A New Method of Representing Constituent Structures

D. TERENCE LANGENDOEN
Department of Linguistics
University of Arizona
Tucson, Arizona 85721

YEDIDYAH LANGSAM
Brooklyn College
Brooklyn, New York 11210

THE PROBLEM

DESPITE the great progress that has been made in recent years in linguistic theory and the grammatical analysis of natural languages, many fundamental questions about the nature and use of language remain unanswered. One of the most important and difficult of these questions is how people are able to understand the expressions they hear or read in the languages they know in essentially no more time than they require to process them phonetically or visually. We have, however, one important clue as to the nature of this ability. People cannot, under ordinary conditions, comprehend any expression in a language which contains a deeply center-embedded constituent, whereas they can comprehend almost every short or moderately long expression in that language which contains no deeply center-embedded constituent (Chomsky 1963; Miller and Chomsky 1963). This suggests that human comprehension of the expressions of a natural language under ordinary conditions can be modelled by a finite automaton or transducer (Chomsky 1959).

Now, we may suppose that one aspect of the comprehension of natural-language expressions is the determination of the arrangements of the parts of those expressions into constituents (Fodor, Bever and Garrett 1974). Many linguists, in fact, following Chomsky (1965), assume that at least two distinct levels of constituent structure must be determined, a level of 'deep structure' and a level of 'surface structure'. However, even if we suppose that it is sufficient for purposes of comprehension to determine constituent structure at one level only, say the level of S-structure ('enriched' surface structure) in the theory of Chomsky (1980), the following, seemingly insuperable, problem arises. The constituent structures of only finitely many expressions without coordinate

* This work was supported in part by a grant from the PSC-CUNY Faculty Research Award Program.
compounding can be recognized by a finite transducer, if those structures are represented in the standard form of labelled bracketing or tree diagrams (Langeleno 1975). In other words, no finite transducer is capable of recognizing the constituent structures in standard notation of all the comprehensible expressions in any given natural language.

The problem is easily illustrated. Consider the following English sentence (E). (Some of the rules of English punctuation are suspended in the citing of English examples, in order to clarify the relation between those examples and the representations of their grammatical structures.)

(E) the principal believed that the teacher knew that the student thought that the alarm meant that school was out

At the very least, any representation of the grammatical structure of this example must include the information that it is a sentence which contains four subordinate sentences (clauses), each of which, in turn, contains the next. Further, while the five sentences begin at different points, they all end at the same point. If we use a labelled bracketing structure to represent this information, we find that it looks something like (B).

(B) { the principal believed that { the teacher knew that { the student thought that { the alarm meant that { school was out } } } } } } } Clearly, in order to recognize structures of this sort, a parser would have to keep track of the number of sentences that are 'open' at any given time, in order to 'close' them when they are finished. But the ability to keep track of a potentially unbounded number of open constituents is beyond the capacity of any finite-state device. Hence if human beings represent the grammatical structures of the expressions of the languages they know as labelled bracketing structures of the type B (or as the equivalent tree structures), they should find multiply right-branching sentences of the type E exceedingly difficult to understand. Since they do not, it must be the case either that human parsing mechanisms are more powerful computationally than finite-state devices, or people represent grammatical structures in a form that allows a finite-state device to compute them whenever the corresponding expressions have no more than some small, fixed degree of center embedding. Since, as we have already pointed out, it seems reasonable to assume that human comprehension of natural languages can be modelled by finite transducers, let us look for alternative methods of representing grammatical structures which can be computed by finite-state devices whenever the expressions themselves can be.

TOWARD A FORMAL SOLUTION TO THE PROBLEM

To begin with, suppose that L is a set of expressions of a language of finite length, that L is unambiguous, and that L is adequately described by a context-free phrase-structure grammar G all of whose rules are of the form A → X or A → a, where A is a category, X is a nonnull string of categories, and a
branching and which in right branching, and there is no effective procedure for doing that (Chomsky 1963). However, though no general effective procedure for jointly satisfying (P1) and (P2) exists, it may be possible to do so for a linguistically significant subset of phrase-structure grammars. In the four following subsections, we show how affix grammars can be constructed ad hoc for a variety of phrase-structure grammars of different formal types. Then in the next section, we consider how affix grammars can be effectively constructed for phrase-structure grammars meeting certain linguistically significant conditions.

Unambiguous Noncenter-Embedding Grammars

Let GI be the rules of an unambiguous noncenter-embedding grammar that generates the language (L1). Note that since (GI) is noncenter embedding, (L1) = (L1(0)).

(L1) \{a^m c b^n : m, n \geq 0 \}.

(GI) a. S \rightarrow A S
b. S \rightarrow C

(GI)* is another noncenter-embedding grammar, constructible from (GI), which is an affix grammar for (L1). Since (GI)* is constructible from (GI) and is noncenter embedding, it satisfies both principles (P1) and (P2). In an affix grammar, the starred symbols are nonterminal symbols and the unstarred symbols are terminal symbols, the capitalized ones being affixes. The axiom of (GI)* is S*.

(GI)* a. S* \rightarrow S A S*
b. S* \rightarrow C* S
c. C* \rightarrow C* B* C

Among the expressions generated by (GI) is (E1); the affixed structure generated by (GI)* that corresponds to (E1) is (A1). The substrings on affixes in (A1) and in subsequent affixed structures merely serve to individuate tokens of a given affix type and are not generated by the affix grammar.

(E1) a a c b b b

(A1) S1 A1 a S2 A2 a C1 b C2 b C3 b C4 b C5 b C6 S6

In order for affixed structures such as (A1) to be considered representations of the constituent structures of expressions with respect to the grammars that generate them, rules for the interpretation of affixes in those structures must be applied. Such rules are given in (R1)–(R3).

(R1) Suppose that (i) K is an affix that is immediately followed by a terminal symbol k, and (ii) K \rightarrow k is a rule of the original (unstarred) grammar. Then K may be a prefix whose daughters are those occurrences of L1, ..., Lm.

(R2) Suppose that (i) K is an affix that is not immediately followed by a terminal symbol, (ii) K is followed by affixes L1, ..., Lm in that order (m \geq 1), and (iii) K \rightarrow L1, ..., Lm is a rule of the original grammar. Then K may be a postfix whose daughters are those occurrences of L1, ..., Lm.

The structure (A1) is interpreted from left to right according to (R1)–(R3) as in (II). (In these interpretive tables, the first column lists the affixes as they are encountered from left to right; the second column, the type of affix (prefix or postfix); the third column, the daughter(s) of that affix; the fourth column, the number of the applicable interpretive rule; the fifth column, the number of the rule of grammar used in the interpretive rule; and the sixth column, the number of the step in the interpretive process.)

(II) Affix Type Daughters I-Rule G-Rule Step
S1 Pre A1, S2 R2 Gla 1
A1 Pre a R1 Gla 2
S2 Pre A2, S3 R2 Gla 3
A2 Pre a R1 Gla 4
C1 Pre c R1 Gla 5
B1 Pre b R1 Gla 6
C2 Post C1, B1 R3 Gla 7
C3 Post C2, B2 R3 Gla 8
B3 Post b R1 Gla 9
C4 Post C3, B3 R3 Gla 10
S4 Post C4 R3 Gla 12

With the interpretation given in (II), (A1) represents the constituent structure of (E1) with respect to (GI).

Finally, note that not only can the affixed structures generated by (GI)* be recognized by a finite automaton, but the interpretations of those structures as representations of the structural descriptions of the expressions generated by (GI) can be assigned by a finite transducer, because it is always the first appropriate sequence of elements to the right of a given prefix or to the left of a given postfix that is identified as its daughters. This ability of a finite-state device to interpret the structures generated by (GI)* rests on the fact that that grammar is itself not center embedding.

Unambiguous Center-Embedding Grammars

Affix grammars satisfying (P1) and (P2) can be constructed for center-embedding grammars as well. Consider first the unambiguous center-embedding grammar (G2) that generates the language (L2).

(L2) \{a^n c b^n : n \geq 0 \}

(G2) a. S \rightarrow A D
b. D \rightarrow S B
c. A \rightarrow a
d. B \rightarrow b
e. S \rightarrow c
(G2)* is another unambiguous center-embedding grammar, constructible from (G2), that generates affixed structures for the expression in (L2). It is easily verified that the degree of center embedding of affixed structures is the same as that of the corresponding expressions in (L2), and hence that (G2) satisfies both (P1) and (P2).

\[(G2)* a. \ S^* \rightarrow S \ A^* \ D^* \\
\quad b. \ D^* \rightarrow S^* \ B^* \ D \]

\[c. \ A^* \rightarrow A \ a \\
\quad d. \ B^* \rightarrow B \ b \\
\quad e. \ S^* \rightarrow S \ c \]

Among the expressions generated by (G2) is (E2); the corresponding affixed structure generated by (G2)* is (A2).

\[(E2) \ a \ a \ c \ b \ b \]

\[(A2) \ S, \ A, \ a \ S, \ A_2 \ a \ S, \ b \ B, \ b \ D, \ b \ D_2 \]

We find that the structure (A2) cannot be straightforwardly interpreted from left to right according to (R1)–(R3) as the constituent structure of (E2) with respect to (G2). For example, the prefix S, in (A2) has as its daughters not the next following affixes A, and D, but rather A, and D; D is a daughter of S, in (A2). Similarly, the postfix D, in (A2) has as its daughters not the next preceding affixes S, and B, but rather S, and B; S, is a daughter of D, in (A2). Here is a formulation of the necessary revision of (R2) to accommodate the interpretation of affixed structures for multiply center-embedding constructions; the revision of (R3) is similar.

\[(R2) \text{ (Revised) Suppose that (i) } K \text{ is an affix that is not immediately followed by a terminal symbol, (ii) } K \text{ is followed by affixes } L_1, \ldots, L_m \text{ in that order (} m \geq 1\text{), and (iii) } K \rightarrow L_1, \ldots, L_m \text{ is a rule of the original grammar. Then either the shortest substring following } K \text{ containing } L_1, \ldots, L_m \text{ that are not daughters of another constituent itself contains } K \text{ or it does not. Suppose it does not. Then } K \text{ may be a prefix whose daughters are those occurrences of } L_1, \ldots, L_m. \text{ Otherwise apply (R2) to the first repetition of } K \text{ in that string. Upon determination of its daughters, mark that occurrence of } K \text{ to distinguish it from the original one and reapply (R2) to the original } K. \]

We may suppose that this complication in the application of rules (R2) and (R3) accounts in large measure for the difficulty people have in the processing of multiply center-embedding constructions.

**Ambiguous Noncenter-Embedding Grammars**

Ambiguity, both lexical and structural, presents other kinds of difficulty for the determination of constituent structures by means of finite-state processes that operate from left to right. First, consider the ambiguous noncenter-embedding grammar (G3) that also generates (L1), but in such a way that the final string of b's together with the medial c forms a constituent (as in the case of expressions generated by (G1), the initial string of a's together with the medial c forms a constituent, or a proper substring of a's together with the medial c forms a constituent. For convenience, we refer to the language generated by (G3) as (L3).

\[(L3) \ a^n \ c \ b^n : m, \ n > 0 \]

\[(G3) \ a. \ S \rightarrow A \ S \\
\quad b. \ S \rightarrow C \\
\quad c. \ S \rightarrow D \ B \\
\quad d. \ C \rightarrow C \ B \\
\quad e. \ D \rightarrow D \ B \\
\quad f. \ D \rightarrow E \\
\quad g. \ E \rightarrow A \ E \\
\quad h. \ A \rightarrow a \\
\quad i. \ B \rightarrow b \\
\quad j. \ C \rightarrow c \\
\quad k. \ E \rightarrow c \]

\[(G3)* \text{ is another noncenter-embedding grammar satisfying (P1) and (P2) that generates affixed structures for the expressions in (L3) with respect to (G3).}

\[(G3)* \ a. \ S^* \rightarrow S \ A^* \ S^* \\
\quad b. \ S^* \rightarrow C^* \ S \\
\quad c. \ S^* \rightarrow D^* \ B^* \ S \\
\quad d. \ C^* \rightarrow C^* \ B^* \ C \\
\quad e. \ D^* \rightarrow D^* \ B^* \ D \\
\quad f. \ D^* \rightarrow D \ E^* \\
\quad g. \ E^* \rightarrow E \ A^* \ E^* \\
\quad h. \ A^* \rightarrow A \ a \\
\quad i. \ B^* \rightarrow B \ b \\
\quad j. \ C^* \rightarrow C \ c \\
\quad k. \ E^* \rightarrow E \ e \]

\[(G3)* \text{ generates several affixed structures for the expression (E1) (repeated here as (E3)), including (A1) (repeated here as (A3a)) and (A3b).}

\[(E3) \ a \ a \ c \ b \ b \]

\[(A3a) \ S \ A \ a \ S \ A \ a \ c \ b \ C \ b \ C \ B \ c \ S \\
\quad (A3b) \ D \ E \ A \ a \ E \ a \ e \ C \ b \ D \ b \ D \ b \ S \]

Tree-diagram representations of the constituent structures of (E3) with respect to (G3) are shown in (T3). It is not difficult to verify that (A3a) and (A3b) represent those constituent structures when interpreted according to rules (R1)–(R3). Nor is it difficult to see how a finite-state parser for (G3) could assign either of these affixed structures (and their interpretations) to (E3).

![Tree-diagram for (T3a)]
An affix grammar for (G4) that satisfies (P1) and (P2) is given in (G4)*.

\[
\begin{align*}
(G4)^* \text{ a. } & S^* \rightarrow S A^* F^* (E^*) \quad \text{ e. } B^* \rightarrow B b \\
& F^* \rightarrow F B^* C^* (E^*) \quad \text{ f. } C^* \rightarrow C c \\
& E^* \rightarrow E D^* E^* \quad \text{ g. } D^* \rightarrow D d \\
& A^* \rightarrow A a \quad \text{ h. } E^* \rightarrow E e
\end{align*}
\]

The expression (E4) is ambiguous with respect to (G4). Tree representations of its two structural descriptions are given in (T4a) and (T4b).

\[
\begin{align*}
\text{(T4a)} \quad \text{(T4b)}
\end{align*}
\]

Nevertheless, (G4)* generates only one affixed structure corresponding to (E4), namely (A4).

\[
(A4) \quad S A a F B b C c E d E e
\]

Indeed, for any ambiguous expression in (L4), (G4)* generates only one affixed structure. In other words, (G4)* provides what seems to be an impoverished basis for determining the constituent structures of the ambiguous expressions of (L4).

It may be thought that the difficulty lies not with affixed-structure notation, but rather with the particular choice of affix grammar to generate the structural descriptions of the expressions of (L4) with respect to (G4). For example, the affix grammar (G4X)* does generate distinct affixed structures for the distinct interpretations of ambiguous expressions in (L4).

\[
\begin{align*}
(G4X)^* \text{ a. } & S^* \rightarrow A^* F^* S \quad \text{ f. } A^* \rightarrow A a \\
& b. S^* \rightarrow S A^* F^* E^* \quad \text{ g. } B^* \rightarrow B b \\
& F^* \rightarrow B^* C^* F \quad \text{ h. } C^* \rightarrow C c \\
& E^* \rightarrow F B^* C^* E^* \quad \text{ i. } D^* \rightarrow D d \\
& A^* \rightarrow A a \quad \text{ j. } E^* \rightarrow E e
\end{align*}
\]
(G4X)* generates the affixed structures (A4Xa) and (A4Xb), corresponding to the tree structures (T4a) and (T4b) respectively.

(A4Xa) A a F b c c E₁, D d E₂, e S
(A4Xb) S a a B c c F E₁, D d E₂, e

However, the decision to introduce the affixes S and F as postfixes in rules (G4X)*c and as prefixes in rules (G4X)*b,d is arbitrary. Furthermore, the method that we develop in the next main section below for determining whether to introduce an affix as a prefix or a postfix in a given rule of an affix grammar precludes the possibility of affix grammars like (G4X)*.

In fact, the single affixed structure A₄ generated by (G4)* does provide an adequate basis for determining the structural descriptions of (E4) with respect to (G4), since (R1)–(R3) may be used to provide two distinct interpretations for (A₄) corresponding to the tree diagrams in (T4). The prefix S, by (R2), may have as daughters either the affixes A and F or the affixes A, F and E₁. Similarly, the prefix F, by (R2), may have as daughters either the affixes B and C or the affixes B, C and E₁. In order for (A₄) as a whole to have an interpretation with respect to (R1)–(R3), either S or F must take E₁ as its right daughter, but not both.

**Ambiguous Center-Embedding Grammars**

A somewhat different, but related, problem arises in the case of certain ambiguous center-embedding phrase-structure grammars such as (G₅). This grammar generates a language (L₅), which we do not attempt to characterize explicitly.

(G₅) a. S → A D
b. S → C B
c. C → S A
d. D → B S

(G₅)* is another center-embedding grammar satisfying (P1) and (P2) (with the constant c = 0) that generates affixed structures for the expressions of (L₅).

(G₅)* a. S* → A D*
b. S* → C* B S
c. C* → S A C
d. D* → D B S

Among the expressions of (L₅) are (E₅a) and (E₅b). The affixed structures for these expressions are respectively (A₅a) and (A₅b), both of which are generated by (G₅)* without center embedding.

(E₅a) A a B d
(E₅b) c a b
(A₅a) S a A D B b S a A a D D
(A₅b) C C B b S A a C B b S

It is easily verified that these affixed structures may be interpreted by (R1)–(R3) to represent the constituent structures of (E₅a)–(E₅b) with respect to (G₅).

Now consider the expression (E₅c), which is generated by (G₅) with one degree of center embedding.

(E₅c) a b a d a b

(E₅c) is ambiguous with respect to (G₅). Tree representations of its two structural descriptions are given in (T₅ci) and (T₅cii).

Nevertheless, (G₅)* generates only one affixed structure for (E₅c), namely that given in (A₅c).

(A₅c) S a a D B b S a a D d A a C B b S

Thus (G₅)* is like (G₄)* in generating impoverished representations of the structural descriptions of expressions with respect to the grammar on which it is based. Like (G₄)*, (G₅)* may be revised so as to distinguish among the various structural descriptions of ambiguous expressions of (L₅); in partic-
ular, if (G5)* were revised so as to introduce only prefixes or only postfixes (with the exception of the prefixes used to categorize terminal elements of (G5)), the resulting grammars would succeed in doing so. However, such grammars would violate principle (P2), since there would be no bound on the degree of center embedding of derivations of structural descriptions of expressions of (L5) with respect to them. Hence, (G5)* cannot be improved upon if principle (P2) is to be preserved.

Once again, we are left with what seems to be an inadequate basis for representations of the grammatical structures of expressions like (E5c) with respect to (G5). However, (A5c), like (A4), may be interpreted in two different ways according to (R1)–(R3), corresponding to the two structural descriptions of (E5c) with respect to (G5). The two interpretations of (A5c) (repeated here for convenience, with subscripts on the affixes to indicate their order of appearance from left to right) provided by (R1)–(R3) are given in (I5ci) and (I5cii).

(A5c) S1, A, a D1, B, b S2, A2 a D2, A3 a C, B3 b S3

(I5ci) Affix Type Daughters I-Rule G-Rule Step
S1 Pre A1, D1 R2 G5a 1
A1 Pre a R1 G5e 2
D1 Pre B1, S1 R2 G5d 3
B1 Pre b R1 G5f 4
S2 Pre A2, D2 R2 G5a 5
A2 Pre a R1 G5e 6
D2 Pre d R1 G5h 7
A3 Pre a R1 G5e 8
C1 Post S1, A2 R3 G5c 9*
B2 Pre b R1 G5f 10
S3 Post C1, B3 R3 G5b 11

* S1 cannot be selected as the left daughter of C1 because the former is selected as the right daughter of D1 in step 3.

(I5cii) Affix Type Daughters I-Rule G-Rule Step
S1 Pre A1, D1 R2 G5a 1
A1 Pre a R1 G5e 2
D1 Pre B1, S1 R2 G5d 3*
B1 Pre b R1 G5f 4
S2 Pre A2, D2 R2 G5a 5
A2 Pre a R1 G5e 6
D2 Pre d R1 G5h 7
A3 Pre a R1 G5e 8
C1 Post S2, A3 R3 G5c 9
B2 Pre b R1 G5f 10
S3 Post C1, B3 R3 G5b 11

* S1 cannot be selected as the right daughter of D1 because the former is selected as the left daughter of C1 in step 9.

It will be observed that the two interpretations of (A5c) depend on whether D1 or C1 obtains S1 as a daughter; whichever one does not obtains S2 as a daughter instead. (Compare steps 3 and 9 in each interpretation.) The two affixes D1 and C1 can be thought of as 'competing' for S2 as a daughter, since the latter is the nearer one to each affix and is thereby the more accessible to each by the interpretive rules. Accordingly, (A5c), as interpreted by (R1)–(R3), does represent the distinct grammatical structures of (E5c); the ambiguity turns on precisely how rules (R2) and (R3) are applied to determine what the daughters of D1 and C1 are in (A5c).

THE CONSTRUCTION OF AFFIX GRAMMARS FOR NATURAL LANGUAGES

In the second section, above, we pointed out that there is no general, effective procedure for constructing an affix grammar G* meeting (P2) for an arbitrary context-free phrase-structure grammar G. However, if affix grammars are psychologically real; that is, if people use them to assign constituent structures from left to right, in the form of affixed structures, to the expressions they hear or produce in the languages they know, we may be able to infer certain procedures for constructing them from linguistically significant phrase-structure grammars.

For simplicity, suppose that the acoustical or visual signals that transmit linguistic expressions are segmented into words and significant parts of words (the elements of the terminal vocabulary of the grammar). Suppose also that, as each word or word part is recognized, it is immediately assigned to a category, say W, or simultaneously to several categories, if it is lexically ambiguous. (Henceforth, suppose also, for simplicity, that each word or word part is categorically unambiguous.) This suggests that, as in the illustrative examples in above, the category W is assigned as a prefix to the word or word part in the development of the affixed structure.

Next, we suppose that higher-level categorial decisions are also made as quickly as possible (Langendoen and Langsam 1984). In particular, if L can be contextually determined to be the left daughter of another constituent K, then K is prefixed to L. Similarly, if K is contextually determined to be the left daughter of J, distinct from both K and L, then J is prefixed to K, and so on, until no more prefixes can be assigned. On the other hand, if L is the left daughter of K, but cannot be so determined contextually, then K is postfixed to its right daughter, and similarly for J.

While this method of assigning prefixes and postfixes from left to right seems well motivated, it is not guaranteed to yield affixed structures that can be generated by affix grammars meeting (P2). Fortunately, to obtain this result, we need only make certain relatively noncontroversial assumptions about the nature of phrase-structure systems of grammar for natural languages.

Suppose that (i) $K \rightarrow k$, or (ii) $K \rightarrow L_1 \ldots L_n$ are the rule types in the grammar of a language. Suppose we also adopt the conventions of the X-bar theory of constituent structure (Chomsky 1970; Jackendoff 1977), with the following modifications. First, we make no stipulation regarding the maximum number of bars in the projections of lexical categories. Second, we let the
symbol XP (where X is a lexical category) stand for the maximal projection of the lexical categories N (noun), V (verb), A (adjective), and P (preposition). As a result of this convention, in the sample grammar (G6) below, the symbols VP and PP denote V and P, respectively, but NP denotes N. Third, we limit the term 'specifier' to a lexically closed class of elements, which may be thought of as 'predictors' of the head, and we permit a specifier to appear either to the left or to the right of the head. Accordingly, we consider the article the to be a specifier in the English noun phrase the child, but the possessive phrase the child's in the child's teacher not to be. Similarly, we consider the adverb yesterday to be a specifier in the English verb phrase saw the child yesterday. With respect to any given rule of grammar, one of the three mutually exclusive and exhaustive cases in (C1)–(C3) holds.

(C1) L, is the head of K with one less bar than K or the daughter of K is K, which is its lexical head.
(C2) L, is a specifier of K.
(C3) Neither (C1) nor (C2) holds.

Now let x be an expression derived from K in the grammar (i.e., K \rightarrow x in G). If (C1) holds, then the left branch under K may be expected to descend nonrecursively to the lexical head K, as in the English verb phrase saw the child and prepositional phrase on the shelf (Huang 1982). By the procedure just sketched for the single-pass, left-to-right computation of affixed structures, the chain of categories on the left branch from K to the lexical head or specifier of K will appear as a string of prefixes in the affixed structure representing the constituent structure of x. Similarly, if (C2) holds, then since L is a lexical category which signals the beginning of K, both K and L will appear as prefixes in the affixed structure representing the constituent structure of x, as in the English noun phrase the child.

Finally, in case (C3), K will normally appear as a suffix in the affixed structure representing the constituent structure of x, since the first element in x will not provide sufficient evidence that x is a constituent of category K, as in the English sentence the child saw the teacher. (The left descendent of this sentence, the article the, does not provide evidence that a sentence has been initiated, but only that its subject noun phrase the child has.) However, in the immediate context in which x appears, there may be sufficient evidence to lead one to suppose that x is a constituent of category K, as in the subordinate clause of the English sentence the principal knew that the child saw the teacher, in which the complementizer that signals that a clause follows. In such cases, K may appear as a prefix in the affixed structure representing the constituent structure of x.

Now suppose that K is a recursive but noncenter-embedding category in the derivation of x. If either (C1) or (C2) holds, then K recurs under the right daughter of K and therefore is a right-branching category; and the corresponding category in the derivation of the affixed structure representing the structural description of x is also right branching. For example, NP is a recursive noncenter-embedding category in the derivation of the English phrase the friend of the child, and since (C1) holds (the left daughter of NP in this case is D, which dominates the specifier the), NP is right branching. Similarly, the derivation of the affixed structure corresponding to this phrase is right branching, since only prefixes are used.

(C3) is more problematic. First, if L is the right daughter of K) is a specifier of K, then K recurs under L and hence is left branching; and the corresponding category in the derivation of the affixed structure representing the constituent structure of x is also left branching, assuming that there is no immediate external evidence that x is a constituent of category K. An example is the English verb phrase saw the child there yesterday, in which each adverb can be construed as a specifier of the contained verb phrase to its left. Second, if L is the head of K, and L branches nonrecursively to the lexical head of K, then K recurs under L, and hence is left branching, like the corresponding category in the derivation of the affixed structure representing the constituent structure of x. An example is the English noun phrase the child's teacher's friend, in which the heads of the noun phrases in its derivation appear as right daughters of those phrases. Third, if L is the head of K, and L contains a complement or modifier within which the category K is introduced as a right daughter, then K is right-branching; and if it appears as a postfix in affixed structures, then the corresponding category in the derivation of the affixed structure representing the structural description of x is center embedding. An example is the English sentence the child said that the teacher knew that the principal hated the superintendent, in which the category S, which introduces the main and both subordinate clauses is right branching. To limit the degree of center embedding in the derivation of affixed structures in this case, let us assume that for all but at most the first occurrence of a string of category K on such a right branch, there is immediate contextual evidence that that string is a constituent of category K, so that from that point the corresponding derivation of the affixed structure is also right branching.

With these assumptions, we can now state a general method that satisfies both (P1) and (P2) for constructing an affix grammar G* from a phrase-structure grammar G that conforms to the principles of the theory of X-bar syntax as far described. If the derivation of an expression by G is noncenter embedding, then the derivation by G* of the corresponding affixed structure has at most first degree of center embedding. The method has four parts, which we label (M1)–(M4).

(M1) \( K \rightarrow k \) is a rule of G. Then \( K^* \rightarrow K \) k is a rule of G*.

(M2) \( K \rightarrow L_1, \ldots, L_m \) is a rule of G, and either C1 or C2 holds. Then \( K^* \rightarrow K L^*_1, \ldots, L^*_m \) is a rule of G* (where \( \ldots \) is a string of nonterminals in G* corresponding to the string of nonterminals in the associated rule of G).

(M3) \( K \rightarrow L_1, \ldots, L_m \) is a rule of G, C3 holds, and \( L_m \) is a specifier of K. Then \( K^* \rightarrow L^*_1, \ldots, L^*_m K \) is a rule of G*.

(M4) \( K \rightarrow L_1, \ldots, L_m \) is a rule of G, C3 holds, and \( L_m \) is the head of K. Then:
   a. \( K^* \rightarrow K L^*_1, \ldots, L^*_m \) is a rule of G* for those left contexts that provide evidence for the occurrence of K.
   b. \( K^* \rightarrow L^*_1, \ldots, L^*_m K \) is a rule of G for those left contexts that do not provide such evidence.

We now apply this method to construct an affix grammar that generates the affixed structures that represent the constituent structures of expressions
generates by the phrase-structure grammar (G6). (A slightly different version of this grammar is given in Langendoen and Langsam 1984.) The notation we use follows the conventions given in this section. The categories used in (G6) are, for the most part, commonplace within the conventions of the X-bar theory; however, C, rather than S, is used to represent the category of a complement clause, on the assumption that the complementizer C is the head of that category; and G is used to represent the category of the genitive (or possessive) modifier of a noun. The head of G is G, the category of the genitive ending's. ADS represents the class of sentence adverbs, as opposed to ADV, which represents the class of verb-phrase adverbs.

(G6) a. S → S ADS
   b. S → (NP, C) VP
   c. VP → VP ADV
   d. VP → V (NP) ((PP, C))
   e. NP → Q NP
   f. NP → (D) N
   g. C → C S
   h. G → NP G
   i. N → N (PP)
   j. PP → P NP

(G6) generates a language (L6) consisting of an infinite subset of English expressions with both right and left branching and with center embedding.

Next, in (G6)*, we give the rules of an affix grammar that generates the affixed structures for the expressions generated by (G6) and that satisfies (P1) and (P2) with c = 1; that is, the degree of center embedding of an affix structure is at most one greater than the degree of center embedding of the corresponding sentence.

(G6)* a. S* → S* ADS* S
   b. i. S* → S (NP, C *) VP* / S that ______
      ii. S* → (NP*, C *) VP* S / elsewhere
   c. VP* → VP* ADV* VP
   d. VP* → V (NP*) ((PP*, C *))
   e. NP* → NP Q* NP*
   f. NP* → NP (D*) N *
   g. NP* → G * N * NP
   h. C * → C C* S *
   i. G * → NP* G* G
   j. N * → N N* (PP*)
   k. PP* → PP P* NP *
   l. V * → V {knew, . . .}
   m. N* → N {child, . . .}
   n. C* → C that
   o. D* → D the
   p. G* → G's
   q. P* → P {of, . . .}

Given the affix grammar (G6)* just presented, one can construct a procedure which assigns affixed strings to expressions from left to right and which behaves as a finite transducer for all expressions drawn from (L6). An elegant implementation of such a procedure has been written in PL/1 by Maria Edelstein and Ethel Fisch.

Rather than analyze examples with respect to this grammar here, we call the reader's attention to our paper (Langendoen and Langsam, 1987), which gives a detailed analysis of examples generated by a similar grammar. Because of publishing delays, that paper appeared before this one, even though it was written earlier; it provides a somewhat different analysis of the problem, but with the same result, namely the ability of a finite transducer to parse sentences up to a fixed finite degree of center embedding. The transducer also "loses" certain structural information about expressions with center embedding, much like the affixed string notation developed here.

SUMMARY

The class of acceptable expressions (with no restrictions on length) of a natural language appears to form a regular set. The class of grammatical structures that corresponds to that class of expressions, however, does not, if those structures are represented in tree or bracketing form. We propose a parenthesis-free method of representing grammatical structures which permits the class that corresponds to the acceptable expressions of a natural language to constitute a regular set. The results of this method can be considered the basis
of a viable model of human representation of grammatical structure for two reasons. First, the grammatical structures of acceptable expressions may be computed in tandem with the processing of those expressions from left to right; and, second, this computation can itself be carried out by a finite automaton. We apply this method to a fragment of English and show how it works for that fragment. A portion of this application has been implemented in PL/I.

REFERENCES

CHOMSKY, N.

Fodor, J. A., T. Bever, and M. Garrett

HUANG, J. C. T.

Jackendoff, R. S.

Langendoen, D. T.

Langendoen, D. T. and Y. Langsam

Miller, G. A. and N. Chomsky