of course, but of a substantial subset of it. It analyzes all of the basic kinds of phrases and many quite complex ones as well, and is extendable in a principled way independently of a particular domain of application. The list that follows contains an illustrative sample of the kinds of syntactic constructions that are successfully analyzed at present.

Sample Sentences

[Constituent analyses of the starred sentences appear in Appendix D.]

* She was given more difficult books by her uncle.
Don't do that now.
Be very careful not to do it too quickly.
Do it quickly but be careful!
* Be careful not to break the vase when you put it down.
Tell me what he told you.
Tell me when he came.
Don't tell me any more about it!
Read these books, if you have the time.
If you have the time, read these books.
Some of them were on the table.
They must have been very big.
There are three on the table now.
* There are some men here from the city.
Those are the books she reads.

Did he give them to her?
Were they given to her?
Could she have been given many of them?
Which ones did he give her?
How much is it?
What should I do then?
I saw many more books than she did.
This is hard for me to do.
I saw that they were there.
The books that were on the table are difficult to read.
Her uncle gave the girl several books.
He gave several books to her in September.
They gave the girl some books by her uncle.
She was given some books by her uncle.
Two of them were by her uncle.
The books were given her by her uncle.
He gave up the books to her.
He broke the vase her uncle had given up to him.
The vase could break, if you aren't careful.
You could break this vase with that hammer.
That hammer could break this vase easily.
It could easily break this vase.
The vase could easily break.
It could break up easily.
You could break it up.
It might possibly have been broken by someone.
Some of us have looked into it, but it is difficult.
We saw it coming.
They found her gone.
On arriving, we had something to eat.
Could he have been here in September?
Do you think it can be assembled easily?
Have there been many people who tried it?
Didn't those people want him to try to do it?
Are there any more left?
* Is this any harder for him to do than that was?
Who gave her that book?
Who wrote it?
By whom was it written?
Who was it written by?
Who was made the leader?
Who made him the leader?
What was it written with?
Where did he try to go?
What could he have been doing there?
Why did you want John to attach it?
* Why do you think he wanted to go?
Who is there now?
What did they tell you Mary said John tried to do?
* How many more of them do you want him to have?
How long will it take you to go there?

All grammars leak and no grammar for a reasonably extensive subset of English can claim to provide all and only the correct analyses for the sentences for which it is applicable. What is claimed is that every structurally distinct semantic interpretation of the sentences shown receives a corresponding, structurally distinct, syntactic analysis. A simple example may make the point. The sentence:

She was given a book by her uncle

has four different semantic interpretations. In one, her uncle gave the book; in a second, he wrote the book; in a third, the giving took place in the vicinity of (nearby) her uncle; in a fourth, the book was located in the vicinity of her uncle. Two of the differences arise from the lexical ambiguity of "by," which can be read as either a locative or an agentive preposition. The other two correspond systematically to two different syntactic analyses, one in which the prepositional phrase is an immediate constituent of the verb phrase "given NP by her uncle" and one in which it is an immediate constituent of the noun phrase "a book by her uncle." DIAGRAM successfully provides the two appropriate syntactic analyses and rates them as equally likely. For this sentence and many other like it, DIAGRAM provides all and only the syntactic analyses that correspond to semantic ambiguities not arising from lexical ambiguity. However, for some sentences there will be incorrect analyses along with correct ones. The reasons for this are relatively well understood and will be described later, together with methods for eliminating or controlling multiple analyses without losing correct ones. (For a more detailed analysis by DIAGRAM of an ambiguous sentence, see Appendix D.)

This paper is an explanatory overview of the grammar. The explanation is aimed at two audiences: linguists who are interested in how decisions about the syntactic analysis of sentences may be affected by considerations of how sentences are used to communicate, and computer scientists who are interested in artificial intelligence (AI) systems that can process language. It should also serve as a guide to any who may want to extend and adapt the grammar to their own uses.

DIAGRAM itself has been constructed with two aims. On the one hand, its rules have been written with the aim of capturing the insightful generalizations about language that appear in the literature of theoretical linguistics. On the other hand, they have been written with a constant awareness of the functional roles that syntactic phenomena play in communication. As a result, DIAGRAM contains some nonstandard definitions of syntactic categories and constituent boundaries along with much that is widely accepted. These definitions will be explained, but not always justified. "Justifying" implies a claim that the rules are superior to any currently conceivable alternatives, and justification requires an argumentative style more suited to the exploration of moot points than to the presentation of an overview of a large grammar. Also, DIAGRAM is not proposed as a static set of rules to be set forth and justified once and for all. Its rules and the procedures for applying them have undergone and will continue to undergo revisions as better ways are found to accommodate the tension between the requirements for capturing linguistic generalizations and for designing rules that are applicable to the process of interpreting actual utterances in a computer system. Consequently, our explanation describes an intermediate stage that is interesting because it incorporates the results of experience with other formulations of rules and procedures and because its unresolved problems establish a direction for significant future research as well as for revision.

The next section presents an overview of DIAGRAM as a component in a larger system for analyzing and interpreting utterances in dialogues. The more detailed explication of DIAGRAM that follows begins with a characterization of the category S (the category of independent sentences) and the subtypes of S, before describing the categories of words and phrases that are the constituents of S. The characterization of sentences includes a discussion of embedded sentences and nonfinite but sentence-like phrases, such as infinitives and gerunds, whose categorization poses certain problems. The smaller constituents and the minor syntactic categories that appear in them are not described in detail, but a fairly complete, though simplified, list of a current set

of categories and constituent structure rules is provided in Appendix A. By starting top down in the order of description, a sense of the whole will be conveyed first so that it can shed light on how constituent parts function with respect to that whole. Indeed, the description will be as much in terms of functions as in terms of syntactic categories. "Function," here, is to be understood in two senses: the sense in which a phrase is capable of serving a purpose in discourse and the sense in which it serves one of the traditional syntactic functions such as subject, predicate, object, and modifier. Context should make clear which is meant.

II OVERVIEW

A. What DIAGRAM Does

For every expression it defines (or accepts) as syntactically well-formed, DIAGRAM provides an annotated description that makes explicit the structural relations holding among the constituent words and phrases. It identifies the syntactic categories of each word and phrase and specifies properties that each contributes to the whole expression or that each acquires as a result of its relationship to the whole. For example, applied to the analysis of

The fish are in the river

it will not only label "the fish" as a noun phrase but also as definitely determined, plural, and the subject of the sentence. Note that the last two properties are not inherent in "the fish" in isolation, but are resultant properties of its use in the whole expression. (Compare "John caught the fish," where the syntactic number is indeterminate for "fish.")

All these properties provide important pieces of information for the other parts of the system that interpret the phrase when it is used in ongoing discourse. From the fact that the noun phrase is plural, it follows that more than one of a set of discrete entities are being referred to. From the fact that it is definitely determined, it can be inferred that the speaker assumes the entities to be already identified or readily identifiable by the hearer. From the fact that it is the subject, it can be inferred that the focus of attention of the speaker is on the fish rather than on the river or on other elements in the context of the discourse. (Compare "The river contains many fish.")

B. How DIAGRAM Does It

Appendix C contains a few of the rules and a sample lexicon. Here we are concerned with the formal properties of DIAGRAM rather than with details of its format.

1. Rules and Procedures

DIAGRAM is an augmented phrase-structure grammar. Its constituent structure rules are augmented by procedures that constrain the application of a rule, add information to the structure created by the rule, and assign one or more interpretations to the resulting enriched structural analysis. The procedures for each rule apply at different points during processing. One procedure, called a constructor, applies immediately as the constituents of a phrase are being assembled. After the entire utterance has been analyzed, a second procedure, called a translator, is applied. The following simplified example should clarify the interactions of the constituent structure statements with the constructors and translators.

Rule SB1 SDEC = NP BE PRED ;

Constructor:
(1) If NP and BE do not agree in syntactic number, then REJECT the analysis.
(2) If either BE contains "not" or NP directly dominates a determiner phrase containing "no," then SDEC is Negative; else it is Affirmative.

Translator:
(1) If the syntactic number of NP is null (undefined), then set it equal to the syntactic number of BE.
(2) Set NP as Subject of SDEC.

The constituent-structure portion states that a declarative sentence may be composed of a noun phrase followed by an auxiliary BE phrase followed by an adjective phrase, as in "the fish are very big."

Within the constructor, the first (sub)procedure accepts or rejects a string of constituents for rule application according to the results of a number-agreement test. This will prevent the rule from applying to "the girls was very tall" in the course of analyzing "the

father of the girls was very tall." The second procedure sets the value of an attribute for the dominating SDEC node created by applying the rule. Notice that the constructor has access to information about the words in the constituents. It can "look down" the parse tree, so to speak, in making decisions about acceptance and in setting attributes. Alternatively, the constructor may set the value of an attribute for a phrase equal to the value for one of its constituents, so that the phrase inherits the properties of the words and phrases that compose it. The translator has similar capabilities, but it is also allowed to "look up" the tree and set attributes on the constituents of a phrase, resolving ambiguities not hitherto resolved or adding other information that cannot be determined without knowing the context in which a phrase is embedded. Thus, in the example, "the fish" acquires attributes of number and subject status.

The translator procedure for a rule can also derive one or more semantic interpretations of the phrase built by the rule. In deriving an interpretation, it can access all of the attributes whose value have been set by translators of rules that have analyzed the larger phrases in which the phrase to be interpreted forms a part. If the SDEC constructed by applying the rule in (1) is embedded in another SDEC, for instance, the translator for the embedding rule will add the information about its embedding to the embedded SDEC node. Such information is necessary for proper semantic interpretation. As an independent sentence, "the fish are in the river" will be interpreted as an assertion about the world. Embedded as part of a sentence beginning with "it is doubtful that..." or "I see that...," its interpretations will be very different.

There is also a third procedure, called an integrator, which we will only mention here, as it is more domain dependent than the other procedures. The integrator for a rule has access to facts about the world, or to a data base, and can render decisions about the pragmatic well-formedness or appropriateness of a semantic interpretation.

A constructor also may make semantic interpretations, which can serve as tests for acceptance or rejection of an analysis, according to whether the constituents obey local selectional restrictions; that is, whether or not they are semantically as well as syntactically compatible. The capability for applying tests at any of three points during the analysis of a sentence--during application of the constructor, or the translator, or the integrator procedures--raises theoretical and practical questions concerning processing. The major question is: When should semantic and pragmatic criteria be invoked to avoid false parsing paths? The question involves matters of computational efficiency and also of psycholinguistic processing, on the assumption that people store a certain amount of incoming text before they process it for its meaning. We speculate that experiments with computational efficiency, using DIAGRAM or some other sufficiently large grammar, may help us understand the psycholinguistic phenomena and conversely.

In brief, the procedures of constructors and translators are designed to provide the overall system in which DIAGRAM participates with the capability for context-sensitive compositional semantic interpretations.

2. Attributes and Factors

The attributes that are set and tested by rule procedures may be divided into two types: general and specific. Specific attributes are associated with specific categories of words and phrases. They are often introduced in the lexical entries for words and inherited by the phrases in which the words are constituents. Others are introduced in the rule procedures themselves. One very important attribute, the attribute of syntactic number, is introduced by both means. Syntactic number is primarily an attribute of the category NOUN, and most NOUNs are formed by a rule that optionally combines a noun stem N with the plural suffix "-s." If the suffix is present, the value of the attribute is "plural" (PL); otherwise it is "singular" (SG). However,

some words that do not combine with the suffix are inherently plural; e.g., "people." Such words are entered into the lexicon as members of the category NOUN with the value PL for the attribute @ NBR. The ambiguous word "fish" is also a NOUN, but it has no attribution for syntactic number. The tests for number agreement accept null values (meaning that the attribute is undefined) as agreeing with any value, so that the NP "the fish," which has not inherited an attribute of number from its NOUN constituent, agrees with either a singular or plural auxiliary.

The syntactic categories assigned by DIAGRAM to words and phrases are listed in Appendix A. Appendix B lists the principal syntactic attributes associated with specific categories.

Among the general attributes that are referenced and used in the procedures is one called @ SPELLING. @ SPELLING is an attribute of every word in the lexicon and its value is the literal form of the word. This attribute is referenced in rules that attach particles and inseparable prepositions (Ps) as constituents of the inner core of verb phrases. For example the lexical entry for "give" will list the form UP as a possible value for the attribute @ PARTICLE, and the rules (Rules VP1 and VP2) that analyze the phrases "give up the book" and "give it up" will test for agreement of the @ SPELLING attribute of the preposition "up" with the forms listed as possible particles for the verb "give." The @ SPELLING attribute is also used to check other "function" words and affixes.

Every phrase also has as attributes the categories of its immediate constituents. These may be referenced to any desired depth in the procedures that constrain the application of a rule. For example, noun phrases may contain modifying phrases that follow the central core (NOMHD) of the NP. Such modifying phrases belong to the category NCOMP, and NPs that contain them as immediate constituents will have an attribute @ NCOMP with the value T, meaning that the constituent is present. The NCOMP itself may have a prepositional phrase constituent, in which case it has the attribute @ PP.

DIAGRAM makes frequent use of such general attributes in the subprocedures called factors that assign likelihood scores to analyses. Instead of simply accepting or rejecting an analysis, a rule may accept it with some assessment of the probability of its correctness. For example, the probability that a prepositional phrase is part of a nominal modifier NCOMP rather than a verbal modifier may depend on what the preposition is (i.e., how it is spelled), and on whether numerous other modifiers have already been added to NCOMP. Factors can specify the probabilities and attach scores to phrases that meet their specifications. Then when several possibilities for attaching the prepositional phrase result in multiple analyses of a sentence, the analyses can be ordered according to their likelihood by combining the judgments of the various factors for the various combinations. Less likely combinations can subsequently be ignored by semantic and pragmatic interpreters.[1]

To give a simple example, consider the use of factors in assessing the probability that a prepositional phrase is part of an NCOMP. A factor in the appropriate rule NCOMP4 stipulates that the probability is low (UNLIKELY) if the NCOMP already has a relative clause constituent, SREL. The designation for the attribute is (@ SREL NCOMP) and its value is T, meaning "true" or "present," for NCOMP modifiers like "that I saw" in:

the man that I saw

In analyzing the larger phrase:

the man that I saw with the telescope

DIAGRAM currently assigns a higher score to the analysis in which "with the telescope" is part of the SREL "that I saw with the telescope," but will allow the alternative analysis in which "with the telescope" independently modifies "man."

It is worth noting that factors can be changed to adjust the analyses to particular styles and discourse domains without rewriting the grammar [J. Robinson 1975].

---

[1] Methods for combining scores of factors and ordering multiple analyses are described in Paxton [1977, pp. 132-138].

III PHRASE CATEGORIES AND CONSTITUENTS

A. Sentences

We begin an account of the grammar itself with an explication of DIAGRAM's characterization of the category S, the category of independent sentences. Sentences, viewed functionally, perform many different tasks. Among other things, they encode propositional content that makes claims about the way the world is; they arrange that content so it meshes with what the speaker assumes the hearer knows and is currently attending to; and they express the speaker's attitudes toward the content and toward the hearer. The same propositional content may be encoded with different words in different syntactic structures. The differences reflect the influence of the other functions the sentence serves when it is used. Consequently, devising a grammar for recognizing and making explicit the significant variations in the syntactic structures of sentences and other phrases is a complicated affair. Decisions as to how some string of words should be structured and categorized in order to simplify the problem of interpreting them in a consistent and general way need to be guided by considerations of their communicative functions. These considerations should be kept in mind as part of the background in which various decisions about the way to analyze English sentences were embodied in DIAGRAM.

B. Sentence Classification

Sentences are traditionally classified, according to their modalities, as imperative and indicative. Indicatives are subclassified as declarative and interrogative; interrogatives are further subclassified into two types according to whether the truth of a total proposition is being questioned or whether some argument of a proposition is marked as unknown. These two types of interrogatives are sometimes called Yes/no (or polar) questions and WH-questions.

This hierarchical concept is not preserved overtly in DIAGRAM. All of the distinctions are treated as equally important and although the special relatedness of WH-questions to Yes/no questions is implicitly recognized in the constituent structure assigned to the WH-questions (see below), DIAGRAM explicitly distinguishes four basic sentence types. These are given the mnemonic category names:

SIMP	Imperative sentence
SDEC	Declarative sentence
SQ	Propositional interrogative
SWHQ	Argument interrogative

The syntactic basis for this treatment is that each type has a distinctive initial category. SIMPs lack an initial NP. They typically begin with an uninflected auxiliary DO or BE or with an uninflected form of a verb. SDECs typically begin with a noun phrase that functions as the subject of the sentence. SQs begin with an auxiliary followed by an NP. SWHQs begin with a noun phrase marked by the presence of a WH-word (how, who, what, which, where, when, whose) in initial position or by an initial prepositional phrase with such a noun phrase as its object.[2] Examples are:

SIMP	Put the apples in the basket. Don't drop them.
SDEC	He put the apples in the basket. They are there now. There are several of them.
SQ	Did he put them in the basket? Are they in the basket or in the sack? Are there more than two there?
SWHQ	How many apples did he put there? Which basket did he put them in? How many of them are there?

The differences in the types of messages each class of sentence is suited to convey are also equally important, conceptually. SIMP represents a directive speech act in which, to use Searle's terms

---

[2] WH-questions can also include sentences with noninitial WH-marked phrases, as in "He put them where?" DIAGRAM does not accept this kind of sentence, but extending it to do so should pose no major problems.

[Searle 1977], the direction of fit of the propositional content is from word to world; that is, the speaker asks the hearer to perform some action (putting apples in a basket) that will bring about a state of affairs in the world that realizes (makes true) the propositional content expressed in the words (the hearer has "put the apples in the basket"). For SDEC, the direction of fit is from world to word; that is, the speaker asserts that the propositional content expressed in the words is true ("they are there now"). SQ is a request for the hearer to judge whether the words fit the world and tell the speaker if the propositional content is true (if "he" did indeed put "them" in the basket). SWHQ presupposes the truth of some propositional content but requests the hearer to supply a missing argument or quantifier. Extending Searle's terminology somewhat, we can say that WH-interrogatives (SWHQs) presuppose speaker and hearer agree that the words of the propositional content fit the world--the apples are in some basket--but that the speaker is missing a piece of the content that the hearer is requested to supply ("which basket").

Although the four sentence types are distinguished by their initial constituents, all of them are analyzable in terms of three basic functions: a subject, a predicate, and an indicator of mood. In most sentences, these functions are served by overtly present constituents. For example:

Subject	Mood Indicator	Predicate
They	were	going there
That	is	a big bird
Your home	is	very beautiful
The cake	must be	in the oven

Under special circumstances, there may be no constituent to represent the mood; for example, in "they go here," where the sentence is affirmative, not interrogative, the predicate is a verb, the subject is plural, and there is no auxiliary to indicate perfective or progressive aspect. And of course, imperatives lack subjects. Nevertheless, the association of constituent phrases with these functions is strong. In general, subjects are NPs, mood is indicated by various positions of

auxiliary phrases, and predicates are phrases headed by one of the four major parts of speech--VPs, NPs, ADJPs or PPs.

Transformational grammars capture these generalizations by setting up canonical underlying forms of sentences in which each function is represented by an appropriate constituent category, arranged in declarative order. If we were writing such a grammar, we might represent the rules for S in a single schema like:

S = NP AUX {VP / NP / ADJP / PP}

and subsequently derive our four subtypes from the canonical forms by moving and deleting constituents to form interrogatives and imperatives. But although it employs some local string transformations on the sentences presented to it for parsing, DIAGRAM represents the various sentence types separately. Imperatives, for instance, are not derived from some appropriately restricted canonical form by deleting an initial NP realized as "you". They are analyzed directly by a rule (see Appendix A, Rule SIMP1). Some relevant facts for interpreting imperatives are: that the subject is not overtly expressed but is understood to be the hearer, that the mood is indicated by the absence of subject, tense and aspect, and that the predicating expression directs the hearer to perform some action to make the world fit the words. These facts are represented in the the rule's translator, which assigns semantic interpretations to the syntactic analysis the rule provides for strings it accepts. This approach achieves the same effect as that achieved by formulating an imperative "transformation" that is constrained to operate on just those "underlying" canonical forms that obey the restrictions on realizing the subject as "you," omitting tense and aspect markers, and confining the predicates to certain semantic subclasses.

Nevertheless, we feel that there is some loss of generality in writing so many separate rules that have so many elements in common, and we are therefore exploring the possibility of deriving some rules from other rules. [See Paxton 1977, pp. 258-260 and Gazdar 1979b.] This approach to achieving generality promises to avoid inefficiencies in the

traditional transformational approach that requires first building and then transforming syntactic tree structures, some of which may have elaborate and redundant substructures that are not only moved around, but are subsequently deleted.

To sum up at this point, then, DIAGRAM does not derive sentences from underlying canonical syntactic forms, but directly defines four basic kinds of sentences. The syntactic bases for classification are the presence or absence of subjects, of tense and aspect auxiliaries, of a specially marked WH-phrase, and their structural arrangements with respect to each other when present. The classification is expressed in the first sentence rule, SO. which defines the root category of independent sentences as any of four types, with or without certain marks of punctuation.[3]

Rule SO S = { SDEC / SIMP / SQ / SWHQ } (ENDPUNCT) ;

C. The Two Interrogatives

Mention has been made of an implicit recognition of the special relatedness of the two types of interrogatives, SQ and SWHQ. The recognition is implied by the analysis of SWHQs as containing SQ constituents in all cases except those in which the questioned argument is the subject of the sentence. The latter cases are handled by a distinct rule.

The rules defining SWHQs are:

SWH1 SWHQ = { WHNP / WHPP / WHADJP } SQ
SWH2 SWHQ = WHNP (AUX) { VP / BE PRED }

Examples are:

(i)	WHNP[Which boat]	SQ[did you buy]	Rule SWH1
(ii)	WHPP[From whom]	SQ[did you buy it]	Rule SWH1
(iii)	WHADJP[How wide]	SQ[is it]	Rule SWH1
(iv)	WHNP[Who]	SQ[did you buy it from]	Rule SWH1
but (v)	Who bought it	(No embedded SQ)	Rule SWH2

---

[3] Braces enclose a list of alternatives, separated by "/". Parentheses enclose (a list of) optional elements.

One might expect, from the foregoing explanation and examples, that "what is that thing" would be analyzed by Rule SWH2 rather than by Rule SWH1, on the basis that the initial WHNP "what" is the subject of the sentence. Contrary to such expectations, Rule SWH1 does analyze such sentences and the reason is that "what" is not the subject but the predicate. Just as "how wide" is the predicating adjective phrase and "it" is the subject in "how wide is it," so "what" is the predicating nominal phrase and "that thing" is the subject in "what is that thing." As an informal demonstration of the validity of this claim, consider the possible order of constituents in the declaratives that could answer the questions:

How wide is it?
  It is very wide. *Very wide it is.

What is that thing.
  That thing is a wrench. *A wrench is that thing.

What is a wrench.
  A wrench is a tool. *A tool is a wrench.

The order of the declarative answer to SWHQs with two nominal constituents reveals that it is the second, non-WH nominal that functions as the subject.

When the SWHQ contains an embedded SQ, the initial WH-marked constituent also "stands for" (in some sense) a constituent that is lacking in the SQ. In other words, an embedded SQ is always semantically incomplete or underspecified. When the missing constituent is an obligatory object of a verb or preposition, the SQ is syntactically incomplete too; that is, it is not a well-formed independent sentence. Such syntactically incomplete SQs are exemplified in the SWHQ examples above by "did you buy" and "did you buy it from."

This type of incompleteness is also present in relative clauses, and a discussion of its theoretical and practical implications for the analysis and interpretation of embedded SQs is deferred to the treatment of relatives and WH questions in Section III I.

D. Sentence Types and the Packaging of Information

Since the sentences analyzed by DIAGRAM are assumed to be utterances in a dialogue, it may be of interest to consider briefly how the information structures of sentences are related to their syntactic analysis. The preceding description of WH-interrogatives provides some illustrative examples.

The SQs that are embedded in SWHQs by Rule SWH1 contain the propositional material that is presupposed; that is, the information that is ostensibly "given" or "already activated" [Chafe 1976], and shared by both speaker and hearer. For SWHQs analyzed by Rule SWH2, the presupposed content is in the auxiliary and predicating constituents. Thus, "which boat did you buy" presupposes "you bought a boat," and "who bought it" presupposes "someone bought it." In both cases, the initial WH-marked constituent contains a pointer to the kind of information the speaker lacks and assumes that the hearer can supply. In declarative sentences, on the other hand, the initial constituent generally points to or contains already activated, shared information and what follows is likely to add information that the speaker assumes is not already activated and shared by the hearer. The most striking exception to this generalization occurs when the declarative is an answer to a WH-interrogative in which the subject is the WH-marked constituent. For example, in the first of the following exchanges, the declarative answer to the question packages the new information in the subject constituent, "John," which is likely to be stressed, and the already activated information is in the predicate that follows.

Who bought the boat?
John bought it.

What did John buy?
He bought a boat.

This is just what we would expect from the way the preceding question serves to reverse the order in which given and new information is packaged. In the second example, the declarative answer contains a constituent in its predicate that is lacking in the predicate of embedded SQ of the SWHQ, a constituent that also contains the kind of

information indicated by the WH-marked initial constituent of the interrogative. As we might expect, therefore, the predicate of the declarative answer contains the new information, which is also likely to be stressed in speaking.

These considerations of how information is packaged become important in designing computer systems for interacting with human users. For instance, it is important that mistaken presuppositions be corrected and that new information be appropriately related to already activated information. Otherwise, there is risk of misunderstanding or incoherence. The risk cannot be avoided if only the propositional content of sentences is considered; the syntactic packaging must be related to the propositional content in analyzable ways.[4]

E. Subtypes of the Basic Types of Sentences

1. The BE/DO Dichotomy

Each of the four basic sentence types is divided into two classes according to whether or not an auxiliary BE is required. Only sentences with verb phrase predicates can occur without any member of the BE paradigm. Such sentences require a form of DO if they are interrogatives or if they are negated or emphatic; otherwise, they may omit an auxiliary constituent entirely, except for the suffix "-s" marking present tense in sentences with third person singular subjects. Examples of the first class are "it is here" and "it could be a bear." Examples of the second class are "they go," "they didn't go," and "he goes."

The rule schema for imperatives, shown below in simplified form, reveals the differences in the two subtypes. Rule SIMP1 allows an imperative with a VP predicate to occur without any auxiliary. When an auxiliary is present, it is the infinitive form of DO. An imperative with a non-VP predicate requires the infinitive form BE, but occurs with an infinitive DO in emphatic and negated forms.[5]

---

4 For a fuller discussion, see Hobbs and Robinson 1979.

Rule SIMP1 SIMP = ("DO" (NOT)) { "BE" PRED / VP }

Examples: Don't be an idiot Don't try
Do be careful Do try
Be here by noon Try again

Only a relatively small number of non-VP predicates are semantically appropriate as imperatives. Intuitively, the DO auxiliary and verbs are both associated semantically with actions, while BE is associated semantically with states, as are adjectives, nouns (static objects) and prepositions. Only those states that are conceived of as requiring an act of will to maintain can appropriately be expressed with non-VP predicates in imperatives.

Except for imperatives, sentences that require a BE form in the auxiliary cannot also occur with a form of DO in the auxiliary. For example, *"He doesn't be an idiot" is not well-formed in modern English. DIAGRAM emphasizes the distinction by defining separate rules for the two subtypes of SDEC, although they could be represented in a single rule schema with some complication of the constructor and translator procedures. The two constituent-structure statements are:

Rule SD1 SDEC = NP (AUXD) VP

Rule SB1 SDEC = NP (AUX) BE PRED

2. Existential Sentences

Sentences requiring a form of BE in the auxiliary have a special subtype that is also distinguished in DIAGRAM. These are the existential sentences like "there is someone here," "are there any more oranges in that basket," and "how many are there." The rules defining existential declaratives and interrogatives form a separate set, in

---

[5] Quotation marks around an item mean that the item is interpreted literally; otherwise, it is interpreted as the name of a syntactic category. Some categories, (e.g., NOT, OF) may contain only one or two members, however. DO and BE are categories of the auxiliary forms, including inflected forms such as "did" and "is". Instead of quoting "DO" and "BE," it would be possible to name the categories and then reject all inflected forms by performing a test in the translator for the rule, but this would clearly be inefficient.

which "there" is explicitly quoted as a constituent, thus:[6]

```
Rule STHERE1
      SDEC = "THERE" (AUX) BE (NP ({ING {VP / "BE" PRED} / PRED}))
             Examples:  There could be two.
                        There is a man being held prisoner.
                        There is a man running away.
                        There are some oranges on the table.

  Rule STHEREQ1
        SQ = BE "THERE" (NP ({ING {VP / "BE" PRED} / PRED}))
             Examples:  Are there?
                        Are there any going?
                        Are there any more here?
                        Are there some oranges left?
```

Separating the rules for existential sentences from the others allows much more economical treatment of certain special syntactic constraints. For example, unlike the other declaratives, number agreement in existential declaratives holds between the auxiliary and the NP that follows it, rather than one that precedes it. Compare: "this is a group" / "these are a group" / "there is a group." Also, the NP is typically indefinite, rather than definite and this fact can easily be stated in the constructors for the rules. Moreover, the NP is always a referring expression unless the sentence contains a negation, whereas in typical declaratives, the NP immediately following the auxiliary is a predicating expression. (Compare: "he is a man" vs. "a man is here.")

These special syntactic and semantic facts stem from the special function that existential sentences perform in discourses. Some sentences in a discourse are analyzable as containing both "old" (or "given") information and "new" information [Chafe 1976], but in existential sentences, all the information is presented to the hearer as "new". They are introductory sentences, quite literally, serving to introduce new objects into the domain of the discourse and focus the hearer's attention on them [Grosz 1980, Grosz and Hendrix 1980]. This is true even when the NP is definite. For example, "There's the book by Chomsky" (with the existential reading for "there") is appropriate only

---

[6] Note the use of "({" and "})" to enclose choices that include the null choice; i.e., the constituent is not required.

if there has been some previous denial or questioning of the existence of certain kinds of objects, as in "there aren't any good books on syntax." In this case, the previous speaker is reminded rather than introduced to the object (Chomsky's book), but in any case, that object is newly introduced into the discourse.

It is fairly common to overlook the special discourse functions of sentences when devising grammars, and appeals are frequently made to the principle that one should try to "capture" generalizations, when arguing for one syntactic analysis over another. On such grounds, many linguists have claimed that existential "there" is to be categorized as NP and existential sentences are to be defined by the same constituent-structure rules that define the other sentence types. However, I know of only one property of the existential "there" that it shares with NPs, and that is its ability to function as the subject of a sentence.[7] But to say that "there" is the subject is to describe its syntactic function rather than its syntactic category. To claim that it is a NP is to ignore the fact that it cannot appear as a predicate nominal nor as the object of a preposition. If it is categorized as NP, this means that every occurrence of "there" in a non-subject position will have to be checked for in the rules for phrases with NP constituents. Otherwise, semantic filters must reject inappropriate analyses that will be produced, for example, by parsing sentences like "John is there" with the same rules that parse "John is a linguist."

In arguing against syntactic overgeneralization, I am not claiming that the current rules for generating existential sentences are sufficiently general. DIAGRAM does not provide analyses for certain embeddings of existentials, such as "there appear to be many oranges on that table," although it handles "it appears that there are many oranges on that table" and "many oranges appear to be on that table."

---

[7] That it does have this property is shown by its behavior in tags, where "there are some men here, aren't there" is comparable to "John is here, isn't he." This syntactic behavior is plausibly described by saying that the subject of the sentence is reduplicated in the tag.

F. Sentence Conjunction and Modification

Larger sentences can be created by conjoining sentences or by modifying them.

1. Conjunction

DIAGRAM's rules for conjoining should be taken primarily as experimental placeholders, awaiting the time when linguistic theory yields more insight into the nature of conjunction and the constraints to be obeyed. Currently, Rules SX1 and SX2 define sentence conjunction and accept sentences like:

I went there but he wasn't there. [Rule SX1]
He came, he saw, and he conquered. [Rule SX1 followed by Rule SX2]

These rules do not cover sentences like "He came, he saw, he conquered." Also, DIAGRAM does not have rules capable of accepting sentences involving "gapping", like

John gave but would have preferred to sell Mary a book

where the sequence "Mary a book" is not a single constituent, but rather two constituents, serving as indirect and direct objects of both "gave" and "sell". However, the main problem with DIAGRAM's conjoining rules is not their failure to analyze some legitimate sentences; it is the problem that all syntax-based conjoining rules are prone to introduce. They assign too many structures. Attempts to constrain such rules syntactically will usually bar them from recognizing some legitimate structures.

The necessary constraints undoubtedly involve some notion of semantic parallelism. Meanwhile, DIAGRAM's conjoining rules for sentences are loosely constrained not to conjoin imperatives with non-imperatives. The constraint, which appears in the constructors, rejects the parsing of sentences like

I like to jog, swim and play tennis

as the conjunction of a declarative "I like to jog" with an imperative interpretation of "swim and play tennis." However, it also rejects one possible correct parsing of the ambiguous sentence:

I like her but don't tell her that.

2. Modification

Sentence modification poses less severe problems. Rules SX3 and SX4 analyze sentences like

At midnight, satisfied, he left.
[Rule SX4 followed by Rule SX3]
After they arrived, he left.
[Rule SX3]

However, modifiers that follow the sentences they modify are attached as parts of the predicates of the four basic sentence types, rather than as modifiers of the S that dominates them. That is, in

He left after they arrived

the phrase "after they arrived" is analyzed as part of the verb phrase. (See Rule VP6.) This attachment is theoretically justifiable, if one considers that Rules SX3 and SX4 define structures in which the modifiers have been fronted from a normally following position, analogously to the fronted object noun phrase "that man" in "that man I don"t like." Even so, the rules for sentence modification will not at present allow a sentence modifier to modify jointly all sentence constituents of a conjoined sentence. For example, in parsing

He arrived and she left at two o"clock

the modifier "at two o"clock" is attached only to the second SDEC of the conjoined pair. One remedy would be to add the rule

S = S (",") PP

which is symmetrical to Rule SX3, but to do so would increase the already large number of parsings due to trailing prepositional phrases. For this reason, adding such a rule has been postponed, pending the development of semantic routines that can be called to check the compatibility of potential modifiers with the heads of the phrases for which they are proposed as constituents.

G. Sentence Embedding

Languages achieve great complexity of expression in large part through providing for the embedding of sentences inside other sentences. In DIAGRAM, the root category of independent sentences, S, may be conjoined with other Ss or it may be modified. When it is conjoined or modified, it is embedded in a higher S, as in the structures S[PP S] and S[S CONJ S]. This self-embedding differs, however, from the embeddings in which sentences are constituents of phrases of a different, nonsentential category. In DIAGRAM, the root category, S, is embedded only as an immediate constituent of a higher S, and SQ, as we have seen, is an immediate constituent of some SWHQs. The only sentence type to be embedded as an immediate constituent of a nonsentential phrase category is the declarative type, SDEC.

Various rules define phrases in which SDEC is a constituent. It occurs as a complement of a verb phrase, as an object of a preposition, as a constituent in the complement of a comparative phrase and as a constituent of a relative clause, as in:

	[SDEC]	Embedding Rule
(i)	I saw (that) SDEC[he had arrived].	Rule VP3
(ii)	After SDEC[he had arrived], they left.	Rule PP1
(iii)	It is wider than SDEC[it is high].	Rule THANCOMP
(iv)	He is as tall as SDEC[she is].	Rule ASCOMP
(v)	I saw the boat SDEC[you bought].	Rule SREL1

Recall that SWHQs cannot be embedded except as immediate constituents of the root category S. Therefore, the expression "who just came in," in "I recognize the person who just came in" is analyzed as a distinct relative clause type and not as an embedding of SWHQ, although it would be so analyzed in isolation. (Compare Rules SREL1 and SREL2 in Appendix A.) Like the sentence types SIMP and SQ, SWHQ is in general disqualified on both syntactic and semantic grounds from being a constituent of relative clauses. That is, *"the man go home" (with "go home" as SIMP), *"the man did you see him," *"the man to whom did you give it," and *"the man which man came" are all ill-formed. Only those SWHQs that have "who" or "which" as subjects have the right shape, so to

speak, to be relative clauses, as in "who just came in" or "(the thing) which fell." But this seems as fortuitous as the fact that the string "open the door" could be analyzed as an imperative and then combined with the plural NP, "they," to form the declarative "they open the door." It seems more general therefore, to exclude all SWHQs from relative clause constituency and account for those SRELs in which the subject of the embedded clause is "who" or "which" by means of a separate rule, SREL2.

H. Nonfinite Phrases

In addition to the SDEC sentence type that can occur as an embedded constituent of a matrix sentence, certain sentence-like phrases also occur as constituents of complex expressions. These include infinitive phrases and gerunds, like:

for John to have been going there yesterday
and John's having been going there yesterday

which resemble embedded declaratives in having subjects, predicates and aspectual auxiliaries, but which do not contain any finite (tensed or modal) auxiliary or verb forms. They may also lack subjects, as in "to have gone there." One way of recognizing the similarities while acknowledging the differences is to say that infinitives and gerunds, like some embedded declaratives, may have the same propositional content as an independent sentence but that a propositional commitment is lacking. Compare, for example, the difference between the nonfinite phrases exemplified above with "John could have been going there yesterday," uttered as an independent sentence. Both the gerund and the infinitive can be embedded in contexts that relate them to reality in different ways, as in:

For John to have gone there was impossible.
John's going there was impossible.

For John to have gone there was regrettable.
John's going there was regrettable.

The possibility of John's going is denied in the first two examples; in the second two, its occurrence is presupposed. Embedded declaratives behave similarly, as shown by

It is impossible that John could have gone there.
It is surprising that John could have gone there.

I. Relatives and WH-Questions

The SDECs embedded in relative clauses and the SQs embedded in SWHQs have a common feature. They have "holes" in them; that is, some normally present constituent is missing. In

I saw the boat you bought

the SDEC "you bought" is incomplete; the transitive verb "bought" lacks an object. The missing object is supplied by a constituent outside the SDEC; namely, by the nominal "the boat," which is a constituent of the NP in which the SDEC is embedded. Similarly, the SQ "did you buy it from" in "who did you buy it from," lacks the NP object of the preposition "from," which is supplied by the WHNP of the embedding matrix sentence. Holes occur in other phrasal categories as well. The problem of specifying where holes may occur and where the constituent may be found that provides the missing content is complicated by the fact that holes may occur at arbitrary depths and distances from the constituents that fill the semantic gaps. In addition to the examples above, in which the holes are in the next level of embedding, we also have

I saw the boat that Mary told me John claimed it was not likely that you were going to buy ( ).

and

Who(m) did they think it was possible for you to manage to buy it from ( ) if you needed it?

In other words, a phrase with a hole in it may depend on an arbitrarily distant constituent for its syntactic and semantic well-formedness.

The constituent-structure statements in DIAGRAM currently allow many constituents to be optional even in contexts where they may in fact be obligatory. The semantic interpretation procedures in the constructors and translators have the task of locating constituents that can fill holes in the interpretations of incomplete phrases or--in case of failure to find such constituents--to reject the analysis. This works tolerably well, but throws unwarranted burdens on semantic

analysis. A current project is to adapt the formalism in which DIAGRAM's rules are written so that it is easy to define and apply something like Gazdar's linking rules or a hold mechanism. This will allow the rules to distinguish between constituents that are always optional and those that can be omitted only in case a linking rule is applicable to them or a suitable constituent is contained in the hold register. Some constructions, however, will probably be handled by special mechanisms in the parser. Most likely, they will be those involving conjunction. Meanwhile, many of the restrictions found in the constructors of the present rules are motivated by the need to reduce the burden on semantics and eliminate unwarranted parsings by syntactic means before trying to interpret the results.

IV CONSTITUENTS OF SENTENCES

There are three important decisions that a grammarian has to make in proposing rules for analyzing sentences. They are: (1) what the ultimate unanalyzed constituents, the basic elements, are; (2) how the elements in the sentence are grouped into phrases (where the boundaries are): and (3) what the categories are. The decisions are not independent of one another, and all are influenced by knowledge of the grammatical and communicative functions that are conceptually associated with a phrase of a given category. In this section, some of the problems and issues that arose in making these decisions in DIAGRAM are explored, focusing on the relationship of categories to the syntactic functions of predication and modification.

A. Words and Basic Elements

It is customary to think of sentences as ultimately decomposed into words (i.e., forms that when written are delimited by blank spaces or marks of punctuation) and of phrases as composed of sequences of words. It is also customary to think of words as appearing only in a lexicon, where they are assigned to word-class categories, and of phrases as built out of word-class and phrase categories by rules that contain only categories and not actual word forms. These concepts are valid only as generalizations and none of them holds in a precise way in DIAGRAM.

Some categories (e.g., N, V, ADJ, P) appear in the lexicon, are never defined by rules, and consist mostly of single words (technically "word stems"). These are sometimes referred to as "lexical" or "word" categories. But some words (e.g., "of") and parts of words (e.g., the plural suffix "-s") appear in rules. Word sequences like "out of" can appear in the lexicon and be assigned to the same lexical category as that of a single word like "from". A single word may be assigned by the

lexicon to a category that is also defined by the rules. For example, "John" is assigned to the category of noun phrases (NP). In addition, an idiom like "kick the bucket" may appear as a multi-word lexical entry, assigned to the category VP that is also defined by rule. However, any category in DIAGRAM that is designated as XP, where X ranges over various word categories, is a "phrasal" category, and there will be one or more rules defining it. Some reasons for these treatments of words, affixes, and phrases should become clear from the following description of the analysis of sentence constituents.

The sentences DIAGRAM analyzes have had suffixes stripped from inflected words like "bolts," "bolted," "running," and "taken". The suffixes are transposed to precede their stems, thus: "-s bolt," "-ed bolt," "-ing run" "-en take". (Reasons for preposing the suffixes will appear shortly.) The uninflected forms appear in the lexicon, where they are assigned to the appropriate categories. It is not necessary, however, to enter "bolts," "bolting" and "bolted," since these words are defined by rule. A single rule statement suffices to combine any member of the class N with a plural suffix, allowing the meaning of the combination to be computed as the meaning of the N plus the meaning of plurality by the semantic interpretation procedures. The lexicon also supplies regular forms for irregularly inflected words like "children," which is given as "child -s" and "took," which is given as "take -ed".

Affix stripping is a common practice in lexicography, and the economies it offers are obvious. DIAGRAM extends the practice to irregular forms. Only inflectional suffixes are stripped. Derivational suffixes, which may effect major changes in the category of the base to which they are attached, are untouched, so that "civil" (ADJ), "civilize" (V), and "civilization"(N), for example, will have individual entries. Derived words not only differ in category from their bases, the meaning of a word composed of a base combined with a derivational suffix cannot be regularly stated as a composite of the meanings of the base and the suffix, whereas the meaning of a base and an inflectional suffix can. A noun stem N plus the plural suffix regularly produces a

NOUN meaning "more than one N," whereas the relationship of the meaning of "civilization" to the base "civil" is less straightforward.

The participle suffix PPL is a borderline case. It is regularly suffixed to verb stems preceded by the auxiliary HAVE, to mark the perfective aspect as in "they have broken it." In this use, it is inflectional. Suffixed to transitive Vs, however, it can arguably be said to change the category of the stem from V to ADJ. Thus "broken," unlike "break," can modify a noun ("the broken vase," but not *"the break vase"), and can itself be modified by words like "very" and "completely" and the prefix "un-," which typically associate with adjectives. Compare: "completely broken" and "unbroken" with "completely happy" and "unhappy". Moreover, as a predicate, "broken" requires a form of BE in the auxiliary and does not occur with a form of "do". Compare: "It is broken" and "is it broken" with "they break it" and "do they break it," as well as with the intransitive sense of "break" in "it breaks" and "it did break."

Much of what has just been said of the suffix PPL applies also to the suffix ING, with added complications. Parallel to "the broken vase," there is "the breaking vase," though not *"the completely breaking vase." In addition, Vs suffixed with ING occupy syntactic positions and serve syntactic functions associated with NPs, as in "the singing was excellent but the dancing was poor." In spite of the shifts in syntactic function, and possibly in syntactic category, however, the meanings of the stem Vs appears to remain constant. It is this constancy of meaning that DIAGRAM seeks to preserve by stripping the two participial suffixes from verb stems.

B. Suffixes and the Auxiliary Constituent

The decomposition of words with inflectional suffixes has consequences for DIAGRAM's definitions of the immediate constituents of phrases. One reason for preposing inflectional suffixes is to regularize the form of the whole phrase in which the stem is an immediate constituent and which has a constant interpretation. Consider the sentences

	NP	AUX	VP
He gives her a book.	[He	-s	give her a book]
He is giving her a book.	[He	is -ing	give her a book]
He has given her a book.	[He	-s have -en	give her a book]
He gave her a book.	[He	-ed	give her a book]
He may give her a book.	[He	may	give her a book]

The VP "give her a book" remains constant in DIAGRAM's analyses of the three different sentences. The difference in meaning is ascribed to the differences in the auxiliary constituents.

Notice that in this analysis, the entire auxiliary, including tense and aspectual affixes, is an immediate constituent of the sentence rather than of the verb phrase. Functionally, the sentence is represented as composed of three constituents, with the subject NP and the predicate VP representing the timeless or neutral propositional content, while the auxiliary represents the speaker's attitude or perspective. It indicates whether the propositional content is considered to hold in the past or the present or in some possible world and whether it is considered to be ongoing or completed.

DIAGRAM's treatment of the form and meaning of auxiliaries and of their syntactic place in the sentence resembles that of Langacker [1978], although independently conceived. Both treatments differ strongly from those in which suffixes are not stripped or preposed and in which modals, HAVE, and BE are categorized as verbs that accept VPs as complements. [Cf. Jackendoff 1977b and Gazdar et al. forthcoming.] Whether or not DIAGRAM's current treatment of auxiliaries and its use of "affix hopping" can be claimed to represent some kind of psychological reality or to result in computational efficiencies in processing dialogue remains moot.[8]

---

[8] It should be noted also that the advantages of simplicity of statement afforded by this affix "hopping" are more easily achieved for some affixes than for others. Transposing the plural suffix to precede a noun does not prefix it to the entire phrase headed by the noun, and does not lead to the simplifying statements possible for the analysis of VP constituents. Although nouns and verbs are often homographic in English, it may still be practical to strip the homographic suffix "-s" from the category V and not from the category N.

C. The Active-Passive Relationship

DIAGRAM's analysis of suffixes and V stems also affects its analysis of the regularities in form and meaning that appear in active-passive sentences. DIAGRAM's input for two passives sentences with "give" as predicate are:

She is given a book.	[she	is	-en	give a book]
A book is given her.	[a book	is	-en	give her]

In the course of analyzing these sentences, "give a book" and "give her" will be parsed as VPs by the same rule, Rule VP1, that applies to the active versions in the preceding examples. The constituent structures to which the rule applies include: a single V, a V followed by an object NP, and a V followed by two object NPs. Only those Vs whose attributes qualify them to take two object NPs will be recognized as predicates of well-formed active declaratives like "he gives her a book." As previously mentioned, DIAGRAM accepts such Vs as sole constituents of VPs when they occur without an obligatory object, to permit recognition of sentences like "what did he give" or "what did he give her." When a verb like "give" is parsed with its two object NPs to form a VP, the constructor procedure of Rule VP1 simply checks for the appropriate attributes, named DIROBJ and INDIROBJ whose values are either T for "present" or null. Here we are interested in the cases in which "give" occurs with only one or with none of its object NPs. In those cases, the procedure marks the resulting VP as lacking its full complement by setting the attribute DIROBJ of the V as an attribute of the VP. (In other words, the VP inherits the attribute of its "head" V.) Subsequently, a rule PRED1 applies when such a VP is preceded by the participle PPL. The relevant parts of Rule PRED1 state the possible constituents of PRED and the conditions on well-formedness, thus:

PRED1 PRED = { ADJP / NP / PP / PPL VP } ;
Constructor: (Informally) If VP is a constituent of PRED, then VP must have the attribute DIROBJ and cannot have a prepositional phrase (PP) constituent.

In this manner, the passive construction joins the categories of NP, ADJP, and PP as a member of the inclusive category PRED, which is

essentially "stative" in contrast to the predicates formed with active verbs.[9] However, it is easy to distinguish subcategories of PREDs on the basis of the categories of its constituents because DIAGRAM's formalism assigns, to each phrase analyzed by a rule, a set of attributes whose names are the categories of its immediate constituents. A PRED formed with PPL and VP will have the attributes PPL and VP that can be tested for when their presence or absence is relevant.

D. The Major Categories

The phrases that are the major constituents of sentences are extensions of the four word (i.e., word stem) categories, noun (N), verb (V), adjective (ADJ), and preposition (P). These categories contain most of the words in a dictionary. In particular, they include the so-called content words--words that, in some intuitive sense, denote entities, actions, processes, qualities, and spatial and temporal properties. [Cf. Lyons 1966.]

With the possible exception of the prepositions, these are large classes that can easily grow when the introduction of new objects and activities into a culture creates the need for new words for talking about them. They contrast with "function" words, whose semantic content is very difficult to specify. The word "of" is an example. Although customarily classed as a preposition, "of" is strikingly different from words like "in," "over," "with," etc. In effect, it is treated as an inflectional prefix of the NP rather than as a preposition. The infinitival "to," as in "he wants to go" is another example. In contrast to the directional preposition "to" in "he came to the office," it carries no semantic information. "Of" and infinitival "to" are not

---

[9] Another rule, Rule PRED2, analyzes prepositional phrase attachments to PREDs. The whole question of the treatment of the active-passive relationship raises many issues beyond the scope of this paper, issues that are also being vigorously raised in revisions of the theory of transformational grammar; e.g., by Bresnan [1978] and Wasow [1978]. We are examining them and exploring alternative ways of associating information about logical forms or predicate-argument structures with variations in syntactic arrangements of subjects, objects, and prepositional phrase complements.

assigned to categories in DIAGRAM; like some affixes, they are cited directly in the rules.

As their names indicate, the categories NP, VP, ADJP, and PP are regarded as phrasal extensions of the categories of the content words. Intuitively, each XP contains a word of category X that is the nucleus or head of the phrase.[10] In:

those very young children requiring an escort

the head is clearly the noun "children". Most of the other phrases within the NP are said to "depend on" or "modify" the head. These latter terms are not well defined and not matched in any simple way with syntactic categories. "Very young" (ADJP) and "requiring an escort" (ING VP) are modifiers, limiting the range of denotation of the head noun. In general, there is a tendency to call modifiers that follow a head "complements". However, the following modifiers or complements of VPs and PPs, which are typically NPs, are usually called "objects".

Function words are seldom classed with modifiers. They are variously called "specifiers", "determiners", or "complementizers". For example, "these" and "those" are not modifiers like "young," but belong to a small, closed class of function words labelled DDET in DIAGRAM. Elliptical headless NPs like "those requiring an escort" are possible when a function word specific to NPs is an initial constituent. Their full interpretation, of course, requires discourse context.

The foregoing description should provide a basis for understanding the interactions of the rules with the lexicon and with each other. Explication of a specific rule will further illustrate the principles involved.

---

[10] Cf. discussions of the X-Bar theory in the theory of transformational grammar in J. Robinson 1970 and Jackendoff 1977 and references therein.

E. A Rule for the Analysis of NP

The rule given below is a simplified version of a rule appearing in Appendix A. It analyzes noun phrases like: "water", "boys", "the boy", "younger children than that", "those very young children requiring an escort", "those boys whom I had expected to see", "every school boy on our block who has been driven to school by his parents". It will reject, as ill-formed: "many water," "much boy," "a boys," "two boys than that," and "those children who has been driven to school". It will mark as UNLIKELY an NP consisting solely of a singular count noun like: "boy," "bolt " etc., as in "I saw boy," but allow it, in case no better analysis can be found. The judgments of likelihood are partly, perhaps even largely, intuitive and subject to alteration in different discourse environments; for example, in an environment where the participants use a "telegraphic" style and omit function words like "the".

```
(NP1      NP = (D={A / DDET / DETQ}) NOMHD (NCOMP) ;

    CONSTRUCTOR (PROGN (COND
                       ((@ D)
                  1.     [COND
                           ((MASS? D)
                              (OR (MASS? NOMHD)
                                   (F.REJECT (QUOTE F.MASS]
                  2.     [COND
                           ((MASS? NOMHD)
                            (OR (NOT (@ A))
                                (F.REJECT (QUOTE F.MASS]
                  3.     [COND
                           ((@ NCOMP)
                            (@SET NBR(@INTERSECT NBR D NOMHD NCOMP)))
                            (T (@SET NBR (@INTERSECT NBR D NOMHD]
                  4.      ((AND (SG? NOMHD)
                                (NOT (MASS? NOMHD)))
                            (@FACTOR (QUOTE F.NODET)
                                     UNLIKELY))
                  5.      ((@ NCOMP)
                             (@SET NBR (@INTERSECT NBR NCOMP NOMHD)))
                  6.       (T (@FROM NOMHD NBR)))
                  7.   [AND (@ THANCOMP NCOMP)
                            (OR (@ THANCOMP NOMHD)
                                (F.REJECT (QUOTE F.THANC]
                  8.   (@FROM NOMHD TYPE)))
```

The first line specifies a possible constituent structure for a well-

formed NP as consisting of a nucleus containing a head noun (NOMHD). The nucleus is optionally preceded by a determiner (D) consisting of the word "a" or a definite determiner (DDET) like "the" or "those" or a determiner of a different type (DETQ) like "much" or "many". It is optionally followed by a complement (NCOMP).

The constructor contains several statements, numbered here for ease of reference, including statements of conditions that must be met for well-formedness. The first two require that any determiner must agree with the NOMHD constituent with respect to the property MASS, thereby rejecting such combinations of a determiner like "much" with a count noun like "child". Statements 3 and 5 require agreement in syntactic number (NBR) of all constituents and also assign the value of the property NBR to the NP on the basis of the values for the constituents. Statement 6 assigns the value of NBR from the NOMHD constituent when it is the sole constituent. Statement 4 rates as UNLIKELY the probability that an NP will lack a determiner if its head noun is a singular count noun like "child". Statement 7 rejects as ill-formed such sequences as "two boys than that," although "taller boys than that" is acceptable. Statement 8 assigns to the analyzed NP the value COUNT or MASS for the attribute TYPE from the value assigned to the NOMHD constituent.

Other NP rules analyze elliptical NPs like "those requiring an escort," comparative NPs like "more tall boys than that" and "as many more tall boys than that as there are girls," and several other types.

V SYNTACTIC CATEGORIES AND SYNTACTIC FUNCTIONS

A. Predication and Modification

The NP rules analyze phrases that can function as subjects of sentences and objects in VPs and PPs. They can also function as predicates; i.e. as members of the category PRED. In:

Children are sleeping

the NP "children" functions as sentential subject. In:

Those boys are children

"children" functions as sentential predicate applied to the subject NP "those boys" in stating that they are members of the set of children. In:

Those boys are the children of my friend

the NP "the children of my friend" is the predicate applied to "those boys" in stating that they are identical to the set of children of a friend of the speaker.

We have already shown how DIAGRAM relates the major phrasal categories to the function of predication, making an important distinction between VP predication and predication by other categories. Because VPs can function as predicates without any accompanying auxiliary constituent, the term "predicate" is sometimes taken to be commensurate with "verb". This usage mixes two levels of description of sentences, the categorial and the functional. Moreover, within the categorial level, it obscures the distinction between lexical heads of phrases (V) and phrases (VP). One consequence of such usage appears in the proposal that prepositions may be indistinguishable from verbs at some abstract level of description [Becker and Arms 1969]. On similar grounds, Ross [1969] has argued that adjectives, verbs, and nouns have deep similarities that outweigh their superficial differences. The mixed properties of passive constructions heighten the confusion.

The same kind of confusion shows up with respect to "modifier", which tends to be equated with "adjective" and applied to any word that modifies a noun in prenominal position. However, three of the four major word categories appear in the function of prenominal modifier, as in:

a stormy sea	(ADJ)
a storm coat	(N)
a threatening sea	(ING V)

All of the major phrasal categories, including prepositional phrases, occur in postnominal modifiers or complements, as in:

a woman of courage	(OF NP)
a book on the table	(PP)
the cat lying on the mat	(VP)
diamonds as big as your thumb	(ADJP)

B. The Shifting of Categories

The pervasiveness of the association of particular categories with particular functions and notions ("a noun is the name of a person, place or thing"; "adjectives modify nouns"), is so notable that it cannot be ignored. Tesniere's notion of translation affords an interesting way of thinking about the interactions of syntactic categories with various functions of the kinds exemplified above.[11] Both function words and derivational affixes are considered to be translatives that serve to shift ("translate") a governor from one category to another, whenever it and the group of words it governs serve one of the functions of the category to which it is shifted. At first glance, these words and affixes would appear to be like the derivational suffixes that change the categories of word bases, previously exemplified in "civil,"

---

[11] Tesniere's syntactic theories, first published posthumously in 1959 [Tesniere 1976, 2nd ed.], center around the concept of the head of a phrase, which he calls a "governor". Governors are content words, categorized (roughly speaking) as nouns, verbs, adjectives, and adverbs. In analyzing a sentence, other words are attached to governors as "dependents" (hence the name "dependency theory"). Each category of governors has a typical set of syntactic functions. Verbs function as predicates, and they alone can occur as independent governors of an entire sentence. Nouns function as subjects and objects, adjectives as modifiers of nouns, and adverbs as modifiers of adjectives and verbs.

"civilize," and "civilization". The similarity is significant, but so are the differences. Translatives include words as well as affixes and they can shift the category of a phrase.[12] In constituent-structure terminology, one would say that translatives are immediate constituents of phrases, even when they form parts of words.

There is a clear example of such a phenomenon in English. In NPs like "the mayor of Boston's hat," the constituent "the mayor of Boston's" is properly analyzed as having two subconstituents, "the mayor of Boston" and "-'s". In effect, the NP "the mayor of Boston" is now a determiner (DDET) of the larger NP of which it is a part. One could say that it had been "translated" from an NP to serve the function of a determiner by combining with the suffix. One of DIAGRAM's rules for analyzing determiner phrases could be so interpreted, namely Rule DDET2, whose constituent-structure statement is:

DDET2 DDET = NP "GEN"

where "GEN" is the lexical replacement for the genitive suffix. Preposed genitives in English are customarily analyzed in this fashion. DIAGRAM's extension of the concept to passive constructions in Rule PRED1 (infra) is not customary, but the extension seems quite natural.

It is relevant to consider how far one might want to push the process of category-changing by translatives in order to capture generalizations about the syntactic distributions and syntactic functions of phrases headed by lexical categories, and what constraints one would want to impose on its use. This is an issue we propose to explore in the future. Specifically, it is interesting to test the consequences of treating the auxiliary BE as a translative for NP, ADJP, and PP predicates, converting them to VP. This would seem to be a reasonable alternative to a more customary concept of copular BE as a "main verb". At present, BE is always an auxiliary, but it is not a translative.

---

[12] Although he often speaks as if only the governor were involved, Tesniere is specific on this point: "C'est le mot ou le groupe de mots resultant de la translation." [Op. cit. , p. 367]

VI DIAGRAM IN THE FUTURE

In its present form, DIAGRAM will be extended in two ways. It will make greater use of the facility provided by constructors and translators for setting contextual constraints on the basis of vertical as well as horizontal context; that is, in terms of what dominates a string as well as what composes it. It will also make greater use of factors to indicate the likelihood that an analysis provided by a rule will ultimately prove to be successful when larger context is available. In addition to providing predictions as to whether a particular analysis will succeed or fail in the long run, factors are even now being used to order multiple parsings, when they occur, so that the less likely ones need not be processed for semantic and pragmatic well-formedness. However, this use of factors is now only rudimentary. Their development will become more urgent as the coverage of the grammar is expanded.

In expanding the coverage, use will be made of syntactic redundancy rules to capture generalizations and reduce the number of statements in the rule procedures. As mentioned in Section III B on sentence classification, there is a loss in efficiency along with a loss in generalization when separate rules are written that require duplicating much of the constructor and translator procedures. The duplication of tests for number agreement of subject NP and AUX or VP is a case in point. A single redundancy rule modifying all of the S-rules to add agreement tests could capture a significant generalization about English. The current metalanguage in which DIAGRAM is written already provides functions for such an operation on rule and category definitions before compilation [See Paxton 1977, especially pp. 256-268.] The functions can also be used to derive new constituent-structure rules from a set of basic rules. One program for deriving constituent-structure rules from other rules has already been implemented by Konolige and its results are being studied.

The most promising as well as the most challenging uses of redundancy rules are related to the problem of holes or traces, noted in Section III I. It was pointed out in that section that holes could be arbitrarily far from the constituents that filled them semantically, as in "I saw the boat that Mary told me John claimed it was not likely that you were going to buy ( )."

In a transformational analysis, the holes were originally occupied by constituents that were moved by transformations to an initial position or deleted when they were coreferential with some other constituent. The moved elements, according to the theory, leave a trace in their original positions. The trace is interpreted as a variable that is bound by the moved element or by the coreferential constituent. For example, "the man I saw" is represented with a trace as:

NP[the man SDEC[I saw t]

where t is "bound" by "the man" [Dresher and Hornstein 1979].

Gazdar [1979b] has recently proposed a non-transformational analysis that accounts for the same phenomena. His method is to derive, from the basic rules, a finite set of rules, each dominating one empty node; i.e., a hole. In addition, the grammar will have a set of "linking" rules for eliminating the derived nodes in an analysis by (in effect) "cancelling" them with the appropriate constituent that occurs elsewhere.

A more familiar alternative in computational linguistics and artificial intelligence systems is to use "hold" registers to store constituents that may be needed to complete a phrase with a hole in it. The grammar might then define the syntactic sites at which the contents of a hold register could be "emptied", and a sentence could be accepted as syntactically well-formed only if nothing remains in the register at the end of an analysis. Woods [1970] appears to have been the first to use a hold register for such a purpose in parsing natural language, but the tests for well-formedness were semantic rather than syntactic. Paxton (1977) has proposed a combination of redundancy rules with null rules to derive SWHQs with holes in the appropriate places and a common

procedure to establish the connection between the fronted WH-phrase and a corresponding null phrase, using a hold mechanism. Other alternatives have recently been advanced as well, some of them involving special parsing mechanisms rather than rules.[13]

DIAGRAM is now a large and complex grammar whose analyses are reasonably constrained. It has been employed in systems developed for different domains of use. One is a system designed for questioning a large data base of information about ships--their properties, locations, destinations, etc. [See Konolige 1979.] Another is an "expert" system designed for guiding a user (an "apprentice") through various steps in the repair and maintenance of small appliances. [See A. Robinson 1978.] It is currently being used in yet another expert system for guiding a novice in the use of computer programs. [See Mann 1979.] The grammar has performed well in these various environments and has benefitted by being tested through use.

Extending DIAGRAM will inevitably have a perturbing effect as the new categories and rules interact with the old ones in unforeseen ways. These perturbations and the methods for controlling them will be worth studying for the light they shed on the English language, or more precisely, on a grammarian's intuitions about the English language.

---

[13] Several proposals were discussed at length at a workshop on non-transformational theories of syntax, held at Stanford University, January 1980, sponsored by the Sloan Foundation.

VII REFERENCES

Becker, A.L. and Arms, D.G. 1969. Prepositions as predicates. In Papers from the Fifth Regional Meeting of the Chicago Linguistic Society, April 18-19, 1969. University of Chicago, Chicago, Illinois.

Bresnan, J. 1978. A realistic transformational grammar. In Halle, M., Bresnan, J., and Miller, G.A. (eds.), Linguistic theory and psychological reality. The MIT Press, Cambridge, Mass.

Chafe, W.L. 1976. Givenness, constrastiveness, definiteness, subjects, topics, and point of view. In Li, ed., 1976.

Culicover, P. W., Wasow, T., and Akmajian, A. (eds.) 1977. Formal syntax. Academic Press, New York, 1977.

Dresher, B.E., Hornstein, N. 1979. Trace theory and NP movement rules. Linguistic Inquiry, 10, Winter, 1979.

Gazdar, G. 1979a. Constituent structures. Ms. January 1979.

Gazdar, G. 1979b. English as a context-free language. Ms. April 1979.

Gazdar, G., Pullum, G.K., and Sag, I. (forthcoming) A context-free phrase structure grammar for the English auxiliary system.

Grosz, B. 1980. Focusing and description in natural language dialogues. In Joshi, A. K., et al., (eds.), Elements of Discourse Understanding. Cambridge University Press, Cambridge, England, 1980.

Grosz, B. and Hendrix, G. 1980. A computational perspective on indefinite reference. Technical Note No. 181, SRI International, Menlo Park, California.

Hobbs, J. and Robinson, J. 1979. "Why Ask," in Roy O. Freedle, ed., Discourse Processes, Vol. 2, Number 4, October-December 1979.

Jackendoff, R. 1977a. X-Bar syntax: A study of phrase structure. The MIT Press, Cambridge, Massachusetts.

Jackendoff, R. 1977b. Constraints on phrase structure rules. In Culicover et al. (eds.), pp 249-283.

Konolige, K. 1979. A framework for a portable natural-language interface to large data bases. Technical Note No. 197, SRI International, Menlo Park, California.

Langacker, R.W. 1978. The form and meaning of the English auxiliary. Language, 54:853-882. December 1978.

Li, C.N., ed. 1976. Subject and topic. Academic Press, New York.

Lyons, J. 1966. Towards a "notional" theory of the "parts of speech". Journal of linguistics, 2:209-236. October 1966.

Lyons, J. 1977. Semantics. 2 vols. Cambridge University Press, Cambridge.

Mann, W.C. 1977. Design for dialogue comprehension. Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, August 11-12, 1979. University of California at San Diego, La Jolla, California.

Robinson, A. 1978. Investigating the process of natural-language communication: A status report. Technical Note No. 165, SRI International, Menlo Park, California.

Robinson, J. 1975. A tuneable performance grammar. American Journal of Computational Linguistics, Microfiche 34.

Robinson, J. 1970. Dependency structures and transformational rules. Language, 46:259-285. June 1970.

Ross, J.R. 1969. Adjectives as noun phrases. In Reibel, D.A. and Schane, S.A., eds. Modern studies in English. Prentice Hall, Englewood Cliffs, New Jersey. 1969.

Searle, J.R. 1977. A classification of illocutionary acts. In Rogers, et. al. (eds.) Proceedings of the Texas Conference on Performatives, Presuppositions and Implicatures. Center for Applied Linguistics, Arlington, Va.

Tesniere, L. 1976. Elements de syntaxe structurale. 2nd ed. Editions Klincksieck, Paris, 1976.

Wasow, T. 1977. Transformations and the lexicon. In Culicover et al. (1977).

Wasow, T. 1978. Remarks on processing, constraints, and the lexicon. In Theoretical Issues in Natural Language Processing-2 (TINLAP-2), University of Illinois at Urbana-Champaign, July 25-27, 1978.

Appendix A

WORD AND PHRASE CATEGORIES AND CONSTITUENT STRUCTURE RULES

Appendix A

WORD AND PHRASE CATEGORIES AND CONSTITUENT STRUCTURE RULES

Phrase categories are followed by the constituent structure rules that analyze them. Rule names are distinguished from category names by numerical identifiers.

A — Indefinite Determiners
a, an, no.
(Small class of function words, constituents of NPs.)

ADJ — Adjective Stems
big, difficult, allowable, frequent, careful, quick, great, tall, short, far, near, close, operational, necessary, possible.
(Some, but not all adjectives can be inflected "-er" or combined with "as", "more", "too" in comparisons of degree.)

ADJCOMP — Adjective Complements
for him, to go there (as in: difficult for him to go there; too heavy for me); that he went (as in: possible that he went).

ADJCOMP1 ADJCOMP = (ENOUGH) (FOR NP) (INFINITIVE)
ADJCOMP2 ADJCOMP = "THAT" SDEC

ADJP — Adjective Phrases
too big, more difficult than that, allowable for him to go there, largest, as ready to go as he is, possible to do, necessary that it is done.

ADJP1 ADJP = ({(QPP) QDET / TDET}) ADJ (ADJCOMP)
ADJP2 ADJP = (DETQ) ("NO") ER ("MUCH") ADJP (THANCOMP (ADJCOMP))
ADJP4 ADJP = AS ADJP (ASCOMP)
ADJP6 ADJP = EST ("MUCH") ADJ (ADJCOMP)

ASCOMP — Adjective Complements for Equality Comparisons
as that, as he appears to be.

ASCOMP1 ASCOMP = AS {NP / SDEC}

ADV — Adverbs
frequently, soon, carefully, quickly, often, necessarily, possibly.
(Words formed from adjectives by adding the suffix "-ly" or that occupy the same syntactic positions as words so formed.)

ADVP — Adverb Phrases
as frequently as that, too often, more carefully than John did it, most rapidly, after he came, when he is here, if they come.

```
  ADVP1     ADVP = ((QPP) QDET) ADV
  ADVP2     ADVP = (DETQ) ER ("MUCH") ADV (THANCOMP)
  ADVP4     ADVP = AS ADVP (ASCOMP)
  ADVP6     ADVP = EST (Q) ADV
  ADVP7     ADVP = {P / "IF"} {SDEC / {PPL / ING} VP}

AUX         Auxiliary Phrases
            couldn't have been -ing, has, is, -ed, -s.
            (May consist of a single suffix.)
  AUX1      AUX = (MODALP) (HAVEP PPL) (BEP ING)

AUXD        DO-Type Auxiliary Phrases
            could have -en, did, -ed.
            (See also DO and DOP.)
  AUXD1     AUXD = C={AUX / DOP}

BE          Forms of BE
            be, am, is, are, was, were.

BEP         BE-Auxiliary Phrases
  BEP1      BEP = BE (NOT)

CONJ        Conjunctions
            and, or, but, nor, then.
            (Includes only coordinating conjunctions.)

DET         Definite Determiners
            the, this, those, your, her, their, its.

DDET        Definite Determiner Phrases
            the,  the many, all,  all five, those, this,  these two, the
            next man's.
  DDET1     DDET = ((NOT) "ALL") DET ({NUMBER / QPP})
  DDET2     DDET = NP "GEN"

DETERQ      Compared Determiner Phrases
            more, that much more, too many more.  (See also DETQ.)
  DETERQ1   DETERQ = (DETQ) ER Q

DETQ        Determiner/Quantifier Phrases
            many, much, too many, any, any two, two, no two, some, most.
            (Indefinite determiners,  including indefinite  quantifiers.
            See also categories Q, QDET, QPP.)
  DETQ1     DETQ = (C={TDET / QDET / A}) Q (QPP)
  DETQ2     DETQ = ({"ANY" / "NO"}) NUMBER
  DETQ3     DETQ = ("THE") EST Q

DO          Forms of the DO Auxiliary
            do, does, did.

DOP         DO-Auxiliary Phrases
  DOP1      DOP = ({"-S" / PAST}) (DO (NOT))

ENDPUNCT    End Punctuation
            (Sentence-final marks of punctuation.)

GERUND      Nonfinite Phrases Formed with "-ing"
```

having gone, being informed, not having realized, having been very careful about that.

GERUND1 GERUND = (NOT) ING (HAVE PPL) (BE ING) VP

GERUND2 GERUND = (NOT) ING (HAVE PPL) BE PRED

HAVE Forms of Auxiliary HAVE

HAVEP HAVE-Auxiliary Phrases

HAVEP1 HAVEP = ({"-S" / PAST}) HAVE (NOT)

INFINITIVE Infinitive Phrases

to have gone, to be informed, not to have realized, to have been very careful about that.

INFINITIVE1 INFINITIVE = (NOT) TO (HAVEP PPL) (BEP1 ING) {VP / BEP2 (PRED)}

INFINITREL Nonfinite Relative Clauses

to whom to give it, for you to do, being attached.
(As in: the man to whom to give it; the thing for you to do; the part being attached.)

INFINREL1 INFINITREL = {({FOR NP / P RELPRO}) INFINITIVE / ING ("BE" PPL) VP}

ING Suffix "-ing"

(Sometimes called the "present participle" suffix. Used for progressive aspect and for forming gerunds.)

MIDPUNCT Nonfinal Marks of punctuation

MODAL Modal Auxiliaries

MODALP Modal Auxiliary Phrases

MODALP1 MODALP = MODAL (NOT) (ADV)

N Noun Stems

man, woman, box, foot, inch, water, length, U.S., United States, time, doctor.

NCOMP Noun-Phrase Complements

of tea, on the corner, for you to do, that I saw.
(Modifying phrases that follow a head nominal or NOMHD).

NCOMP1 NCOMP = {OF / P} NP

NCOMP2 NCOMP = {INFINITREL / SREL / ADJP / THANCOMP}

NCOMP3 NCOMP = PPL VP

NCOMP4 NCOMP = NCOMP PP

NHD Pre-Noun Modifier Heads

twenty-one gun, three mile, stone
(As in: twenty-one gun salute; three mile swim; stone house. NHDs are constituents of compound nominals in which a noun stem (N) modifies a noun.)

NHD1 NHD = NUMBER ("-S") N

NHD2 NHD = NOMHD

NOMHD Nominal Heads

very big task, broken vases, running streams, more difficult task, fish, cats, people.

(Constituents of NPs, containing the head noun and pre-nominal modifiers, but not determiners or quantifiers of the NP.)

NOMHD1 NOMHD = NOUN ({(",") "AND" / ","} NOMHD)
NOMHD2 NOMHD = {(QDET) ADJ / {ING / PPL} V} NOMHD
NOMHD3 NOMHD = (QPP) ER ("MUCH") ADJ NOMHD

NOUN Nouns
(N stems with or without a fronted plural affix; e.g., cat, -s cat, -s child (children), fish. May be pre-modified by NHD or NOMHD as in: cat food containers. Unlike Ns, NOUNS have an attribute of syntactic number, SG or PL, unless they are ambiguous with respect to number.)

NOUN1 NOUN = ("-S") N
NOUN2 NOUN = NHD NOUN

NP Noun Phrases
water, cats, very big cats that lie on mats, the length of that board, a length of two feet, as big a box as you could carry, two of them, more ships, a more difficult task than that, what he did, he, her, the best ones, the best I could find, his having gone there yesterday, those, those few, some, all of them that are here, John, Mary, September.
(The most complicated category in the grammar.)

NP1 NP = (D={A / DDET / DETQ}) NOMHD (NCOMP)
NP2 NP = D={A / DDET / DETQ} (NCOMP)
NP3 NP = DETERQ (NOMHD) (NCOMP)
NP4 NP = {WHAT / WHEN / WHERE} SDEC
NP5 NP = PRO
NP6 NP = DDET EST ("MUCH") ADJ (NOMHD) (NCOMP)
NP7 NP = {DET / NP GEN} GERUND
NP8 NP = AS QPP ({OF NP / NOMHD}) (ASCOMP)
NP9 NP = AS ((QPP) ER ("MUCH")) ADJ (OF) NP (THANCOMP) (ASCOMP)

NUMBER Numbers
two, twenty, 2, 20.

PAST Past-Tense Suffix
("-ed" as regularized by lexical entry and suffix stripping; e.g., "took" becomes "take -ed.")

PAST1 PAST = "ED"

P Prepositions, Particles, Subordinating Conjunctions
at, in, on, before, after, by, to, for, with.
(Does not include "of." Ps may occur as intransitives or particles, as in: break up the pavement; or as transitives with NP objects, as in: in the box. Combined with sentences and verb phrases, however, they become "translatives", and the resultant phrase is ADVP.

PP Prepositional Phrases
on it, after that, there (i.e., at that place).

PP1 PP = P (NP)

PP2 PP = "WITHIN" NP1 (OF NP2)
PP3 PP = "BETWEEN" NP1 ("AND" NP2)

PPL Past-Participle Suffixes
PPL1 PPL = {"EN" / "ED"}

PRED Predicate Phrases (not VP)
(NPs as in: that is a boy;
ADJPs as in: that could have been very heavy;
PPs as in: that wasn't in the box when I left;
Passives as in: that was attached to it yesterday.)
PRED1 PRED = {ADJP / NP / PP / PPL VP}
PRED10 PRED = PRED (",") {PP / ADVP}

PRO Pronouns
I, you, he, we, me, it, us, her, him, they, them.

Q Indefinite Quantifiers (nonstandard)
(A small set of words, including only: many, much, few, little. For other words usually called quantifiers, see DETQ and DDET. DETQs may have Q as sole constituents.)

QDET Determiners or Specifiers of Q and ADJ
very, too, much too.
(As in: much too many, much too hard, very many, very hard.)
QDET1 QDET = QDET "," QDET)

QPP Q Phrases
much, too much, too many, too few, little.
QPP1 QPP = (QDET) Q (QPP)

RELPRO Relative Pronouns
who, which, that, whom, whose.

S Sentences
(Independent sentences of various subtypes. Includes complex and compound sentences.)
SO S = C={SIMP / SDEC / SQ / SWHQ} (ENDPUNCT)
SX1 S = CONJ S
SX2 S = S1 (MIDPUNCT) S2
SX3 S = {PP / ADVP} (",") S
SX4 S = {PPL / ING} VP "," S

SDEC Declarative Sentences
They went there; they might have been going there then; it could be difficult; he is here; that is the book; there are some apples in that basket.
SB1 SDEC = NP (AUX) (ADV1) BEP (ADV2) (PRED)
SD1 SDEC = NP (ADV) (AUXD) VP
STHERE1 SDEC = "THERE" (AUX) BEP NP
({ING {VP / "BE" PRED1} / PRED2 / SREL})

SIMP Imperative Sentences
Put them on the table; don't go; be careful; don't be difficult!

```
  SIMP1     SIMP = ("DO" (NOT)) {"BE" PRED / VP}

SQ          Yes/No Interrogatives
            Is he going; could he have been going; did he go; is it here
            now; are there any more; was it hard to do?
  SBQ1      SQ = BEP NP ((ADV) (ING "BE") PRED)
  SBQ2      SQ = MODALP NP (ADV) (HAVEP PPL) BEP ((ING "BE") PRED)
  SBQ3      SQ = HAVEP NP (ADV) PPL "BE" ((ING "BE") PRED)
  SDQ1      SQ = DOP NP (ADV) VP
  SDQ2      SQ = MODALP NP (ADV) (HAVEP PPL) (BEP ING) VP
  SDQ3      SQ = HAVEP NP (ADV) PPL (BEP ING) VP
  SDQ4      SQ = BEP NP (NOT) ING VP
  STHEREQ1  SQ = BEP "THERE" (NP)
                 ({ING {"BE" PRED1 / VP} / PRED2 / SREL})
  STHEREQ2  SQ = MODALP "THERE" (NOT) (HAVEP PPL) BEP (NP)
                 ({ING {"BE" PRED1 / VP} / PRED2 / SREL})
  STHEREQ3  SQ = HAVEP "THERE" PPL "BE" (NP)
                 ({ING {"BE" PRED1 / VP} / PRED2 / SREL})

SREL        Relative Clauses
            who went there yesterday; to whom he gave it; that was
            there; you saw yesterday; that you saw yesterday; that he
            attached it to; on which he placed it.
  SREL1     SREL = RELPRO {(AUX) BEP PRED / (AUXD) VP}
  SREL2     SREL = ((P) RELPRO) SDEC
  SREL3     SREL = "," SREL ({MIDPUNCT / ENDPUNCT})

SWHQ        WH-Questions
            who is it; who could it have been; what did he do; what
            could he have done; where did they go; on which table is it;
            which table is it on?
  SWH1      SWHQ = {WHNP / WHPP / WHADJP} SQ
  SWH2      SWHQ = WHNP (AUXD) VP
  SWH3      SWHQ = WHNP (AUX) BEP (PRED)

V           Verb Stems
            go, move, do, have, give, arrive, attach, see, seem, break,
            take, tell, say, know, think, want, try, tend.

VP          Verb Phrases
            go there in September, move it from here to there, arrive,
            seem to be difficult, give her a book, give a book to her,
            give her, give a book, give, tell him that it is here, want
            to go, tend to be careful, look at it, look up a book, look
            it up, look into it, found her a book, saw him leave, saw
            him leaving.
            (Verb stems combined with following objects, particles,
            prepositions, and prepositional phrase modifiers.)
  VP1       VP = V (NP1 ({NP2 / P}))
  VP2       VP = V P (NP)
  VP3       VP = V (NP) ("THAT") SDEC
  VP4       VP = V (NP) INFINITIVE
  VP5       VP = V (NP) {PPL VP / ADJP}
```

```
  VP7       VP = V (NP) (ING) {VP / BE PRED}
  VP8       VP = V (NP) {WHPP / WHNP / WHADJP} {SDEC / INFINITIVE}
  VP9       VP = VP (",") {PP / INFINITIVE / ADVP}

WHADJP      Interrogative Adjective Phrases
            how big, how much  bigger than that, how much more difficult
            to do.
  WHADJP1   WHADJP = HOW ADJP

WHDET       Interrogative Determiners
            how much, how many more, whose, which man's.
  WHDET1    WHDET = HOW Q1 (QPP) (ER Q2)
  WHDET2    WHDET = WHNP "GEN"

WHNP        Interrogative NPs
            how many more women, how much water, whose book.
  WHNP1     WHNP = WHDET (NUMBER) (NOMHD) (NCOMP)

WHPP        Interrogative PPs
            where, when, at what time, in which box, from where.
  WHPP1     WHPP = {P WHNP / C={TO / FROM} "WHERE"}
```

Appendix B

ATTRIBUTES

Appendix B

ATTRIBUTES

The major specific attributes that currently affect DIAGRAM's syntactic analysis of a phrase are listed below, grouped with the relevant categories. The name of the attribute is followed by a list of the values it may assume. Where values are not listed, the attribute is binary: it is either present or absent, and if present, its value is T. See also the description of general attributes, page 10.

Attributes of S:

STYPE: (SIMP, SDEC, SQ, SWHQ). Sentence type; affects coordination.

CONJUNCT: Contains a conjunction; affects coordination.

Attributes of AUX, AUXD, BE, BEP, DO, DOP, HAVE, HAVEP:

TENSE: (PAST, PRESENT) Marked as finite; cannot combine with preceding auxiliaries.

NBR: (SG, PL) Number agreement feature.

INFINITIVE: May be non-finite; compare: may have gone, they have gone.

Attributes of N, NOMHD, NP, PRO, WHNP:

TYPE: (COUNT, MASS)

NBR: (SG, PL) Syntactic number: singular or plural. Used heavily in number-agreement tests.

NOMCASE: Is marked as nominative pronoun; e.g., he (cf. him). Cannot be an object of V or P.

PROPN: Is a proper name; does not accept full range of determiners and complements of NPs. E.g., *every Mississippi.

Attributes of V and VP:

VPPL: (EN, ED). Form of participial ending; e.g., taken, waited.

DIROBJ: Accepts a direct object; e.g., assemble it.

INDIROBJ: Accepts an indirect object: e.g., give them something.

INFOBJ: Accepts an infinitive object; e.g., want John to go.

SOBJ: Accepts an S object; e.g, said that he would go.

INGCOMP: Accepts gerundive complement; e.g., saw her leaving.

PPLCOMP: Accepts a participial complement; e.g., found her gone.

PARTICLE: Accepts any member of a list of Ps as a particle; e.g., give up, give away, give it up, give up something.

INSEPARABLE: Accepts any member of a list of Ps as an inseparable preposition; e.g., look into.

BAREV: Is a VP with no objects or complements. Affects ability to modify nominals or function as a passive predicate PRED.

Attributes of ADJ, ADJP:

SOBJ: Accepts an S complement; e.g., possible that he went.

ACOMP: Has a complement (of any type).

ERCOMP: Accepts a comparative complement, e.g., heavier than that.

Attributes of P, PP:

INGCOMP: Accepts a participial complement to form an ADVP; e.g., on going there.

SCOMP: Accepts S complement to form an adverbial ADVP; e.g., after he left.

BAREPREP: A PP consisting only of P. May be a stranded preposition or a particle. Cannot modify a nominal. E.g., gave the book up.

Attributes of minor categories:

The minor categories have some attributes of nominals and adjectivals, including TYPE, NBR, and THANCOMP.

Appendix C

SAMPLE LEXICON AND RULES

Appendix C

SAMPLE LEXICON AND RULES

Sample Lexical Entries

(If a word W1 has the attribute LIKE with a word W2 as value, then word W1 has the same attributes as word W2 except for those specifically assigned to W1.)

```
Words for N
        (APPRENTICE     (TYPE . COUNT)
                        (GENDER M F))
        (BOY            (TYPE . COUNT)
                        (GENDER  M))
        (FISH           (TYPE . MASS))
        (FOOT           (TYPE . COUNT))
        (GIRL           (TYPE . COUNT)
                        (GENDER  F))
        (MAN            (LIKE . BOY))
        (THING          (TYPE . COUNT))
        (WATER          (TYPE . MASS))
        (WOMAN          (LIKE . GIRL))

Irregular Forms in N
        (FEET           (FOOT -S))
        (MEN            (MAN  -S))
        (WOMEN          (WOMAN -S)) )

Words in NOMHD
        (FISH           (TYPE . COUNT))  (Cf. entry for "fish" as N)
        (PEOPLE         (TYPE . COUNT)
                        (NBR . PL))
        (U.S.           (NBR . SG)
                        (PROPN . T)) )

Words in NP
        (JOHN           (NBR . SG)
                        (DEF . T)
                        (GENDER M)
                        (PROPN . T))
        (MARY           (LIKE . JOHN)
                        (GENDER F))
        ((NEW YORK)     (NBR . SG)
                        (TYPE . COUNT)
                        (PROPN . T)) )
```

```
Words for V
        (ARRIVE)
        (ASSEMBLE       (DIROBJ . T))
        (BREAK          (DIROBJ . T)
                        (PARTICLE UP OUT OFF))
                        (PPL . EN))
        (BUY            (DIROBJ . T)
                        (INDIROBJ . T)
                        (DIRECTION FOR FROM BY))
        (FIND           (ADJOBJ . T)
                        (DIROBJ . T)
                        (INDIROBJ . T)
                        (INFOBJ . T)
                        (INCGCOMP . T)
                        (SOBJ . T)
                        (DIRECTION FOR BY)
                        (PARTICLE OUT))
        (FINISH         (INGCOMP . T)
                        (LIKE . ASSEMBLE))
        (GO             (PPL . EN))
        (GIVE           (LIKE . BUY)
                        (DIRECTION TO BY)
                        (PARTICLE UP)
                        (PPL . EN))
        (LOOK           (DIROBJ . T)        (Transitive with inseparable
                        (INSEPARABLE INTO)   preposition only, but this
                        (ADJOBJ . T)         constraint is not yet imposed.)
                        (PARTICLE UP))
        (TRY            (INFOBJ . T))
        (WANT           (DIROBJ . T)
                        (INFOBJ . T)) )

Irregular Forms in V
        (BOUGHT         (BUY ED))
        (BROKE          (BREAK ED))
        (GAVE           (GIVE ED))
        (GONE           (GO EN))
        (WENT           (GO ED)) )

Words for Q
        (FEW            (LIKE . MANY))
        (LITTLE         (LIKE . MUCH))
        (MANY           (TYPE . COUNT)
                        (NBR . PL))
        (MUCH           (TYPE . MASS)
                        (NBR . SG)) )
Irregular Forms in Q
        (MORE           (1 (MANY ER)))
        (MORE           (2 (MUCH ER))) )
```

Examples of Rules

```
(SD1      SDEC = NP (ADV) (AUXD) VP ;
      CONSTRUCTOR [PROGN (COND
                           [(@ AUXD)
                             (OR (AGREE (@ PPL AUXD)
                                        (@ PPL VP))
                                 (F.REJECT (QUOTE F.PPL)))
                             (OR (AGREE NBR NP AUXD)
                                 (F.REJECT (QUOTE F.NBRSD1]
                           (T (OR (NEQ (@ NBR NP)
                                       (QUOTE SG))
                                  (F.REJECT (QUOTE F.NBR)))
                              (@SET TENSE (QUOTE PRESENT]
      TRANSLATOR (PROGN (@SET ROLE (QUOTE SUBJECT)
                              NP)
                        (TRANSLATE (@ VP))
                        (TRANSLATE (@ NP))))

(NP1      NP = (D={A / DDET / DETQ}) NOMHD (NCOMP) ;
      CONSTRUCTOR (PROGN (COND
                           ((@ D)
                             [COND
                               ((MASS? D)
                                 (OR (MASS? NOMHD)
                                     (F.REJECT (QUOTE F.MASS]
                             [COND
                               ((MASS? NOMHD)
                                 (OR (NOT (@ A))
                                     (F.REJECT (QUOTE F.MASS]
                             [COND
                               ((@ NCOMP)
                                 (@SET NBR (@INTERSECT NBR D NOMHD NCOMP)))
                               (T (@SET NBR (@INTERSECT NBR D NOMHD]
                             (@FROM D DEF NOT))
                           ((AND (SG? NOMHD)
                                 (NOT (MASS? NOMHD)))
                             (@FACTOR (QUOTE F.NODET)
                                      UNLIKELY))
                           ((@ NCOMP)
                             (@SET NBR (@INTERSECT NBR NCOMP NOMHD)))
                           (T (@FROM NOMHD NBR)))
                         [AND (@ THANCOMP NCOMP)
                              (OR (@ THANCOMP NOMHD)
                                  (F.REJECT (QUOTE F.THANC]
                         (@FROM NOMHD TYPE)))

(VP1      VP = V (NP 1 ({NP 2 / P})) ;
      CONSTRUCTOR (PROG ((PARTICLE (@ DIAMOND.SPELLING P)))
                        (COND
                          [(@ NP 1)
                            (OR (@ DIROBJ V)
```

```
                        (F.REJECT (QUOTE F.DIROBJ)))
                    (AND (@ NOMCASE NP 1)
                         (F.REJECT (QUOTE F.NOMCASE)))
                    (COND
                      [(@ NP 2)
                        (OR (@ INDIROBJ V)
                            (F.REJECT (QUOTE F.INDIROBJ)))
                        (AND (@ NOMCASE NP 2)
                             (F.REJECT (QUOTE F.NOMCASE]
                      ((@ P)
                        (OR (FMEMB PARTICLE (@ PARTICLE V))
                            (F.REJECT (QUOTE F.PARTICLE)))
                        (AND (@ PRO NP 1)
                             (@FACTOR (QUOTE F.PARTICLE)
                                      LIKELY))
                        (COND
                          ((@ NCOMP NP 1)
                            (OR (@ NP NCOMP NP 1)
                                (@FACTOR (QUOTE F.PARTICLE)
                                         UNLIKELY))
                            (AND (@ NCOMP NP NCOMP NP 1)
                                 (@FACTOR (QUOTE F.PARTICLE)
                                          UNLIKELY]
                      (T (@SET BAREV T)
                         (@FROM V DIRECTION DIROBJ)))
                    (@FROM V PPL))
TRANSLATOR (PROGN [COND
                    ((@ NP 2)
                      (@SET ROLE (QUOTE DIROBJ)
                            NP 2)
                      (@SET ROLE (QUOTE INDIROBJ)
                            NP 1))
                    (T (AND (@ NP 1)
                            (OR (@ INDIROBJ V)
                                (@SET ROLE (QUOTE DIROBJ)
                                          NP 1]
                    (TRANSLATE (@ NP 1))
                    (TRANSLATE (@ NP 2))))
```

Appendix D

SAMPLE CONSTITUENT STRUCTURE ANALYSES

Appendix D

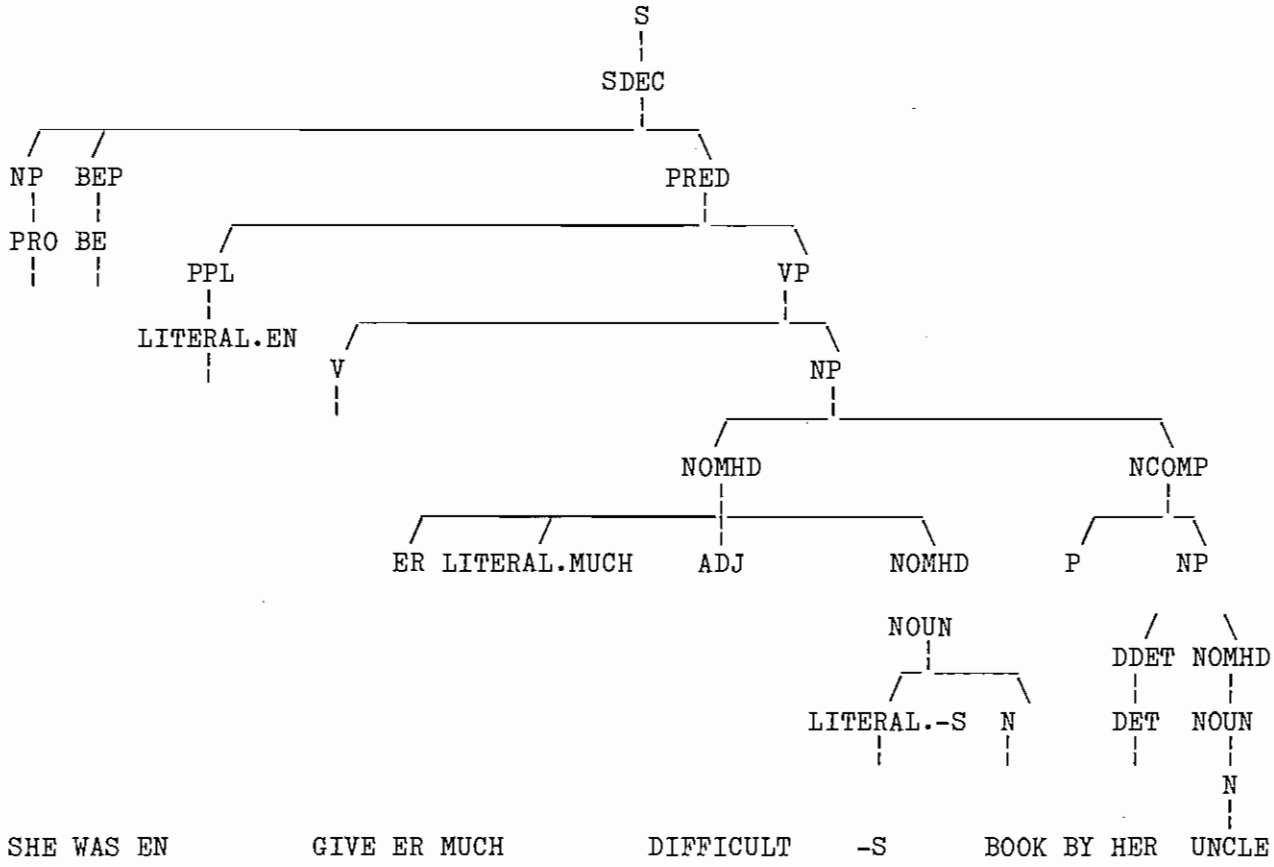
SAMPLE CONSTITUENT STRUCTURE ANALYSES

1. ((SHE WAS GIVEN MORE DIFFICULT BOOKS BY HER UNCLE))

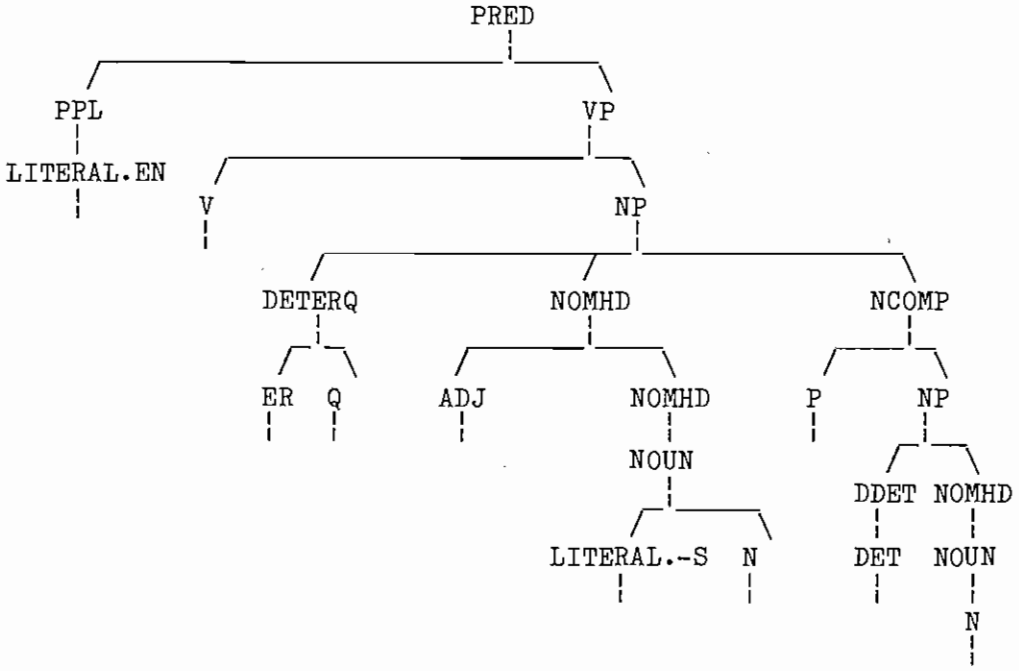
The sentence contains two syntactically ambiguous phrases: either the books were more difficult or there were more books, and either the books were "by" her uncle or the giving was "by" her uncle. These two ambiguities combine to produce four different analyses of the sentence. (There is an additional semantic ambiguity, since "by" has both a locative and an agentive sense.) The first of DIAGRAM's four analyses is given below in two forms. In one, each node of a tree printed vertically is labelled with the name of the rule that was applied. In the second, each node of a tree drawn horizontally is labelled with the category of the phrase assigned by the rule.

(In this reading, the books were more difficult and were by her uncle.)

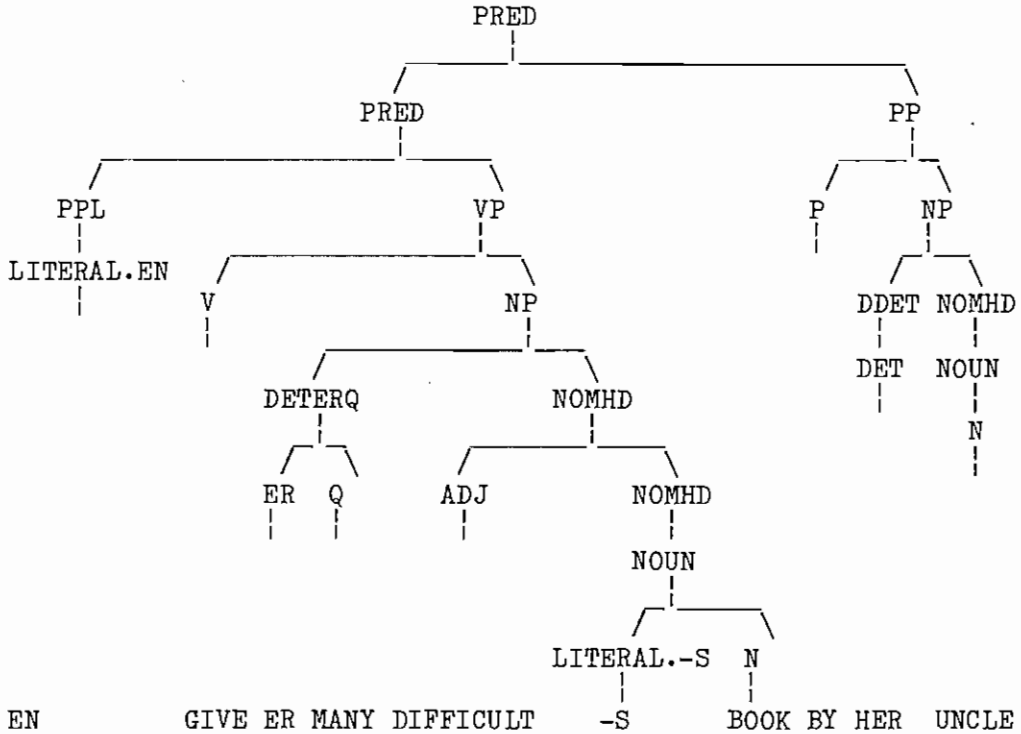
```
140 SO
  139 SB1
    2 NP5
      1 PRO SHE
    4 BEP1
      3 BE WAS
    138 PRED1
      8 PPL1
        7 "EN"
      137 VP1
        9 V GIVE
        136 NP1
          61 NOMHD3
            14 ER ER
            25 "MUCH"
            35 ADJ DIFFICULT
            46 NOMHD1
              45 NOUN1
                43 "-S"
                44 N BOOK
          129 NCOMP1
            69 P BY
            128 NP1
              88 DDET1
                87 DET HER
              127 NOMHD1
                126 NOUN1
                  125 N UNCLE
```


S
SDEC
NP
BEP
PRED
PRO
BE
PPL
VP
LITERAL.EN
V
NP
NOMHD
NCOMP
ER
LITERAL.MUCH
ADJ
NOMHD
P
NP
NOUN
DDET
NOMHD
LITERAL.-S
N
DET
NOUN
N
SHE WAS EN
GIVE ER MUCH
DIFFICULT
-S
BOOK BY HER
UNCLE

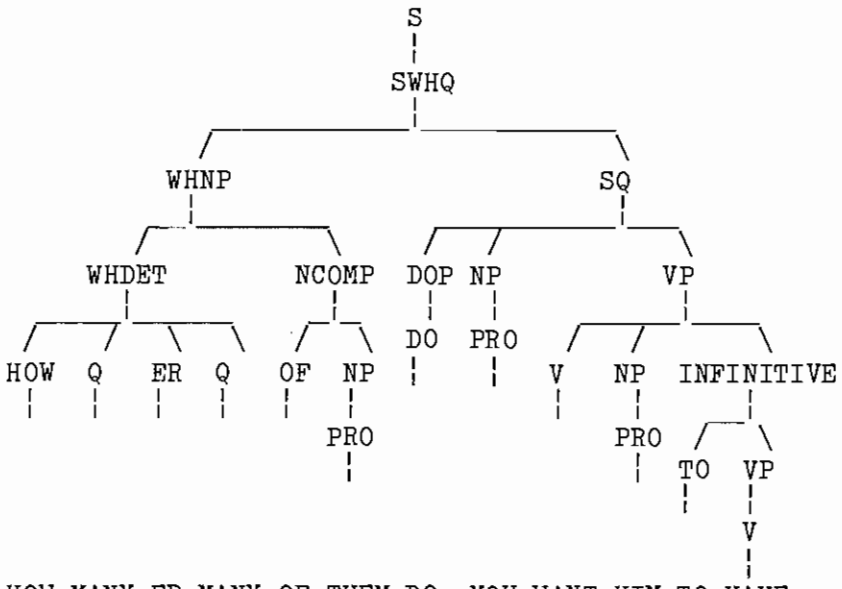
(The next tree shows an analysis of the predicate in which there were more books by her uncle and they were difficult.)



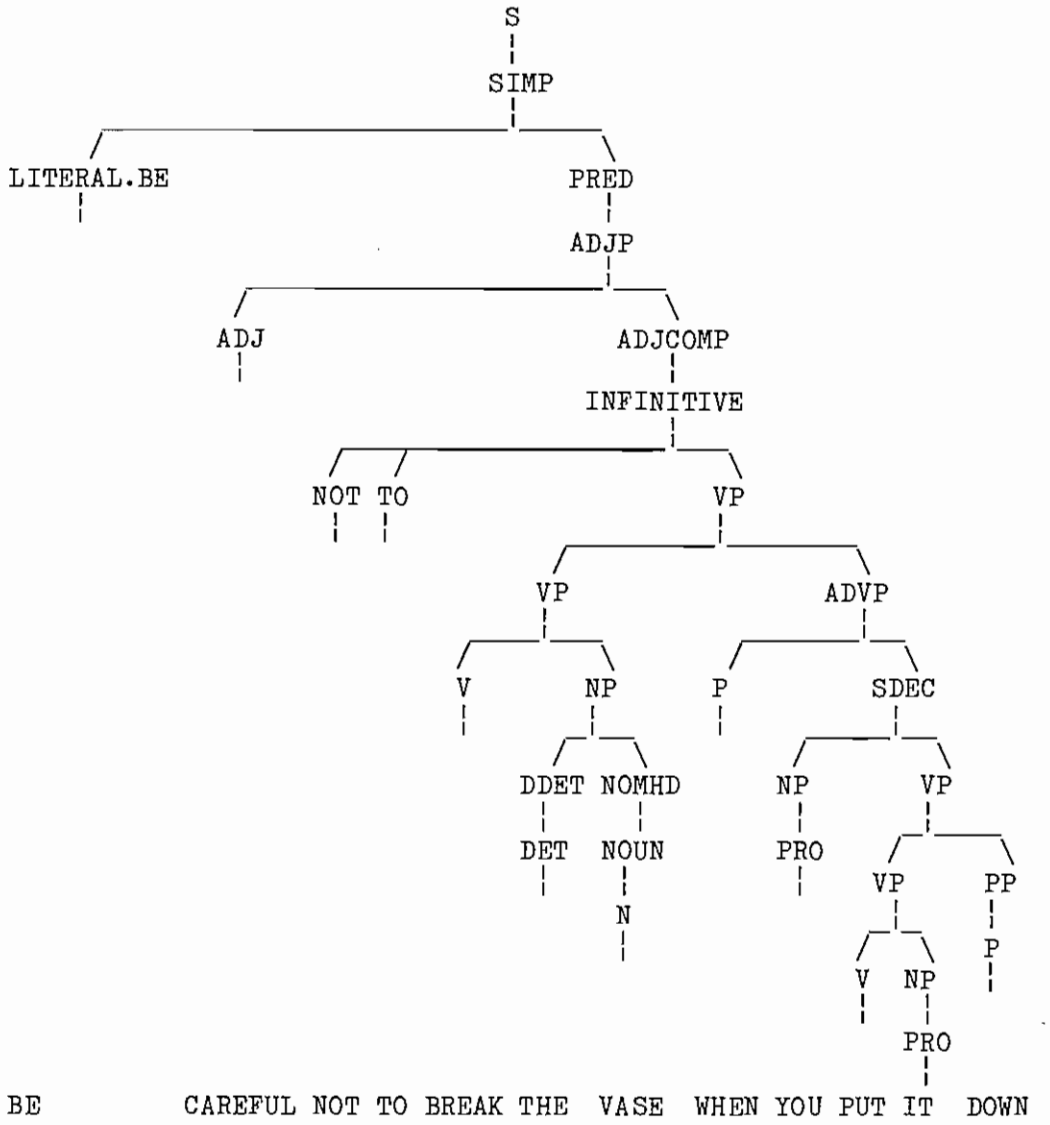
(The next tree shows an analysis of the predicate in which the giving was by her uncle.)



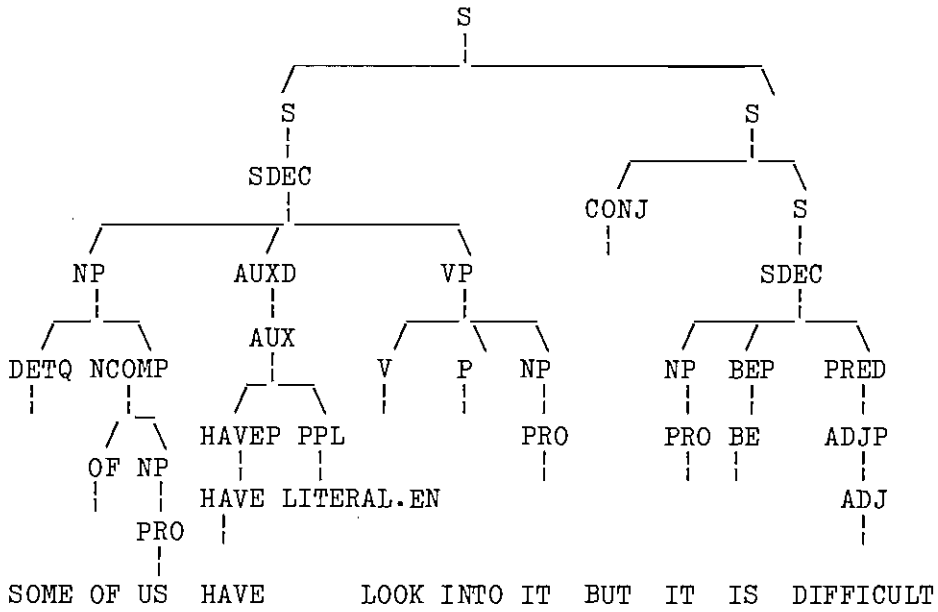
2. ((HOW MANY MORE OF THEM DO YOU WANT HIM TO HAVE))



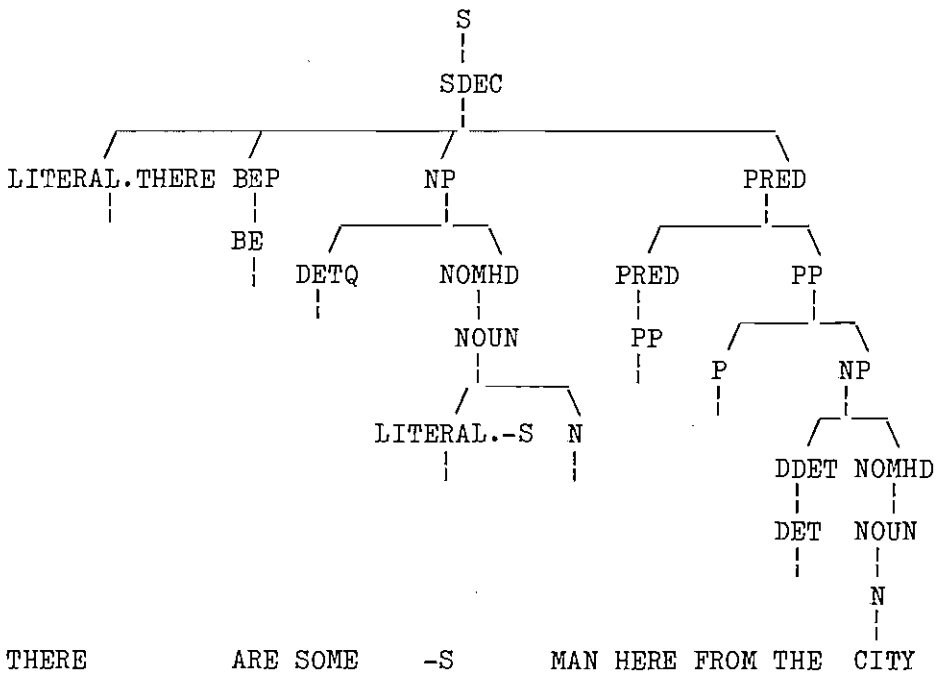
3. ((BE CAREFUL NOT TO BREAK THE VASE WHEN YOU PUT IT DOWN))



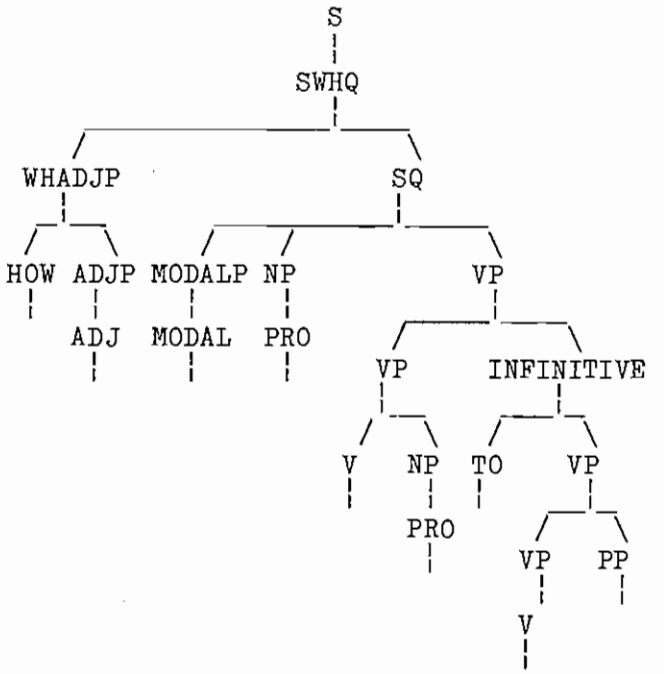
4. ((SOME OF US HAVE LOOKED INTO IT BUT IT IS DIFFICULT))



5. ((THERE ARE SOME MEN HERE FROM THE CITY))



6. ((HOW LONG WILL IT TAKE YOU TO GO THERE))



7. ((IS THIS ANY HARDER FOR HIM TO DO THAN THAT IS?))