Reprinted from The Uses of Linguistics
Volume 583 of the *Annals of the New York Academy of Sciences*
May 25, 1990

A New Method of Representing Constituent Structures[a]

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

[a] 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 (Langendoen 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) $[_S$ the principal believed that $[_S$ the teacher knew that
$[_S$ the student thought that $[_S$ the alarm meant that
$[_S$ school was out $]_S$ $]_S$ $]_S$ $]_S$ $]_S$

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 $A \rightarrow X$ or $A \rightarrow a$, where A is a category, X is a nonnull string of categories, and a

is a terminal element (lexical item). Let L (n) be the subset of L of expressions whose derivations with respect to G have no more than degree n of center embedding. Finally, let C(L) be the set of constituent structures of the members of L and C(L(n)) be the subset of C(L) of expressions of degree n or less of center embedding. We seek a method of representing the constituent structures C(L) of the members of L which has two properties. First, the elements of C(L) should be determinable from L; and second, the degree of center embedding for each element of C(L) should be no more than a fixed, nonnegative constant c greater than that of the corresponding element of L. We restate these properties here as (P1) and (P2).

(P1) C(L) can be directly generated by another phrase-structure grammar G* constructible from G.
(P2) The derivations of the structures C(L(n)) associated with the members of L(n) with respect to G* have no more than degree $n + c$ ($c \geq 0$) of center embedding.

While the method which makes use of full bracketing structures has property (P1), it lacks property (P2). A grammar G* that directly generates the set of full bracketing structures associated with the expressions of L can be obtained from G by replacing each rule of G of the form $A \rightarrow x$ by the rule $A \rightarrow [_A\ x\ ]_A$. Now suppose that L(0) is an infinite set of expressions with left or right branching (by definition, the expressions in L(0) lack center embedding). Then there is no number m such that derivations of all elements of C(L(0)) with respect to G* have less than degree m of center embedding. For given any proposed m, the derivation of the full bracketing structure of any expression of L with degree m of right or left branching will be found to have degree m of center embedding with respect to G*.

Nevertheless, there are representations of grammatical structures that enable both (P1) and (P2) to be satisfied. Such representations are obtained from full bracketing structures by omitting right brackets from right-branching structures and left brackets from left-branching structures. In right-branching structures, the positions of the left brackets that remain indicate where each constituent begins, and the position where each constituent ends can be determined by a finite-state procedure. Conversely, in left-branching structures, the positions of the right brackets that remain indicate where each constituent ends, and the position where each constituent begins can also be determined by a finite-state procedure. Moreover, since brackets no longer occur in matched pairs, the brackets themselves can be omitted, leaving only the category labels. In left-branching structures, these category symbols occur as postfixes; in right-branching structures they occur as prefixes. We call any symbol which occurs as a prefix or postfix in a string that represents the constituent structure of an expression an 'affix', and such strings we call 'affixed structures'. Finally, we call grammars that directly generate the affixed structures for the expressions of a given language an 'affix grammar'.

We cannot provide a fully general method for constructing an affix grammar G* for an arbitrary phrase-structure grammar G that generates a language L, such that (P2) is satisfied, since such a method would depend on our being able to determine which of the productions of G result in left

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 G1 be the rules of an unambiguous noncenter-embedding grammar that generates the language (L1). Note that since (G1) is noncenter embedding, (L1) = (L1(0)).

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

(G1) a. $S \rightarrow A\ S$ d. $A \rightarrow a$
b. $S \rightarrow C$ e. $B \rightarrow b$
c. $C \rightarrow C\ B$ f. $C \rightarrow c$

(G1)* is another noncenter-embedding grammar, constructible from (G1), which is an affix grammar for (L1). Since (G1)* is constructible from (G1) 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 (G1)* is S*.

(G1)* a. $S^* \rightarrow S\ A^*\ S^*$ d. $A^* \rightarrow A\ a$
b. $S^* \rightarrow C^*\ S$ e. $B^* \rightarrow B\ b$
c. $C^* \rightarrow C^*\ B^*\ C$ f. $C^* \rightarrow C\ c$

Among the expressions generated by (G1) is (E1); the affixed structure generated by (G1)* that corresponds to (E1) is (A1). (The subscripts 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) $S_1\ A_1\ a\ S_2\ A_2\ a\ C_1\ c\ B_1\ b\ C_2\ B_2\ b\ C_3\ B_3\ b\ C_4\ S_3$

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 is a prefix whose sole daughter is that occurrence of k.

(R2) Suppose that (i) K is an affix that is not immediately followed by a terminal symbol, (ii) K is followed by affixes $L_1, \ldots, L_m$ in that order ($m \geqslant 1$), and (iii) $K \rightarrow L_1 \ldots L_m$ is a rule of the original grammar. Then K may be a prefix whose daughters are those occurrences of $L_1, \ldots, L_m$.

(R3) Suppose that (i) K is an affix that is not immediately followed by a terminal symbol, (ii) K is preceded by affixes $L_1, \ldots, L_m$ in that order ($m \geqslant 1$), and (iii) $K \rightarrow L_1 \ldots L_m$ is a rule of the original grammar. Then K may be a postfix whose daughters are those occurrences of $L_1, \ldots, L_m$.

The structure (A1) is interpreted from left to right according to (R1)–(R3) as in (I1). (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.)

(I1)

Affix	Type	Daughters	I-Rule	G-Rule	Step
S_1	Pre	A_1, S_2	R2	G1a	1
A_1	Pre	a	R1	G1d	2
S_2	Pre	A_2, S_3	R2	G1a	3
A_2	Pre	a	R1	G1d	4
C_1	Pre	c	R1	G1f	5
B_1	Pre	b	R1	G1e	6
C_2	Post	C_1, B_1	R3	G1c	7
B_2	Pre	b	R1	G1e	8
C_3	Post	C_2, B_2	R3	G1c	9
B_3	Pre	b	R1	G1e	10
C_4	Post	C_3, B_3	R3	G1c	11
S_3	Post	C_4	R3	G1b	12

With the interpretation given in (I1), (A1) represents the constituent structure of (E1) with respect to (G1).

Finally, note that not only can the affixed structures generated by (G1)* be recognized by a finite automaton, but the interpretations of those structures as representations of the structural descriptions of the expressions generated by (G1) 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 (G1)* 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) $\{a^n\ c\ b^n : n \geqslant 0\}$

(G2) a. $S \rightarrow A\ D$ c. $A \rightarrow a$
b. $D \rightarrow S\ B$ 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. $S^* \rightarrow S\ A^*\ D^*$
b. $D^* \rightarrow S^*\ B^*\ D$
c. $A^* \rightarrow A\ a$
d. $B^* \rightarrow B\ b$
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) $S_1\ A_1\ a\ S_2\ A_2\ a\ S_3\ c\ B_1\ b\ D_1\ B_2\ b\ D_2$

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_1 in (A2) has as its daughters not the next following affixes A_1 and D_1, but rather A_1 and D_2; D_1 is a daughter of S_2 in (A2). Similarly, the postfix D_2 in (A2) has as its daughters not the next preceding affixes S_3 and B_2, but rather S_2 and B_2; S_3 is a daughter of D_1 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) (Revised) Suppose that (i) K is an affix that is not immediately followed by a terminal symbol, (ii) K is followed by affixes $L_1, \ldots, L_m$ in that order ($m \geq 1$), and (iii) $K \rightarrow L_1 \ldots L_m$ is a rule of the original grammar. Then either the shortest substring following K containing $L_1, \ldots, L_m$ that are not daughters of another constituent itself contains K or it does not. Suppose it does not. Then K may be a prefix whose daughters are those occurrences of $L_1, \ldots, L_m$. Otherwise apply (R2) to the first repetition of K in that string. Upon determination of its daughters, mark that occurrence of K 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) $\{a^m\ c\ b^n : m, n \geq 0\}$

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

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

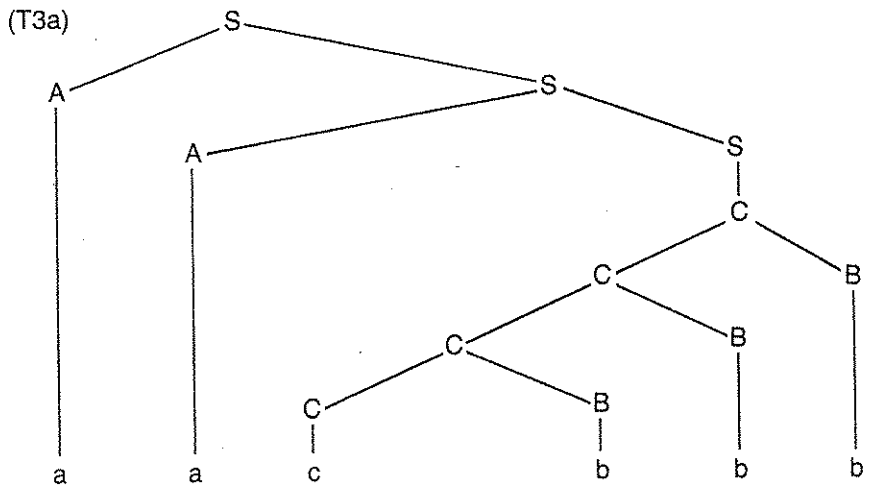
(G3)* a. $S^* \rightarrow S\ A^*\ S^*$
b. $S^* \rightarrow C^*\ S$
c. $S^* \rightarrow D^*\ B^*\ S$
d. $C^* \rightarrow C^*\ B^*\ C$
e. $D^* \rightarrow D^*\ B^*\ D$
f. $D^* \rightarrow D\ E^*$
g. $E^* \rightarrow E\ A^*\ E^*$
h. $A^* \rightarrow A\ a$
i. $B^* \rightarrow B\ b$
j. $C^* \rightarrow C\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

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

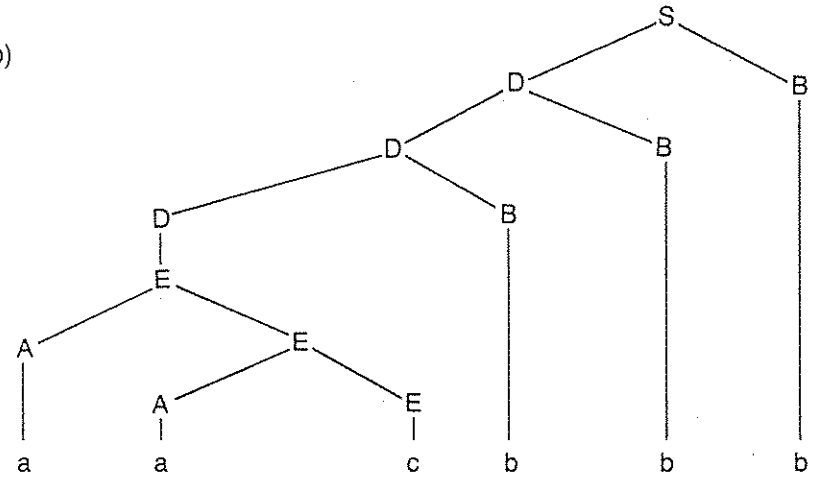
(E3) a a c b b b
(A3a) S A a S A a C c B b C B b C B b C S
(A3b) D E A a E A a E c B b D B b D B b S

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).

(T3a)



(T3b)



However, such a parser, if not carefully designed, could have difficulty with expressions in (L3) of the form a^m c; that is, expressions in which the string of b's is null. Such expressions are unambiguous with respect to (G3), having no derivation in which the string of a's together with c are analyzable as the constituent D. Nevertheless, a left-to-right parser cannot determine for certain that those expressions have no such analysis until the end of the expression is reached. (The only expressions in (L3) which are unambiguous with respect to (G3) are those in which the final string of b's is null.) If a parser should postulate at the beginning of its processing of expressions of this type that it begins with the constituent D, it would not discover its error until the end of the expression is reached (and consequently it would be seriously 'garden pathed'). On the other hand, if a parser never postulates that an expression of (L3) begins with the constituent D, it would never discover that expressions in which the string of b's is nonnull are ambiguous.

Consequently, an effective left-to-right parser for (L3) would have to assume at the outset that the expression it is processing is ambiguous, and only abandon that assumption when the end of the expression is reached, if it is discovered that the string of b's is null. In principle, no amount of look-ahead is sufficient to resolve the ambiguity, since the initial string of a's can be of any finite length whatever, and under neither analysis can any part of that string be grouped into subconstituents until after the element c is reached.

Next, consider the ambiguous noncenter-embedding grammar (G4), which generates the language (L4).

(L4) $\{a\ b\ c\ (d^m\ e)\ (d^n\ e) : m, n \geq 0\}$

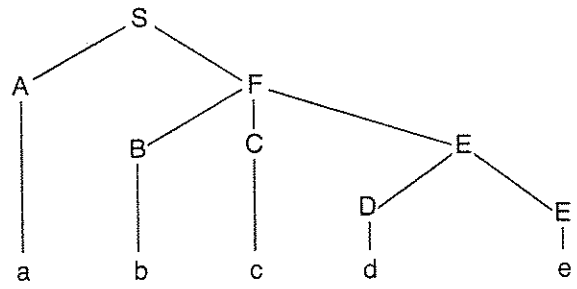
(G4) a. S → A F (E)
b. F → B C (E)
c. E → D E
d. A → a
e. B → b
f. C → c
g. D → d
h. E → e

An affix grammar for (G4) that satisfies (P1) and (P2) is given in (G4)*.

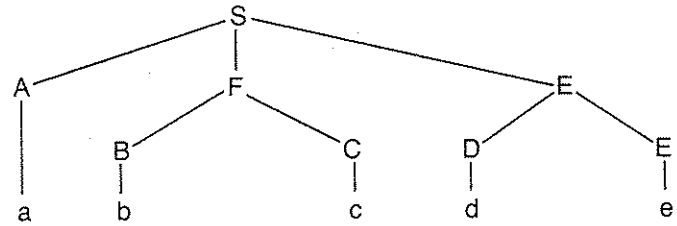
(G4)* a. S* → S A* F* (E*)
b. F* → F B* C* (E*)
c. E* → E D* E*
d. A* → A a
e. B* → B b
f. C* → C c
g. D* → D d
h. E* → E e

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

(T4a)



(T4b)



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

(A4) S A a F B b C c E_1 D d E_2 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).

(G4X)* a. S* → A* F* S
b. S* → S A* F* E*
c. F* → B* C* F
d. F* → F B* C* E*
e. E* → E D* E*
f. A* → A a
g. B* → B b
h. C* → C c
i. D* → D d
j. E* → E e

(G4X)* generates the affixed structures (A4Xa) and (A4Xb), corresponding to the tree structures (T4a) and (T4b) respectively.

(A4Xa) A a F B b C c E_1 D d E_2 e S
(A4Xb) S A a B b C c F E_1 D d E_2 e

However, the decision to introduce the affixes S and F as postfixes in rules (G4X)*a,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 E_1. Similarly, the prefix F, by (R2), may have as daughters either the affixes B and C or the affixes B, C and E_1. In order for (A4) as a whole to have an interpretation with respect to (R1)–(R3), either S or F must take E_1 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
e. A → a
f. B → b
g. C → c
h. D → d

(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* → S A* D*
b. S* → C* B* S
c. C* → S* A* C
d. D* → D B* S*
e. A* → A a
f. B* → B b
g. C* → C c
h. D* → D d

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

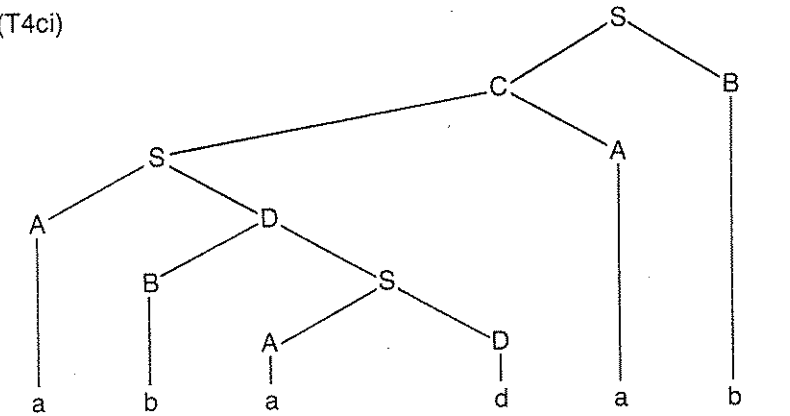
It is easily verified that these affixed structures may be interpreted by (R1)–(R3) to represent the constituent structures of (E5a)–(E5b) with respect to (G5).

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

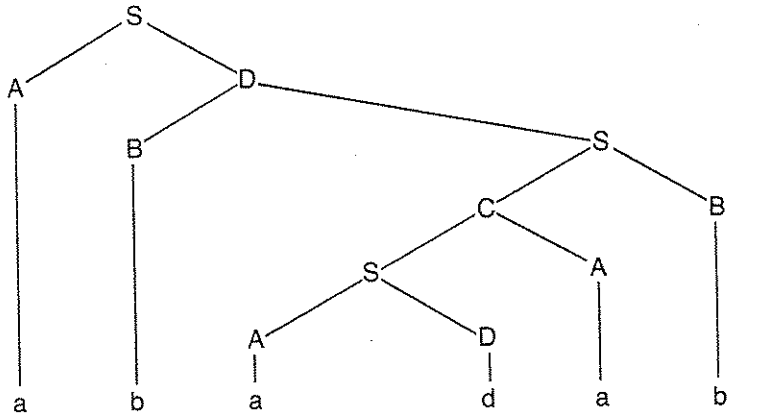
(E5c) a b a d a b

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

(T4ci)



(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 $(L5)_0$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) S_1 A_1 a D_1 B_1 b S_2 A_2 a D_2 d A_3 a C_1 B_2 b S_3

(I5ci)

Affix	Type	Daughters	I-Rule	G-Rule	Step
S_1	Pre	A_1, D_1	R2	G5a	1
A_1	Pre	a	R1	G5e	2
D_1	Pre	B_1, S_2	R2	G5d	3
B_1	Pre	b	R1	G5f	4
S_2	Pre	A_2, D_2	R2	G5a	5
A_2	Pre	a	R1	G5e	6
D_2	Pre	d	R1	G5h	7
A_3	Pre	a	R1	G5e	8
C_1	Post	S_1, A_3	R3	G5c	9*
B_2	Pre	b	R1	G5f	10
S_3	Post	C_1, B_2	R3	G5b	11

*S_2 cannot be selected as the left daughter of C_1 because the former is selected as the right daughter of D_1 in step 3.

(I5cii)

Affix	Type	Daughters	I-Rule	G-Rule	Step
S_1	Pre	A_1, D_1	R2	G5a	1
A_1	Pre	a	R1	G5e	2
D_1	Pre	B_1, S_3	R2	G5d	3*
B_1	Pre	b	R1	G5f	4
S_2	Pre	A_2, D_2	R2	G5a	5
A_2	Pre	a	R1	G5e	6
D_2	Pre	d	R1	G5h	7
A_3	Pre	a	R1	G5e	8
C_1	Post	S_2, A_3	R3	G5c	9
B_2	Pre	b	R1	G5f	10
S_3	Post	C_1, B_2	R3	G5b	11

*S_2 cannot be selected as the right daughter of D_1 because the former is selected as the left daughter of C_1 in step 9.

It will be observed that the two interpretations of (A5c) depend on whether D_1 or C_1 obtains S_2 as a daughter; whichever one does not obtains S_3 as a daughter instead. (Compare steps 3 and 9 in each interpretation.) The two

affixes D_1 and C_1 can be thought of as 'competing' for S_2 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 D_1 and C_1 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) $K \rightarrow L_1 \ldots L_m$ 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 $\overline{V}$ and $\overline{P}$, 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_1 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_1 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 preposition 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_1 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 postfix 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_m (the right daughter of K) is a specifier of K, then K recurs under L_1 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_m is the head of K, and L_m branches nonrecursively to the lexical head of K, then K recurs under L_1 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_m is the head of K, and L_m 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 so 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 . . . 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

generated 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, $\bar{C}$, rather than $\bar{S}$, is used to represent the category of a complement clause, on the assumption that the complementizer C is the head of that category; and $\bar{G}$ is used to represent the category of the genitive (or possessive) modifier of a noun. The head of $\bar{G}$ 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, $\bar{C}$} VP
c. VP → VP ADV
d. VP → V (NP) ({PP, $\bar{C}$})
e. NP → Q NP
f. NP → (D) $\bar{N}$
g. NP → $\bar{G}$ $\bar{N}$
h. $\bar{C}$ → C S
i. $\bar{G}$ → NP G
j. $\bar{N}$ → N (PP)
k. PP → P NP
l. V → {knew . . .}
m. N → {child, . . .}
n. C → that
o. D → the
p. G → 's
q. P → {of, . . .}
r. Q → {both, . . .}
s. ADS → {too, . . .}
t. ADV → {then, . . .}

(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*, $\bar{C}$ *} VP* / C that ____
ii. S* → {NP*, $\bar{C}$ *} VP* S / *elsewhere*
c. VP* → VP* ADV* VP
d. VP* → VP V* (NP*) ({PP*, $\bar{C}$ *})
e. NP* → NP Q* NP*
f. NP* → NP (D*) $\bar{N}$ *
g. NP* → $\bar{G}$ * $\bar{N}$ * NP
h. $\bar{C}$ * → $\bar{C}$ C* S*
i. $\bar{G}$ * → NP* G* $\bar{G}$
j. $\bar{N}$ * → $\bar{N}$ 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, . . .}
r. Q* → Q {both, . . .}
s. ADS* → ADS {too, . . .}
t. ADV* → ADV {then, . . .}

An explanation for the decision whether to introduce an affix as a prefix or as a postfix in each of the rules in (G6)* is given in the table in (X6). (The designations (C1)–(C3) and (M1)–(M4) refer to the cases and methods proposed earlier in this section.

(X6)

Rule	*Case*	*Method*	*Type*
(G6)*a	C3	M3	Post
(G6)*bi	C3	M4a	Pre
(G6)*bii	C3	M4b	Post
(G6)*c	C3	M3	Post
(G6)*d	C1	M2	Pre
(G6)*e	C1	M2	Pre
(G6)*f	C1	M2	Pre
(G6)*g	C3	M4b	Post
(G6)*h	C1	M2	Pre
(G6)*i	C3	M4b	Post
(G6)*j	C1	M2	Pre
(G6)*k	C1	M2	Pre
(G6)*l–t	C1	M1	Pre

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 $(L6)_0$. An elegant implementation of such a procedure has been written in PL/I 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 later; 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.

1959 A note on phrase structure grammars. *Information and Control* 2: 393–395.

1963 Formal properties of grammars. In *Handbook of mathematical psychology*, vol. 2, edited by R.D. Luce, R.R. Bush, and E. Galanter. Reading, MA: Wiley, pp. 323–418.

1965 *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

1970 Remarks on nominalization. In *Readings in English transformational grammar*, edited by R. Jacobs and P.S. Rosenbaum. Waltham, MA: Ginn, pp. 184–221.

1980 *Rules and representations*. New York: Columbia University Press.

FODOR, J. A., T. BEVER, AND M. GARRETT

1974 *The psychology of language*. New York: McGraw-Hill.

HUANG, J. C. T.

1982 *Logical relations in Chinese and the theory of grammar*. Ph.D. dissertation, Massachusetts Institute of Technology.

JACKENDOFF, R. S.

1977 *X-Bar Syntax*. Cambridge, MA: MIT Press.

LANGENDOEN, D. T.

1975 Finite-state parsing of phrase-structure languages and the status of readjustment rules in grammar. *Linguistic Inquiry* 6: 533–554.

LANGENDOEN, D. T. AND Y. LANGSAM

1984 The representation of constituent structures for finite-state parsing. *Proceedings of COLING84*, pp. 24–27.

1987 On the design of finite transducers for parsing phrase-structure grammars. In *Mathematics of language*, edited by A. Manaster-Ramer. Amsterdam and Philadelphia: John Benjamins, pp. 191–235.

MILLER, G. A. AND N. CHOMSKY

1963 Finitary models of language users. In *Handbook of mathematical psychology*, vol. 2, R. D. Luce, R. R. Bush and E. Galanter. Reading, MA: Wiley, pp. 419–491.