THE REPRESENTATION OF CONSTITUENT STRUCTURES FOR FINITE-STATE PARSING

D. Terence Langendoen
Yedidyah Langsam

Departments of English and Computer & Information Science
Brooklyn College of the City University of New York
Brooklyn, New York 11210 U.S.A.

ABSTRACT

A mixed prefix-postfix notation for representations of the constituent structures of the expressions of natural languages is proposed, which are of limited degree of center embedding if the original expressions are noncenter-embedding. The method of constructing these representations is applicable to expressions with center embedding, and results in representations which seem to reflect the ways in which people actually parse those expressions. Both the representations and their interpretations can be computed from the expressions from left to right by finite-state devices.

The class of acceptable expressions of a natural language L all manifest no more than a small, fixed, finite degree n of center embedding. From this observation, it follows that the ability of human beings to parse the expressions of L can be modeled by a finite transducer that associates with the acceptable expressions of L representations of the structural descriptions of those expressions. This paper considers some initial steps in the construction of such a model. The first step is to determine a method of representing the class of constituent structures of the expressions of L without center embedding in such a way that the members of that class themselves have no more than a small fixed finite degree of center embedding. Given a grammar that directly generates that class of constituent structures, it is not difficult to construct a deterministic finite-state transducer (parser) that assigns the appropriate members of that class to the noncenter-embedded expressions of L from left to right. The second step is to extend the method so that it is capable of representing the class of constituent structures of expressions of L with no more than degree n of center embedding in a manner which appears to accord with the way in which human beings actually parse those sentences. Given certain reasonable assumptions about the character of the rules of grammar of natural languages, we show how this step can also be taken.

*This work was partly supported by a grant from the PSC-CUNY Faculty Research Award Program.

Let G be a context-free phrase-structure grammar (CFPSG). First, suppose that the category A in G is right-recursive; i.e., that there are subderivations with respect to G such that A ==> X A, where X is a nonnull string of symbols (terminal, nonterminal, or mixed). We seek a new CFPSG G*, derived from G, that contains the category A* (corresponding to A), such that there are subderivations with respect to G* of the form A* ==> X* A*, where X* represents the constituent structure of X with respect to G. Next, suppose that the category B in G is left-recursive; i.e., that there are subderivations with respect to G such that B ==> B Y, where Y is nonnull. We seek a new CFPSG G*, derived from G, that contains the category B* (corresponding to B), such that there are subderivations with respect to G* of the form B* ==> B* Y*, where Y* represents the constituent structure of Y with respect to G. In other words, given a grammar G, we seek a grammar G* that directly generates strings that represent the constituent structures of the noncenter-embedded expressions generated by G, that is right-recursive wherever G is right-recursive and is left-recursive wherever G is left-recursive.

In order to find such a G*, we must first determine what kinds of strings are available that can represent constituent structures and at the same time can be directly generated by noncenter-embedding grammars. Full bracketing diagrams are not suitable, since grammars that generate them are center embedding whenever the original grammars are left- or right-recursive (Langendoen 1975). Suppose, however, that we leave off right brackets in right-recursive structures and left brackets in left-recursive structures. In right-recursive structures, the positions of the left brackets that remain indicate where each constituent begins; the position where each constituent ends can be determined by a simple counting procedure provided that the number of daughters of that constituent is known (e.g., when the original grammar is in Chomsky-normal-form). Similarly, in left-recursive 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 simply by counting. Moreover, since brackets no longer occur in matched pairs, the brackets themselves can be omitted, leaving only the category labels. In left-recursive structures, these category symbols occur as **postfixes**; in right-recursive structures,

they occur as **prefixes**. Let us call any symbol which occurs as a prefix or a postfix in a string that represents the constituent structure of an expression an **affix**; the strings themselves **affixed strings**; and the grammars that generate those strings **affix grammars**.

To see how affix grammars may be constructed, consider the noncenter-embedding CFPSG G1, which generates the artificial language L1 = a(b*a)*b*a.

(G1) a. S ⟶ S A b. A ⟶ B A

c. A ⟶ a d. B ⟶ b

e. S ⟶ a

A noncenter-embedding affix grammar that generates the affixed strings that represent the constituent structures of the expressions of L1 with respect to G1 is given in G1*.

(G1*) a. S* ⟶ S* A* S b. A* ⟶ A B* A*

c. A* ⟶ A a d. B* ⟶ B b

e. S* ⟶ S a

Among the expressions generated by G1 is E1; the affixed string generated by G1* that represents its structural description is E1*.

(E1) abbaba

(E1*) SaABbABbAaSABbAaS

Let us say that an affix **covers** elements in an affixed string which correspond to its constituents (not necessarily immediate). Then E1* may be interpreted as a structural description of E1 with respect to G1 according to the rules in R, in which J, K, and L are affixes; k is a word; x and y are substrings of affixed strings; and G is a CFPSG (in this case, G1).

(R) a. If K ⟶ k is a rule of G, then in the configuration ... K k ..., K is a prefix which covers k.

b. If J ⟶ K L is a rule of G, then in the configuration ... J K x L ..., in which x does not contain L, J is a prefix which covers K L.

c. If J ⟶ K L is a rule of G, then in the configuration ... K x L y J ..., in which x does not contain L and y does not contain K, J is a postfix which covers K L.

Coverage of constituents by the rules in R may be thought to be assigned dynamically from left to right.

A postfix is used in rule G1*a because the category S is left-recursive in G1, whereas a prefix is used in rule G1*b because the category A is right-recursive in G1. The use of prefixes in rules G1*c-e, on the other hand, is unmotivated if the only criteria for choosing an affix type have to do with direction of recursion. For affix grammars of natural languages, however, one can motivate the decision to use a particular type of affix by principles other than those having to do with direction of recursion.

The use of a prefix can be interpreted as indicating a decision (or guess) on the part of the language user as to the identity of a particular constituent on the basis of the identity of the first constituent in it. Since lexical items are assigned to lexical categories essentially as soon as they are recognized (Forster 1976), we may suppose first that prefixes are used for rules such as those in G1*c-e that assign lexical items to lexical categories. Second, if, as seems reasonable, a decision about the identity of constituents is always made as soon as possible, then we may suppose that prefixes are used for all rules in which the leftmost daughter of a particular constituent provides sufficient evidence for the identification of that constituent; e.g., if the leftmost daughter is either the specifier or the head of that constituent in the sense of Jackendoff (1977). Third, we may suppose that even if the leftmost daughter of a particular constituent does not provide sufficient evidence for the identification of that constituent, a prefix may still be used if that constituent is the left sister of a constituent that provides sufficient evidence for its identification. Fourth, we may suppose that postfixes are used in all other cases.

To illustrate the use of these four principles, consider the noncenter-embedding partial grammar G2 that generates a fragment of English that we call L2.

(G2) a. **S ⟶ NP VP** b. **NP ⟶ D $\bar{N}$**

c. **NP ⟶ $\bar{G}$ $\bar{N}$** d. **$\bar{G}$ ⟶ NP G**

e. **$\bar{N}$ ⟶ N** f. **VP ⟶ V ({NP, $\bar{C}$})**

g. **$\bar{C}$ ⟶ C S** h. **C ⟶ that**

i. **D ⟶ the** j. **G ⟶ 's**

k. **N ⟶ {boss, child ...}**

l. **V ⟶ {knew, saw, ...}**

Among the expressions of L2 are those with both right-recursion and left-recursion, such as E2.

(E2) the boss knew that the teacher's sister's neighbor's friend believed that the student saw the child

We now give an affix grammar G2* that directly generates affixed strings that represent the structural descriptions of the expressions of L2 with respect to G2, and that has been constructed in accordance with the four principles described above.

(G2*) a. i. $S^* \longrightarrow S\ NP^*\ VP^*$ / C that ___

ii. $S^* \longrightarrow NP^*\ VP^*\ S$ / elsewhere

b. $NP^* \longrightarrow NP\ D^*\ \bar{N}^*$

c. $NP^* \longrightarrow \bar{G}^*\ \bar{N}^*\ NP$

d. $\bar{G}^* \longrightarrow NP^*\ G^*\ \bar{G}$

e. $\bar{N}^* \longrightarrow \bar{N}\ N^*$

f. $VP^* \longrightarrow VP\ V^*\ (\{NP^*, \bar{C}^*\})$

g. $\bar{C}^* \longrightarrow \bar{C}\ C^*\ S^*$

h. $C^* \longrightarrow C$ that

i. $D^* \longrightarrow D$ the

j. $G^* \longrightarrow G$'s

k. $N^* \longrightarrow N\ \{$child, house, ...$\}$

l. $V^* \longrightarrow V\ \{$knew, saw, ...$\}$

Rules G2*h-l conform to the first principle, according to which lexical categories generally appear as prefixes. Rules G2*b,e-g conform to the second principle, according to which a category appears as a prefix if its leftmost daughter in the corresponding rule of G2 is its head or specifier. Rule G2*ai conforms to the third principle, according to which a category appears as a prefix if its presence can be predicted from its right sister in G2. Finally, rules G2*aii,c,d conform to the fourth principle, according to which a category appears as a postfix if it cannot appear as a prefix according to the preceding three principles.

The affixed string that G2* generates as the representation of the structural description of E2 with respect to G2 is given in E2*.

(E2*) NP D the $\bar{N}$ N boss VP V knew $\bar{C}$ C that S NP D the $\bar{N}$ N teacher G 's $\bar{G}$ $\bar{N}$ N sister NP G 's $\bar{G}$ $\bar{N}$ N neighbor NP G 's $\bar{G}$ $\bar{N}$ N friend NP VP V believed $\bar{C}$ C that S NP D the $\bar{N}$ N student VP V saw NP D the $\bar{N}$ N child S

E2* can be interpreted as the structural description of E2 with respect to G2 by the rules in R, with the addition of a rule to handle unary nonlexical branching (as in G2e), and a modification of Rc to prevent a postfix from simply covering a sequence of affixes already covered by a prefix. (This restriction is needed to prevent the postfix S in E2* from simply covering any of the subordinate clauses in that expression.) It is worth noting how the application of those rules dynamically enlarges the NP that is covered by the S prefix that follows the words knew that. First the teacher is covered; then the teacher's sister; then the teacher's sister's neighbor; and finally the teacher's sister's neighbor's friend.

The derivation of E2* manifests first-degree center embedding of the category S*, as a result of the treatment of S as both a prefix and a suffix in G2*. However, no derivation of an affixed string generated by G2* manifests any greater degree of center embedding; hence, the affixed strings associated with the expressions of L2 can still be assigned to them by a finite-state parser. The added complexity involved in interpreting E2* results from the fact that all but the first of the NP-VP sequences in E2* are covered by prefix Ss, so that the constituents covered by the postfix S in E2* according to rule Rc are considerably far away from it.

It will be noted that we have provided two logically independent sets of principles by which affixed grammars may be constructed from a given CFPSG. The first set is explicitly designed to preserve the property of noncenter-embedding. The second is designed to maximize the use of prefixes on the basis of being able to predict the identity of a constituent by the time its leftmost descendent has been identified. There is no reason to believe a priori that affixed grammars constructed according to the second set of principles should preserve noncenter-embedding, and indeed as we have just seen, they don't. However, we conjecture that natural languages are designed so that representations of the structural descriptions of acceptable expressions of those languages can be assigned to them by finite-state parsers that operate by identifying constituents as quickly as possible. We call this the **Efficient Finite-State Parser Hypothesis.**

The four principles for determining whether to use a prefix or a postfix to mark the presence of a particular constituent apply to grammars that are center embedding as well as to those that are not. Suppose we extend the grammar G2 by replacing rules G2e and f by rules G2e' and f' respectively, and adding rules G2m-s as follows:

(G2) e'. $\bar{N} \longrightarrow N$ (PP1)

f'. $VP \longrightarrow V$ (NP) $(\{PP2, \bar{C}\})$

m. $NP \longrightarrow NP\ PP2$

n. $PP1 \longrightarrow P1\ NP$

o. $PP2 \longrightarrow P2\ NP$

p. $VP \longrightarrow VP\ \{A, PP2\}$

q. $A \longrightarrow$ yesterday

r. $P1 \longrightarrow$ of

s. $P2 \longrightarrow \{$in, on, ...$\}$

Among the expressions generated by the extended grammar G2 are those in E3.

(E3) a. the boss knew that the teacher saw the child yesterday

b. the friend of the teacher's sister

Although each of the expressions in E3 is ambiguous with respect to G2, each has a strongly preferred interpretation. Moreover, under each interpretation, each of these sentences manifests first-degree center embedding. In E3, the included VP saw the child is wholly contained in the including VP knew that the teacher saw the child yesterday; and in E3b, the included NP the teacher is wholly contained in the including NP the friend of the teacher's sister.

Curiously enough, the extension of the affix grammar that our principles derive from the extension of the grammar G2 just given associates only one affixed string with each of the expressions in E3. That grammar is obtained by replacing rules G2*e and F with G2*e' and f' respectively, and adding the rules G2*m-s as follows.

(G2*) e'. $\bar{N}$* ⟶ $\bar{N}$ N* (PP1*)

f'. VP* ⟶ VP V* (NP*) ({PP2*, $\bar{C}$*})

m. NP* ⟶ NP* PP2* NP

n. PP1* ⟶ PP1 P1* NP*

o. PP2* ⟶ PP2 P2* NP*

p. VP* ⟶ VP* {A*, PP2*} VP

q. A* ⟶ A yesterday

r. P1* ⟶ P1 of

s. P2* ⟶ P2 {in, on, ...}

The affix strings that the extended affix grammar G2* associates with the expressions in E3 are given in E3*.

(E3*) a. NP D the $\bar{N}$ N boss VP V knew $\bar{C}$ C that S NP D the $\bar{N}$ N teacher VP V saw NP D the $\bar{N}$ N child A yesterday VP S

b. NP D the $\bar{N}$ N friend PP1 P1 of NP D the $\bar{N}$ N teacher G 's $\bar{G}$ $\bar{N}$ N sister NP

We contend that the fact that the expressions in E3 have a single strongly preferred interpretation results from the fact that those expressions have a single affixed string associated with them. Consider first E3a and its associated affixed string E3*a. According to rule Rc, the affix VP following yesterday is a postfix which covers the affixes VP and A. Now, there is only one occurrence of A in E3*a, namely the one that immediately precedes yesterday; hence that must be the occurrence which is covered by the postfix VP. On the other hand, there are two occurrences of prefix VP in E3*a that can legitimately be covered by the postfix, the one before saw and the one before knew. Suppose in such circumstances, rule Rc picks out the nearer prefix. Then automatically the complex VP, saw the child yesterday, is covered by the subordinate S prefix, in accordance with the natural interpretation of the expression as a whole.

Next, consider E3b and its associated affixed string E3*b. According to rule Rc, the $\bar{G}$ is a postfix that covers the affixes NP and G. Two occurrences of the prefix NP are available to be covered; again, we may suppose that rule Rc picks out the nearer one. If so, then automatically the complex NP, the teacher's sister, is covered by PP1, again in accordance with the natural interpretation of the expression as a whole.

This completes our demonstration of the ability of affixed strings to represent the structural descriptions of the acceptable sentences of a natural language in a manner which enables them to be parsed by a finite-state device, and which also predicts the way in which (at least) certain expressions with center embedding are actually interpreted. Much more could be said about the system of representation we propose, but time and space limitations preclude further discussion here. We leave as exercises to the reader the demonstration that the expression E4a has a single affixed string associated with it by G2*, and that the left-branching (stacked) interpretation of E4b is predicted to be preferred over the right-branching interpretation.

(E4) a. the student saw the teacher in the house

b. the house in the woods near the stream

ACKNOWLEDGMENT

We thank Maria Edelstein for her invaluable help in developing the work presented here.

REFERENCES

Forster, Kenneth I. (1976) Accessing the mental lexicon. In R.J. Wales and E.T. Walker, eds., New Approaches to Language Mechanisms. Amsterdam: North-Holland.

Jackendoff, Ray S. (1977) X-Bar Syntax. Cambridge, Mass.: MIT Press.

Langendoen, D. Terence (1975) Finite-state parsing of phrase-structure languages and the status of readjustment rules in grammar. Linguistic Inquiry 6.533-54.