SEMANTICS OF CONTEXT-FREE FRAGMENTS OF NATURAL LANGUAGES

by

Patrick Suppes

TECHNICAL REPORT NO. 171

March 30, 1971

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Semantics of Context-free Fragments of Natural Languages[1]

Patrick Suppes

Stanford University
Stanford, California 94305

1. Introduction

The search for a rigorous and explicit semantics of any significant portion of a natural language is now intensive and far-flung--far-flung in the sense that wide varieties of approaches are being taken. Yet almost everyone agrees that at the present time the semantics of natural languages are less satisfactorily formulated than the grammars, even though a complete grammar for any significant fragment of natural language is yet to be written.

A line of thought especially popular in the last couple of years is that the semantics of a natural language can be reduced to the semantics of first-order logic. One way of fitting this scheme into the general approach of generative grammars is to think of the deep structure as being essentially identical with the structure of first-order logic. The central difficulty with this approach is that now as before how the semantics of the surface grammar is to be formulated is still unclear. In other words, how can explicit formal relations be established between first-order logic and the structure of natural languages? Without the outlines of a formal theory, this line of approach has moved no further than the classical stance of introductory teaching in logic, which for many years has concentrated on the translation of English sentences into

first-order logical notation. The method of translation, of course, is left at an intuitive and ill-defined level.

The strength of the first-order logic approach is that it represents essentially the only semantical theory with any systematic or deep development, namely, model-theoretic semantics as developed in mathematical logic since the early 1930's, especially since the appearance of Tarski (1935). The semantical approaches developed by linguists or others whose viewpoint is that of generative grammar have been lacking in the formal precision and depth of model-theoretic semantics. Indeed, some of the most important and significant results in the foundations of mathematics belong to the general theory of models. I shall not attempt to review the approaches to semantics that start from a generative-grammar viewpoint, but I have in mind the work of Fodor, Katz, Lakoff, McCawley and others.

My objective is to combine the viewpoint of model-theoretic semantics and generative grammar, to define semantics for context-free languages and to apply the results to some fragments of natural language. The ideas contained in this paper were developed while I was working with Hélène Bestougeff on the semantical theory of question-answering systems. Later I came across some earlier similar work by Knuth (1968). My developments are rather different from those of Knuth, especially because my objective is to provide tools for the analysis of fragments of natural languages, whereas Knuth was concerned with programming languages.

Although on the surface the viewpoint seems different, I also benefited from a study of Montague's interesting and important work (1970) on the analysis of English as a formal language. My purely extensional

line of attack is simpler than Montague's. I adopted it for reasons of expediency, not correctness. I wanted an apparatus that could be applied in a fairly direct way to empirical analysis of a corpus. As in my work on probabilistic grammars (Suppes, 1970), I began with the speech of a young child, but without doubt, many of the semantical problems that are the center of Montague's concern must be dealt with in analyzing slightly more complex speech. Indeed, some of these problems already arise in the corpus studied here. As in the case of my earlier work on probabilistic grammars, I have found a full-scale analytic attack on a corpus of speech a humbling and bedeviling experience. The results reported here hopefully chart one possible course; in no sense are they more than preliminary.

This paper is organized in the following fashion. In Section 2, I describe a simple artificial example to illustrate how a semantic valuation function is added to the generative machanisms of a context-free grammar. The relevant formal definitions are given in Section 3. The reader who wants a quick survey of what can be done with the methods, but who is not really interested in formal matters, may skip ahead to Section 4, which contains the detailed empirical results. On the other hand, it will probably be somewhat difficult to comprehend fully the machinery used in the empirical analysis without some perusal of Section 3, unless the reader is already quite familiar with model-theoretic semantics. How the results of this paper and the earlier one on probabilistic grammars are meant to form the beginnings of a theory of performance is sketched in Section 5.

2. A Simple Example

To illustrate the semantic methods described formally below, I use as an example the same simple language I used in Suppes (1970). As remarked there, this example is not meant to be complex enough to fit any actual corpus; its context-free grammar can easily be rewritten as a regular grammar. The five syntactic categories are IV, TV, Adj, PN and N, where IV is the class of intransitive verbs, TV the class of transitive verbs or two-place predicates, Adj the class of adjectives, PN the class of proper nouns and N the class of common nouns. Additional nonterminal vocabulary consists of the symbols S, NP, VP and AdjP. The set P of production rules consists of the following seven rules, plus the rewrite rules for terminal vocabulary belonging to one of the five categories.

	Production Rule	Semantic Function
1.	S → NP + VP	Truth-function
2.	VP → IV	Identity
3.	VP → TV + NP	Image under the converse relation
4.	NP → PN	Identity
5.	NP → AdjP + N	Intersection
6.	AdjP → AdjP + Adj	Intersection
7.	AdjP → Adj	Identity

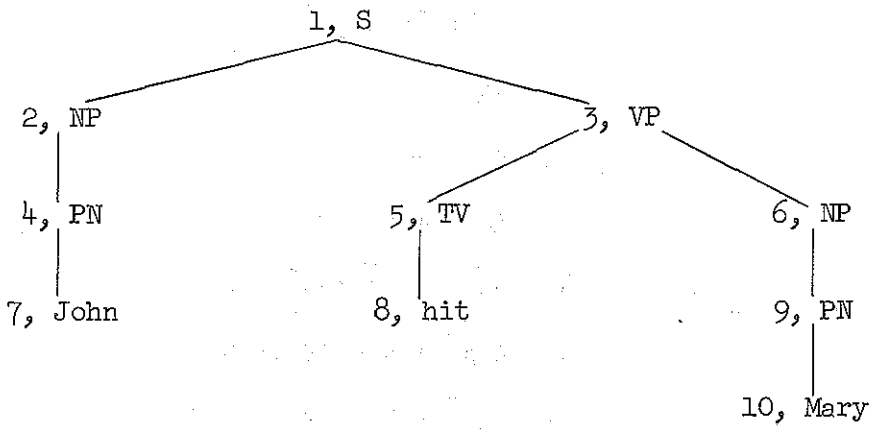
If Adj^n is understood to denote a string of n adjectives, then the possible grammatical types (infinite in number) all fall under one of the following schemes.

Grammatical Type

1. $PN + IV$
2. $PN + TV + PN$
3. $Adj^n + N + V_1$
4. $PN + TV + Adj^n + N$
5. $Adj^n + N + TV + PN$
6. $Adj^m + N + TV + Adj^n + N$

What needs explaining are the semantic functions to the right of each production rule. For this purpose it is desirable to look at an example of a sentence generated by this grammar. The intuitive idea is that we define a valuation function v over the terminal vocabulary, and as is standard in model-theoretic semantics, v takes values in some relational structure.

Suppose a speaker wants to say 'John hit Mary'. The valuation function needs to be defined for the three terminal words 'John', 'hit' and 'Mary'. We then recursively define the denotation of each labeled node of the derivation tree of the sentence. In this example, I number the nodes, so that the denotation function ψ is defined for pairs (n,α), where n is a node of the tree and α is a word in the vocabulary. The tree looks like this.



Let I be the identity function, $\breve{A}$ the converse of A, i.e.,

$$\breve{A} = \{\langle x,y\rangle : \langle y,x\rangle \in A\} ,$$

and f"A the image of A under f, i.e., the range of f restricted to the domain A, and let T be truth and F falsity. Then the denotation of each labeled node of the tree is found by working from the bottom up:[2]

$$\psi(10, \text{Mary}) = v(\text{Mary})$$

$$\psi(9, \text{PN}) = I(v(\text{Mary}))$$

$$\psi(8, \text{hit}) = v(\text{hit})$$

$$\psi(7, \text{John}) = v(\text{John})$$

$$\psi(6, \text{NP}) = I(v(\text{Mary}))$$

$$\psi(5, \text{TV}) = I(v(\text{hit}))$$

$$\psi(4, \text{PN}) = I(v(\text{John}))$$

$$\psi(3, \text{VP}) = \overbrace{I(v(\text{hit}))}\text{"}\, I(v(\text{Mary}))$$

$$\psi(2, \text{NP}) = II(v(\text{John}))$$

$$\psi(1, \text{S}) = f(\psi(2, \text{NP}), \psi(3, \text{VP})) = \begin{cases} T & \text{if } \psi(2, \text{NP}) \subseteq \psi(3, \text{VP}) \\ F & \text{otherwise.} \end{cases}$$

Clearly, the functions used above are just the semantic functions associated with the productions. In particular, the production rules for the direct descendants of nodes 2, 4, 5 and 6 all have the identity function as their semantic function.

One point should be emphasized. I do not claim that the set-theoretical semantic functions of actual speech are as simple as those associated with the production rules given in this section. Consider Rule 5, for instance. Intersection is fine for old dictators, but not for alleged dictators. One standard mathematical approach to this kind of difficulty is to generalize the semantic function to cover the meaning of both sorts of cases. In the present case of adjectives, we could require that the semantic function be one that maps sets of objects into sets of objects. In this vein, Rule 5 would now be represented by

$$\psi(n_1, \text{NP}) = \psi(n_2, \text{AdjP}) \text{ " } \psi(n_3, \text{N}) \ .$$

Fortunately, generalizations that rule out the familiar simple functions as semantic functions do not often occur early in children's speech. Some tentative empirical evidence on this point is presented in Section 4.

3. Denoting Grammars

I turn now to formal developments. Some standard grammatical concepts are defined in the interest of completeness. First, if V is a set, V^* is the set of all finite sequences whose elements are members of V. I shall often refer to these finite sequences as strings. The empty sequence, 0, is in V^*; we define $V^+ = V^* - \{0\}$. A structure $G = \langle V, V_N, P, S \rangle$ is a phrase-structure grammar if and only if V and P

are finite, nonempty sets, V_N is a subset of V, S is in V_N and $P \subseteq V_N^* \times V^+$. Following the usual terminology, V_N is the nonterminal vocabulary and $V_T = V - V_N$ the terminal vocabulary. S is the start symbol of the single axiom from which we derive strings or words in the language generated by G. The set P is the set of production or rewrite rules. If $\langle \alpha, \beta \rangle \in P$, we write $\alpha \rightarrow \beta$, which we read: from α we may produce or derive β (immediately).

A phrase-structure grammar $G = \langle V, V_N, P, S \rangle$ is <u>context-free</u> if and only if $P \subseteq V_N \times V^+$, i.e., if $\alpha \rightarrow \beta$ is in P then $\alpha \in V_N$ and $\beta \in V^+$.[3] These ideas may be illustrated by considering the simple language of the previous section. Although it is intended that N, PN, Adj, IV, and TV be nonterminals in any application, we can treat them as terminals for purposes of illustration, for they do not occur on the left of any of the seven production rules. With this understanding

$$V_N = \{S, NP, VP, AdjP\}$$
$$V_T = \{N, PN, Adj, IV, TV\}$$

and P is defined by the production rules already given. It is obvious from looking at the production rules that the grammar is context-free, for only elements of V_N appear on the left-hand side of any of the seven production rules.

The standard definition of derivations is as follows. Let $G = \langle V, V_N, P, S \rangle$ be a phrase-structure grammar. First, if $\alpha \rightarrow \beta$ is a production of P, and γ and δ are strings in V^*, then $\gamma \alpha \delta \underset{G}{\Rightarrow} \gamma \beta \delta$. We say that β is <u>derivable</u> from α in G, in symbols, $\alpha \overset{*}{\underset{G}{\Rightarrow}} \beta$ if there are strings $\alpha_1, \ldots, \alpha_n$ in V^* such that $\alpha = \alpha_1$,

$\alpha_1 \underset{G}{\Rightarrow} \alpha_2, \ldots, \alpha_{n-1} \underset{G}{\Rightarrow} \alpha_n = \beta$. The sequence $\Delta = \langle \alpha_1, \ldots, \alpha_n \rangle$ is a derivation in G. The language L(G) generated by G is $\{\alpha : \alpha \in V_T^* \ \& \ S \overset{*}{\underset{G}{\Rightarrow}} \alpha\}$. In other words, L(G) is the set of all strings made up of terminal vocabulary and derived from S.

The semantic concepts developed also require use of the concept of a derivation tree of a grammar. The relevant notions are set forth in a series of definitions. Certain familiar set-theoretical notions about relations are also needed. To begin with, a binary structure is an ordered pair $\langle T,R \rangle$ such that T is a nonempty set and R is a binary relation on T, i.e., $R \subseteq T \times T$. R is a partial ordering of T if and only if R is reflexive, antisymmetric and transitive on T. R is a strict simple ordering of T if and only if R is asymmetric, transitive and connected on T. We also need the concept of R-immediate predecessor. For x and y in T, xJy if and only if xRy, not yRx and for every z if $z \neq y$ and zRy then zRx. In the language of formal grammars, we say that if xJy then x directly dominates y, or y is the direct descendant of x.

Using these notions, we define in succession tree, ordered tree and labeled ordered tree. A binary structure $\langle T,R \rangle$ is a tree if and only if (i) T is finite, (ii) R is a partial ordering of T, (iii) there is an R-first element of T, i.e., there is an x such that for every y, xRy, and (iv) if xJz and yJz then x = y. If xRy in a tree, we say that y is a descendant of x. Also the R-first element of a tree is called the root of the tree, and an element of T that has no descendants is called a leaf. We call any element of T a node, and we shall sometimes refer to leaves as terminal nodes.

A ternary structure $\langle T,R,L\rangle$ is an ordered tree if and only if (i) L is a binary relation on T, (ii) $\langle T,R\rangle$ is a tree, (iii) for each x in T, L is a strict simple ordering of $\{y : xJy\}$, (iv) if xLy and yRz then xLz, and (v) if xLy and xRz then zLy. It is customary to read xLy as "x is to the left of y." Having this ordering is fundamental to generating terminal strings and not just sets of terminal words. The terminal string of an ordered labeled tree is just the sequence of labels $\langle f(x_1),\ldots,f(x_n)\rangle$ of the leaves of the tree as ordered by L. Formally, a quinary structure $\langle T,V,R,L,f\rangle$ is a labeled ordered tree if and only if (i) V is a nonempty set, (ii) $\langle T,R,L\rangle$ is an ordered tree, and (iii) f is a function from T into V. The function f is the labeling function and $f(x)$ is the label of node x.

The definition of a derivation tree is relative to a given context-free grammar.

Definition 1. Let $G = \langle V,V_N,P,S\rangle$ be a context-free grammar and let $\mathcal{T} = \langle T,V,R,L,f\rangle$ be a labeled ordered tree. $\mathcal{T}$ is a derivation tree of G if and only if

(i) If x is the root of $\mathcal{T}$, $f(x) = S$;

(ii) If xRy and $x \neq y$ then $f(x)$ is in V_N;

(iii) If $y_1,\ldots,y_n$ are all the direct descendants of x, i.e., $\bigcup_{i=1}^{n}\{y_i\} = \{y : xJy\} \neq \emptyset$, and $y_i L y_j$ if $i < j$, then

$$\langle f(x),\langle f(y_1),\ldots,f(y_n)\rangle\rangle$$

is a production in P.

We now turn to semantics proper by introducing the set Φ of set-theoretical functions. We shall let the domains of these functions be n-tuples of any sets (with some appropriate restriction understood to avoid set-theoretical paradoxes).

Definition 2. Let $\langle V, V_N, P, S \rangle$ be a context-free grammar. Let Φ be a function defined on P which assigns to each production p in P a finite, possibly empty set of set-theoretical functions subject to the restriction that if the right member of production p has n terms of V, then any function of $\Phi(p)$ has n arguments. Then $G = \langle V, V_N, P, S, \Phi \rangle$ is a potentially denoting context-free grammar. If for each p in P, $\Phi(p)$ has exactly one member then G is said to be simple.

The simplicity and abstractness of the definition may be misleading. In the case of a formal language, e.g., a context-free programming language, the creators of the language specify the semantics by defining Φ. Matters are more complicated in applying the same idea of capturing the semantics by such a function for fragments of a natural language. Perhaps the most difficult problem is that of giving a straightforward set-theoretical interpretation of intensional contexts, especially to those generated by the expression of propositional attitudes of believing, wanting, seeking and so forth. I shall not attempt to deal with these matters in the present paper.

How the set-theoretical functions in $\Phi(p)$ work was illustrated in the preceding section; some empirical examples follow in the next section. The problems of identifying and verifying Φ even in the simplest sort of context are discussed there. In one sense the definition should be

strengthened to permit only one function in $\Phi(p)$ of a given number of arguments. The intuitive idea behind the restriction is clear. In a given application we try first to assign denotations at the individual word level, and we proceed to two- and three-word phrases only when necessary. The concept of such hierarchical parsing is familiar in computer programming, and a detailed example in the context of a question-answering program is worked out in a joint paper with Hélène Bestougeff. However, as the examples in the next section show, this restriction seems to be too severe for natural languages.

A clear separation of the generality of Φ and an evaluation function v is intended. The functions in Φ should be constant over many different uses of a word, phrase or statement. The valuation v, on the other hand, can change sharply from one occasion of use to the next. To provide for any finite composition of functions, or other ascensions in the natural hierarchy of sets and functions built up from a domain of individuals, the family $\mathscr{K}'(D)$ of sets with closure properties stronger than needed in any particular application is defined. The abstract objects T (for truth) and F (for falsity) are excluded as elements of $\mathscr{K}'(D)$. In this definition $\mathcal{P}A$ is the power set of A, i.e., the set of all subsets of A.

Definition 3. Let D be a nonempty set. Then $\mathscr{K}'(D)$ is the smallest family of sets such that

(i) $D \in \mathscr{K}'(D)$,

(ii) if $A, B \in \mathscr{K}'(D)$ then $A \cup B \in \mathscr{K}'(D)$,

(iii) if $A \in \mathscr{K}'(D)$ then $\mathcal{P}A \in \mathscr{K}'(D)$,

(iv) if $A \in \mathscr{K}'(D)$ and $B \subseteq A$ then $B \in \mathscr{K}'(D)$.

We define $\mathscr{K}(D) = \mathscr{K}'(D) \cup \{T,F\}$, with $T \notin \mathscr{K}'(D)$, $F \notin \mathscr{K}'(D)$ and $T \neq F$.

A model structure for G is defined just for terminal words and phrases. The meaning or denotation of nonterminal symbols changes from one derivation or derivation tree to another.

Definition 4. *Let* D *be a nonempty set, let* $G = \langle V, V_N, P, S \rangle$ *be a phrase-structure grammar, and let* v *be a partial function on* V_T^+ *to* $\mathcal{K}(D)$ *such that if* v *is defined for* α *in* V_T^+ *and if* γ *is a subsequence of* α, *then* v *is not defined for* γ. *Then* $\mathcal{D} = \langle D, v \rangle$ *is a model structure for* G. *If the domain of* v *is exactly* V_T, *then* $\mathcal{D}$ *is simple.*

We also refer to v as a *valuation function* for G.

I now define semantic trees that assign denotations to nonterminal symbols in a derivation tree. The definition is for simple potentially denoting grammars and for simple model structures. In other words, there is a unique semantic function for each production, and the valuation function is defined just on V_T, and not on phrases of V_T^+.

Definition 5. *Let* $G = \langle V, V_N, P, S, \Phi \rangle$ *be a simple, potentially denoting context-free grammar, let* $\mathcal{D} = \langle D, v \rangle$ *be a simple model structure for* G, *let* $\mathcal{T}' = \langle T, V, R, L, f \rangle$ *be a derivation tree of* $\langle V, V_N, P, S \rangle$ *such that if* x *is a terminal node then* $f(x) \in V_T$ *and let* ψ *be a function from* f *to* $\mathcal{K}(D)$ *such that*

(i) *if* $\langle x, f(x) \rangle \in f$ *and* $f(x) \in V_T$ *then*

$$\psi(x, f(x)) = v(f(x)) ,$$

(ii) *if* $\langle x, f(x) \rangle \in f$, $f(x) \in V_N$, *and* $y_1, \ldots, y_n$ *are all the direct descendants of* x *with* $y_i L y_j$ *if* $i < j$, *then*

$$\psi(x, f(x)) = \varphi(\psi(y_1, f(y_1)), \ldots, \psi(y_n, f(y_n)) ,$$

where $\varphi = \Phi(p)$ and p is the production

$$\langle f(x), \langle f(y_1), \ldots, f(y_n)\rangle\rangle .$$

Then $\mathcal{T} = \langle T, V, R, L, f, \Psi\rangle$ is a simple semantic tree of G and $\mathcal{B}$.

The extension of Definition 5 to semantic trees that are not simple is relatively straightforward, but is not given explicitly here in the interest of restricting the formal parts of the paper. The empirical examples considered in the next section implicitly assume this extension, but the simplicity of the corpus makes the several set-theoretical functions φ attached to a given production easy to interpret.

The function Ψ assigns a denotation to each node of a semantic tree. The resulting structural analysis can be used to define a concept of meaning or sense for each node. Perhaps the most natural intuitive idea is this. Extend the concept of a model structure by introducing a set of situations. For each situation σ $\langle D_\sigma, v_\sigma\rangle$ is a model structure. The meaning or sense of an utterance is then the function Ψ of the root of the tree of the utterance. For example, using the analysis of John hit Mary from Section 3, dropping the redundant notation for the identity function and using the ordinary lambda notation for function abstraction, we obtain as the meaning of the sentence

$$\Psi(1,S) = (\lambda\sigma)f(v_\sigma(\underline{\text{John}}), v_\sigma(\underline{\text{hit}})'' v_\sigma(\underline{\text{Mary}})) ,$$

but this idea will not be developed further here. Its affinity to Kripke-type semantics is clear.

4. Noun-Phrase Semantics of Adam I

In Suppes (1970), I proposed and tested a probabilistic noun-phrase grammar for Adam I, a well-known corpus of the speech of a young boy

(about 26 months old) collected by Roger Brown and his associates--and once again I wish to record my indebtedness to Roger Brown for generously making his transcribed records available for analysis. Eliminating immediate repetitions of utterances, we have a corpus of 6109 word occurrences with a vocabulary of 673 different words and 3497 utterances. Noun phrases dominate the corpus. Of the 3497 utterances, I have classified 936 as single occurrences of nouns, another 192 as occurrences of two nouns in sequence, 147 as adjective followed by noun, and 138 as adjectives alone. The context-free grammar for the noun phrases of Adam I has seven production rules, and the theoretical probability of using each rule in a derivation is also shown for purposes of later discussion. From a probabilistic standpoint, the grammar has five free parameters: the sum of the a_i's is one, so the a_i's contribute four parameters and $b_1 + b_2 = 1$, whence the b_i's contribute one more parameter. To the right are also shown the main set-theoretical functions that make the grammar potentially denoting. These semantic functions, as it is convenient to call them in the present context, are subsequently discussed extensively. I especially call attention to the semantic function for Rule 5, which is formally defined.

Noun-Phrase Grammar for Adam I

	Production Rule	Probability	Semantic Function
1.	NP → N	a_1	Identity
2.	NP → AdjP	a_2	Identity
3.	NP → AdjP + N	a_3	Intersection
4.	NP → Pro	a_4	Identity
5.	NP → NP + NP	a_5	Choice function
6.	AdjP → AdjP + Adj	b_1	Intersection
7.	AdjP → Adj	b_2	Identity

As I remarked in the earlier article, except for Rule 5, the production rules seem standard and an expected part of a noun-phrase grammar for standard English. The new symbol introduced in V_N beyond those introduced already in Section 2 is Pro for pronoun; inflection of pronouns is ignored. On the other hand, the special category, PN, for proper nouns is not used in the grammar of Adam I.

The basic grammatical data are shown in Table I. The first column gives the types of noun phrases actually occurring in the corpus in

---

Insert Table I about here

---

decreasing order of frequency. Some obvious abbreviations are used to shorten notation: A for Adj, P for Pro. The grammar defined generates an infinite number of types of utterances, but, of course, all except a small finite number have a small probability of being generated. The second column lists the numerical observed frequencies of the utterances (with immediate repetition of utterances deleted from the frequency count). The third column lists the theoretical or predicted frequencies when a maximum-likelihood estimate of the five parameters is made (for details on this see the earlier article). The impact of semantics on these theoretical frequencies is discussed later.

The fourth column lists the observed frequency with which the "standard" semantic function shown above seems to provide the correct interpretation for the five most frequent types. Of course, in the case of the identity function, there is not much to dispute, and so I concentrate entirely on the other two cases. First of all, if the derivation uses more than one rule, then by standard interpretation

TABLE I

Probabilistic Noun-Phrase Grammar for Adam I

Noun phrase	Observed frequency	Theoretical frequency	Stand. semantic function
N	1445	1555.6	1445
P	388	350.1	388
NN	231	113.7	154
AN	135	114.0	91
A	114	121.3	114
PN	31	25.6	
NA	19	8.9	
NNN	12	8.3	
AA	10	7.1	
NAN	8	8.3	
AP	6	2.0	
PPN	6	.4	
ANN	5	8.3	
AAN	4	6.6	
PA	4	2.0	
ANA	3	.7	
APN	3	.1	
AAA	2	.4	
APA	2	.0	
NPP	2	.4	
PAA	2	.1	
PAN	2	1.9	

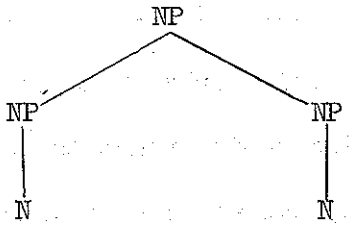
I mean the derivation that only uses Rule 5 if it is necessary and that interprets each production rule used in terms of its standard semantic function. Since none of the derivations is very complex, I shall not spend much time on this point.

The fundamental ideas of denoting grammars as defined in the preceding section come naturally into play when a detailed analysis is undertaken of the data summarized in Table I. The most important step is to identify the additional semantic functions if any in $\Phi(p)$ for each of the seven production rules. A simple way to look at this is to examine the various types of utterances listed in Table I, summarize the production rules and semantic functions used for each type, and then collect all of this evidence in a new summary table for the production rules.

Therefore I now discuss the types of noun phrases listed in Table I and consider in detail the data for the five most frequently listed.

Types N and P, the first two, need little comment. The identity function, and no other function, serves for them. It should be clearly understood, of course, that the nouns and pronouns listed in these first two lines--a total of 1833 without immediate repetition--do not occur as parts of a larger noun phrase. The derivation of N uses only P1 (Production Rule 1), and the derivation of P uses only P4.

The data on type NN are much richer and more complex. The derivation is unique; it uses P5 then P1 twice, as shown in the tree. As before, the semantic function for P1 is just the identity function, so all the analysis of type NN centers around the interpretation of P5. To begin with, I must explain what I mean by the choice function shown above as



the standard semantic function of P5. This is a set-theoretical function of A and B that for each A is a function selecting an element of B when B is the argument of f. Thus

$$\varphi(A,B) = f_A(B) \in B .$$

I used 'A' rather than an individual variable to make the notation general, but in all standard cases, A is a unit set. (I emphasize again, I do not distinguish unit sets from their members.) A standard set-theoretical choice function, i.e., a function f such that if B is in the domain of f and B is nonempty then $f(B) \in B$, is a natural device for expressing possession. Intuitively, each of the possessors named by Adam has such a function and the function selects his (or hers or its) object from the class of like objects. Thus Daddy chair denotes that chair in the class of chairs within Adam's purview that belongs to or is used especially by Daddy. If we restrict our possessors to individuals, then in terms of the model structure $\mathcal{A} = \langle D, v \rangle$, $\varphi(A,B)$ is just a partial function from $D \times \wp(D)$ to D, where $\wp(D)$ is the power set of D.[4]

The complete classification of all noun phrases of type NN is given in Table II. (I emphasize that this classification must be

---

Insert Table II about here

---

regarded as tentative at this early stage of investigation.) As the data in Table II show, the choice function is justly labeled the standard semantic function for P5, but at least four other semantic functions belong in $\Phi(P5)$. One of these is the converse of $\varphi(A,B)$ as defined above, i.e.,

$$\breve{\varphi}(A,B) = f_B(A) \ ,$$

which means the possessor is named after the thing possessed. Here are examples from Adam I for which this interpretation seems correct: part trailer (meaning part of trailer), part towtruck, book boy, name man, ladder firetruck, taperecorder Ursula. The complete list is given in Table II.

The third semantic function is a choice function on the Cartesian product of two sets, often the sets' being unit sets as in the case of Mommy Daddy. Formally, we have

$$\varphi(A,B) = f(A \times B) \ ,$$

and $f(A \times B) \in A \times B$. Other examples are Daddy Adam and pencil paper. The frequency of use of this function is low, however--only 12 out of 230 instances according to the classification shown in Table II.

The fourth semantic function proposed for $\Phi(P5)$ is the intersection function,

$$\varphi(A,B) = A \cap B \ .$$

Examples are lady elephant and lady Ursula. Here the first noun is functioning like an adjective.

TABLE II

Semantic Classification of Noun Phrases of Type NN*

Choice function	
Adam checker	Adam horn
Adam hat	Adam hat
Adam bike	Adam pillow
Moocow tractor	Moocow truck
Catherine dinner	Car mosquito
Newmi book	Newmi bulldozer
Daddy briefcase	Adam book
Adam book	Adam paper
Daddy chair	Daddy tea
Mommy tea	Tuffy boat
Tuffy boat	Adam pencil
Adam tractor	Tuffy boat
Judy buzz	Judy buzz
Ursula pocketbook	Ursula pocket
Daddy name	Daddy name
Daddy Bozo	Daddy Johnbuzzhart
Daddy name	Adam light
Catherine Bozo	Monroe suitcase
Adam glove	Adam ball
Adam locomotive	Daddy racket
Daddy racket	Adam racket
Adam pencil	Joshua shirt
Joshua foot	Adam busybulldozer
Robie nail	Adam busybulldozer
Train track	Adam Daddy
Daddy suitcase	Cromer suitcase
Adam suitcase	Daddy suitcase
Adam doggie	Adam doggie
Choochoo track	Daddy Adam
Adam water	Ursula water
Ursula car	Adam house
Hobo truck	Doctordan circus
Doctordan circus	Joshua book
Daddy paper	Adam Cromer
Cromer coat	Adam pencil
Adam pillow	Mommy pillow
Adam pillow	Daddy pillow
Dan circus	Doctordan circus

*Whenever the type NN appeared in the context of a longer utterance, the entire utterance is printed.

Adam ladder
Adam mouth
Doctordan circus
Adam horn
Adam piece
Adam playtoy
Doggie car
Adam book
Adam shirt
Adam ball
Cromer suitcase
Adam letter
Adam firetruck
Bambi wagon
Like Adam bookshelf
Pull Adam bike
Write Daddy name
Hit Mommy wall
Hit Adam roadgrader (?)
Spill Mommy face
Bite Cromer mouth
Hit Mommy ball
Get Adam ball
Write Cromer shoe
Sit Missmonroe car
Walk Adam Bambi
Adam Panda march (?)
Oh Adam belt
Adam bite rightthere (?)
Fish water inhere
Put Adam bandaid on
Put Missmonroe towtruck (?)
Mommy tea yeah
Adam school tomorrow
Daddy suitcase goget it
Take off Adam paper
No Adam Bambi
That Adam baby
Powershovel pick Adam dirt up

Adam mouth
Daddy desk
Adam sky
Adam baby
Adam candy
Kitchen playtoy
Man Texacostar (?)
Adam paper
Adam pocketbook
Daddy suitcase
Adam suitcase
Adam pencil
Adam firetruck
See Daddy car
Give doggie paper
Read Doctor circus
Write Daddy name
Hit Mommy rug
See Adam ball
Bite Mommy mouth
Bite Ursula mouth
Take Adam car
Sit Adam chair
Sit Monroe car
Walk Adam Bambi
Going Cromer suitcase
Doggies tummy hurt
Yeah locomotive caboose
Adam shoe rightthere
Take lion nose off
Pick roadgrader dirt (?)
Put Adam boot
Adam pencil yeah
Becky star tonight
Adam pocket no
Big towtruck pick Joshua dirt up
Look Bambi Adam pencil
Break Cromer suitcase Mommy
Where record folder go

Converse of choice function

Part trailer
Book boy
Ladder firetruck
Part head
Foot Adam
Car train
Taperecorder Ursula

Part towtruck
Name man
Record Daddy
Part game
Track train
Part broom
Circus Dan

Speghetti Cromer
Part basket
Game Adam
Take piece candy
Excuseme Ursula part broom

Part apple
Piece candy
Time bed (?)
Paper kitty open

Choice function on Cartesian product

Pencil paper
Mommy Daddy
Mommy Daddy
Pencil roadgrader (?)
Busybulldozer truck (?)
Jack Jill come

Paper pencil
Towtruck fire
Record taperecorder
Jack Jill
Give paper pencil
Adam wipeoff Cromer Ursula

Intersection

Lady elephant
Lady Ursula
Toy train

Lady Ursula
Lady elephant
Record box

Identity

Pin Game
Daddy Cromer (?)
Doctor Doctordan

Babar Pig
Mommy Cromer (?)

Unclassified

Joshua home
Train train (Repetition?)
Dog pepper
Suitcase water
Doggie pepper
Daddy home (S)
Door book
Pumpkin tomato
Chew apple mouth (2)
Hit door head (2)
Hit head trash (2)
Show Ursula Bambi (2)
Look car mosquito (2)
Pick dirt shovel up (2)
Ohno put hand glove (2)

Pencil doggie
Adam Adam (Repetition?)
Kangaroo bear
Doggie doggie (Repetition ?)
Kangaroo marchingbear
Ball playtoy (?)
Pumpkin tomato
Put truck window (2)
Hit towtruck knee (2)
Make Cromer Doctordan (2)
Hurt knee chair (2)
Show Ursula Bambi (2)
Daddy Daddy work (Repetition?)
Mommy time bed
Time bed Mommy

Note.--230 utterances of type NN are shown instead of the 231 shown in Table 1, because one of the 231 was incorrectly classified as NN.

The fifth semantic function, following in frequency the choice function and its converse, is the identity function. It seems clear from the transcription that some pairs of nouns are used as a proper name or a simple description, even though each noun is used in other combinations. (By a simple description I mean a phrase such that no subsequence of it denotes (see Definition 4).) Some examples are pin game and Daddy Cromer.

I do not consider in the same detail the next two most frequent types shown in Table I, namely, AN and A. The latter, as in the case of N and P, is served without complications by the identity function. As would be expected, the picture is more complicated for the type AN. Column 4 of Table I indicates that 91 of the 135 instances of AN can be interpreted as using intersection as the semantic function. Typical examples are these: big drum, big horn, my shadow, my paper, my tea, my comb, oldtime train, that knee, green rug, that man, poor doggie, pretty flower. The main exceptions to the intersection rule are found in the use of numerical or comparative adjectives like two or more. Among the 116 AN phrases standing alone, i.e., not occurring as part of a longer utterance, 19 have two as the adjective; for example, two checkers, two light, two sock, two men, two boot, two rug. No numerical adjective other than two is used in the 116 phrases.

I terminate at this point the detailed analysis of the Adam I corpus, but some computations concerning the length of noun phrases in Adam I are considered in the next section.

5. Towards a Theory of Performance

The ideas developed in this paper and in my earlier paper on probabilistic grammars are meant to be steps toward a theory of performance. In discussing the kind of theories of language wanted by linguists, philosophers or psychologists, I have become increasingly aware of the real differences in the objectives of those who want a theory of ideal competence and those who are concerned with performance. Contrary to the opinions expressed by some linguists, I would not concede for a moment that a theory of competence must precede in time the development of a theory of performance. I do recognize, on the other hand, the clear differences of objectives in the two kinds of theories. The linguistic and philosophical tradition of considering elaborate and subtle examples of sentences that express propositional attitudes is very much in the spirit of a theory of competence. The subtlety of many of these examples is far beyond the bulk of sentences used in everyday discourse by everyday folk. The kind of corpus considered in the preceding section is a far cry from most of these subtle examples.

The probabilistic grammars discussed in the preceding section, and elaborated upon more thoroughly in the earlier paper, clearly belong to a theory of performance. Almost all of the linguists or philosophers interested in theories of competence would probably reject probabilistic grammars as being of any interest to such theories. On the other hand, from the standpoint of a theory of performance, such grammars immediately bring to hand a detailed analysis of actual speech as well as a number of predictions about central characteristics of actual speech that are not a part of a theory of competence. Perhaps the simplest and clearest

example is predictions about the distribution of length of utterances. One of the most striking features of actual speech is that most utterances are of short duration, and no utterances are of length greater than 10^4 even though in the usual theories of competence there is no way of predicting the distribution of length of utterance and no mechanism for providing it. A probabilistic grammar immediately supplies such a mechanism, and I would take it to be a prime responsibility of a theory of performance to predict the distribution of utterances from the estimation of a few parameters.

Here, for example, are the theoretical predictions of utterance length in terms of the parameters a_i and b_j assigned to the production rules for Adam I noun phrases. In order to write a simple recursive expression for the probability of a noun phrase of length n, I use ℓ_i for the probability of an utterance of length $i < n$. Thus, for example, one of the terms in the expression for the probability of a noun phrase of length 3 is $2a_5\ell_1\ell_2$. By first using Rule 5 (with probability a_5) and then generating for one NP a noun phrase of length 1, which starting from NP has probability ℓ_1, and generating for the other NP a noun phrase of length 2 with probability ℓ_2, we obtain $2a_5\ell_1\ell_2$, since this can happen in two ways. We have in general the following:

Length of noun phrase	Probability of this length
1	$a_1 + a_2 b_2 + a_4$
2	$a_2 b_1 b_2 + a_3 b_2 + a_5 (a_1 + a_2 b_2 + a_4)^2$
3	$a_2 b_1^2 b_2 + a_3 b_1 b_2 + 2a_5 \ell_1 \ell_2$
⋮	⋮
n	$a_2 b_1^{n-1} b_2 + a_3 b_1^{n-2} b_2 + a_5 \sum_{\substack{1 \le i,j < n \\ i+j=n}} \ell_i \ell_j$

Using the maximum-likelihood estimates of the parameters a_i and b_j obtained to make the theoretical predictions of Table I, we can compare theoretical and observed distributions of noun-phrase length for Adam I. The results are shown in Table III for lengths up to 3.

Insert Table III about here

Because this paper is mainly concerned with semantics, I shall not pursue these grammatical matters further, but turn to the way in which the theory of semantics developed here is meant to contribute to a theory of performance. From a behavioral standpoint it is much easier to describe the objective methods used in constructing a probabilistic grammar, because the corpus of sentences and the classification of individual words into given syntactic categories can be objectively described and verified by any interested person. The application of the theory, in other words, has an objective character that is on the surface. Matters are different when we turn to semantics. For example, it does not seem possible to

TABLE III

Prediction of Length of Noun Phrases for Adam I

Length	Observed frequency	Theoretical frequency
1	1947	2027.1
2	436	314.1
3	51	66.9
> 3	0	25.9
	2434	2434.0

state directly objective criteria by which the classification of semantic functions as described in the preceding section are made. Clearly I have taken advantage of my own intuitive knowledge of the language in an inexplicit way to interpret Adam's intended meaning in using a particular utterance. If the methodology for applying semantics to actual speech had to be left at the level of analysis of the preceding section, objections could certainly be made that the promise of such a semantics for a theory of performance was very limited.

A first naive approach to applying semantics to the development of a more complete theory of performance might have as an objective the prediction of the actual sentences uttered by a speaker. Everyone to whom this proposal is made instantly recognizes the difficulty, if not the impossibility, of predicting the actual utterance made once the structure of the utterance goes beyond something like a simple affirmation or denial. Frequently the next step is to use this common recognition of difficulty as an argument for the practical impossibility of applying any concepts of probability in analyzing actual speech behavior. This skeptical attitude has been expressed recently by Chomsky (1969, p. 57) in the following passage:

> . . . If we return to the definition of 'language' as a "complex of dispositions to verbal behavior", we reach a similar conclusion, at least if this notion is intended to have empirical content. Presumably, a complex of dispositions is a structure that can be represented as a set of probabilities for utterances in certain definable 'circumstances' or 'situations'. But it must be recognized that

> the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term. On empirical grounds, the probability of my producing some given sentence of English--say, this sentence, or the sentence "birds fly" or "Tuesday follows Monday", or whatever--is indistinguishable from the probability of my producing a given sentence of Japanese. Introduction of the notion of 'probability relative to a situation' changes nothing, at least if 'situations' are characterized on any known objective grounds (we can, of course, raise the conditional probability of any sentence as high as we like, say to unity, relative to 'situations' specified on ad hoc, invented grounds).

One can agree with much of what Chomsky says in this passage, but also recognize that it is written without familiarity with the way in which probability concepts are actually used in science. What is said here applies almost without change to the study of the simplest probabilistic phenomenon, e.g., the flipping of a coin. If we construct a probability space for a thousand flips of a coin, and if the coin is approximately a fair one, then the actual probability of any observed sequence is almost zero, namely, approximately 2^{-1000}. If we use a representation that is often used for theoretical purposes and take the number of trials to be infinite, then the probability of any possible outcome of the experiment in this theoretical representation is strictly zero. It in no sense follows that the concept of probability cannot be applied in a meaningful way to the flipping of a coin. A response may

be that a single flip has a high probability and that this is not the case for a single utterance, but corresponding to utterances, we can talk about sequences of flips and once again we have extraordinarily low probabilities attached to any actual sequence of flips of length greater than, say, a hundred. What Chomsky does not seem to be aware of is that in most sophisticated applications of probability theory the situation is the same as what he has described for sentences. The basic objects of investigation have either extremely small probabilities or strictly zero probabilities. The test of the theory then depends upon studying various features of the observed outcome. In the case of the coin the single most interesting feature is the relative frequency of heads, but if we are suspicious of the mechanism being used to toss the coin we may also want to investigate the independence of trials.

To make the comparison still more explicit, Chomsky's remarks about the equal probability of uttering an English or Japanese sentence can be mimicked in discussing the outcomes of flipping a coin. The probability of a thousand successive heads in flipping a fair coin is 2^{-1000}, just the probability of any other sequence of this length. Does this equal probability mean that we should accept the same odds in betting that the relative frequency of heads will be less than 0.6, and betting that it will be greater than 0.99? Certainly not. In a similar way there are many probabilistic predictions about verbal behavior that can be made, ranging from trivial predictions about whether a given speaker will utter an English or Japanese sentence to detailed predictions about grammatical or semantic structure. Our inability to predict the unique flow of discourse no more invalidates a definition of language as a "complex of

dispositions to verbal behavior" than our inability to predict the trajectory of a single free electron for some short period of time invalidates quantum mechanics--even in a short period of time any possible trajectory has strictly zero probability of being realized on the continuity assumptions ordinarily made.

Paradoxically, linguists like Chomsky resist so strongly the use of probability notions in language analysis just when these are the very concepts that are most suited to such complex phenomena. The systematic use of probability is to be justified in most applications in science because of our inability to develop an adequate deterministic theory.

In the applications of probability theory one of the most important techniques for testing a theory is to investigate the theoretical predictions for a variety of conditional probabilities. The concept of conditional probability and the related concept of independence are the central concepts of probability theory. It is my own belief that we shall be able to apply these concepts to show the usefulness of semantics at a surface behavioral level. Beginning with a probabilistic grammar, we want to improve the probabilistic predictions by taking into account the postulated semantic structure. The test of the correctness of the semantic structure is then in terms of the additional predictions we can make. By taking account of the semantic structure, we can make differential probabilistic predictions and thereby show the behavioral relevance of semantics. Without entering into the kind of detailed data analysis of the preceding section, let me try to indicate in more concrete fashion how such an application of semantics is to be made.

I have reported previously the analysis of the corpus of Adam I. We have also been collecting data of our own at Stanford, and we have at hand a corpus of some 20 hours of Erica, a rather talkative 30-month-old girl.[5] We have been concerned to write a probabilistic grammar for Erica of the same sort we have tried to develop for Adam I. The way in which a semantic structure can be used to improve the predictions of a probabilistic grammar can be illustrated by considering Erica's answers to the many questions asked her by adults. For the purposes of this sketch, let me concentrate on some of the data in the first hour of the Erica corpus. According to one straightforward classification, 169 questions were addressed to Erica by an adult during the first hour of the corpus. These 169 questions may be fairly directly classified in the following types: what-questions, yes-no-questions, where-questions, who-questions, etc. The frequency of each type of question is as follows:

What-questions	79
Yes-no-questions	60
Where-questions	12
Who-questions	9
Why-questions	4
How-many-questions	3
Or-questions	1
How-do-you-know-questions	1

By taking account of the most obvious semantic features of these different types of questions, we can improve the probabilistic predictions of the

kind of responses Erica makes without claiming that we can make an exact prediction of her actual utterances. Moreover, the semantic classification of the questions does not depend on any simple invariant features of the surface grammar. For example, some typical yes-no-questions, with Erica's answers in parentheses, are these: Can you sit on your seat please? (O.K.), You don't touch those, do you? (No), Aren't they? (Uh huh. That Arlene's too), He isn't old enough is he? (No. Just Martin's old enough.)

It is an obvious point that the apparatus of model-theoretic semantics is not sufficient to predict the choice of a particular description of an object from among many semantically suitable ones. Suppose John and Mary are walking, and John notices a spider close to Mary's shoulder. He says, "Watch out for that spider." He does not say, "Watch out for the black, half-inch long spider that has a green dot in its center and is about six inches from your left shoulder at a vertical angle of about sixty degrees." The principle that selects the first utterance and not the second I call a principle of minimal discrimination. A description is selected that is just adequate to the perceptual or cognitive task. Sometimes, of course, a full sentence rather than a noun phrase is used in response to a what-question, the sort of question whose answer most naturally exemplifies a minimal principle. Here is an example from Erica: What do you want for lunch? (Peanut butter and jelly), What do you want to drink? (I want to drink peanut butter). In answering what-questions by naming or describing an object, Erica uses adjectives only sparingly, and then mainly in a highly relevant way. Here are a couple of examples: What are you going to ride on? (On a big towel), What are those? (Oboe and clarinet. And a flute. Little bitty flute called a piccolo.). Preliminary

analysis of the Erica corpus indicates that even a relatively crude probabilistic application of the principle of minimal discrimination can significantly improve predictions about Erica's answers. Presentation of systematic data on this point must be left for another occasion.

I want to finish by stressing that I do not have the kind of imperialistic ambitions for a theory of performance that many linguists seem to have for a theory of competence. I do not think a theory of performance need precede a theory of competence. I wish only to claim that the two can proceed independently--they have sufficiently different objectives and different methods of analysis so that their independence, I would venture to suggest, will become increasingly apparent. A probabilistic account of main features of actual speech is a different thing from a theory-of-competence analysis of the kind of subtle examples found in the literature on propositional attitudes. The investigation of these complicated examples certainly should not cease, but at the present time they have little relevance to the development of a theory of performance. The tools for the development of a theory of performance, applied within the standard scientific theory of probability processes, are already at hand in the concepts of a probabilistic grammar and semantics. Unfortunately, many linguists dismiss probabilistic notions out of hand and without serious familiarity with their use in any domain of science.

Quine ended a recent article (1970) with a plea against absolutism in linguistic theory and methodology. It is a plea that we all should heed.

References

N. Chomsky, 'Quine's Empirical Assumptions' in Words and Objections Essays on the Work of W. V. Quine (ed. by D. Davidson and J. Hintikka), Dordrecht, Holland, 1969, pp. 53-68.

D. E. Knuth, 'Semantics of Context-Free Languages', Mathematical Systems Theory 2 (1968) 127-131.

R. Montague, 'English as a Formal Language' in Linguaggi Nella Società e Nella Tecnica (ed. by B. Visentini et al.) Milan, 1970, pp. 189-224.

P. Suppes, 'Probabilistic Grammars for Natural Languages', Synthese 22 (1970) 95-116.

A. Tarski, 'Der Wahrheitsbegriff in Den Formalisierten Sprachen', Studia Philosophica 1 (1935) 261-405.

W. V. Quine, 'Methodological Reflections on Current Linguistic Theory', Synthese 21 (1970) 386-398.

Footnotes

1. This research has been supported by the National Science Foundation under grant NSFGJ-443X. I am indebted to Pentti Kanerva for help in the computer analysis and organization of the data presented in Section 4, and I am indebted to Dr. Elizabeth Gammon for several useful ideas in connection with the analysis in Section 4. D. M. Gabbay and George Huff have made a number of penetrating comments on Section 3, and Richard Montague trenchantly criticized an unsatisfactory preliminary version.

2. I have let the words of V serve as names of themselves to simplify the notation.

3. As Richard Montague pointed out to me, to make context-free grammars a special case of phrase-structure grammars, as defined here, the first members of P should be not elements of V_N, but one-place sequences whose terms are elements of V_N. This same problem arises later in referring to elements of V*, but treating elements of V as belonging to V*. Consequently, to avoid notational complexities, I treat elements, their unit sets and one-place sequences whose terms are the elements, as identical.

4. Other possibilities exist for the set-theoretical characterization of possession. In fact, there is an undesirable asymmetry between the choice function for Adam hat and the intersection function for my hat, but it is also clear that v(my) can in a straightforward sense be the set of Adam's possessions but v(Adam) is Adam, not the set of Adam's possessions.

5. The corpus was taped and edited by Arlene Moskowitz.