The Elimination of Grammatical Restrictions in a String Grammar of English *

M. Salkoff and N. Sager

Institute for Computer Research in the Humanities

New York University, New York

RESUME

In linguistic string analysis[1], the major syntactic structures of English are given by about 200 elementary strings (sequences of word categories) together with a rule of combination on strings to form sentences. The grammatical dependencies are expressed by restrictions on the strings as to the word subcategories which can occur together in a string or in related strings. One may ask whether it is possible to write a string grammar in which all restrictions are eliminated, i.e., all grammatical dependencies are between elements of the same elementary string. We have done this by rewriting in a restrictionless form an existing fairly detailed computer string grammar of English[2]. There results an increase of an order of magnitude in the size of the grammar.

1. Harris, Z. S., String Analysis of Sentence Structure, Papers on Formal Linguistics, No. 1, Mouton and Co., The Hague, 1962.

2. Sager, N., Salkoff, M., Morris, J., Raze, C., Report on the String Analysis Programs, Department of Linguistics, University of Pennsylvania, Philadelphia, Pennsylvania, March 1966.

* Presented at 2ème Conférence Internationale sur le Traitement Automatique des Langues, Grenoble, August 1967.

The Elimination of Grammatical Restrictions in a String Grammar of English

M. Salkoff and N. Sager

Institute for Computer Research in the Humanities

New York University, New York

1. Summary of String Theory

In writing a grammar of a natural language, one is faced with the problem of expressing grammatical dependencies. For example, in the sentence form N V N (N, noun: V, verb), the subject N and the verb V must agree in number: The boy eats the meat; ∄ The boys eats the meat. Or, in the sequence Q N_1 P N_2 (Q a number; P, preposition), e.g., five feet in length, N_1 and N_2 are of particular subclasses: ∄ five feet in beauty. One of the theories of linguistic structure which is particularly relevant to this problem is linguistic string analysis [1]. In this theory, the major syntactic structures of English are stated as a set of elementary strings (a string is a sequence of word categories, e.g., N V N, N V P N, etc.). Each sentence of the language consists of one elementary sentence (its center string) plus zero or more elementary adjunct strings which are adjoined either to the right or left or in place of particular elements of other elementary strings in the sentence.

The elementary strings can be grouped into classes according to how and where they can be inserted into other strings. If $Y = X_1 X_2 \ldots X_n$ is an elementary string, X ranging over the category symbols, the following classes of strings are defined:

ℓ_X left adjuncts of X: adjoined to a string Y to the left of X in Y, or to the left of an ℓ_X adjoined to Y in this manner.

r_X right adjuncts of X: adjoined to a string Y to the right of X in Y, or to the right of an r_X adjoined to Y in this manner.

n_X replacement strings of X: adjoined to a string Y, replacing X in Y.

s_Y sentences adjuncts of the string Y, adjoined to the left of X_1 or after X_i in Y ($1 \leq i \leq n$), or to the right of an s_Y adjoined to Y in this manner.

$c_{Y,i}$ conjunctional strings of Y, conjoined after X_i in Y ($1 \leq i \leq n$), or to the right of a $c_{Y,i}$ adjoined to Y in this manner.

z center strings, not adjoined to any string.

These string-class definitions, with various restrictions on the repetition and order of members of the classes, constitute rules of combination on the elementary strings to form sentences.

Roughly speaking, a center string is the skeleton of a sentence and the adjuncts are modifiers. An example of a left adjunct of N is the adjective green in the green blackboard. A right adjunct of N is the clause whom we met in the man whom we met. A replacement formula of N is, for example, what he said in the sentence What he said was interesting. The same sentence with a noun instead of a noun replacement string might be The lecture was interesting. Examples of sentence adjuncts are in general, at this time, since he left. The c strings have coordinating conjunctions at their head. An example is but left in He was here but left. Examples of center strings are He understood and also We wondered whether he understood.

The grammatical dependencies are expressed by restrictions on the strings as to the word subcategories which can occur together in a string or in strings related by the rules of combination. Thus, in the center string N_1 V N_2, the

grammatical dependency mentioned above is formulated by the restriction: if N_1 is plural, then V does not carry the singular morpheme -s. The string grammar with restrictions gives a compact representation of the linguistic data of a language, and provides a framework within which it is relatively simple to incorporate more linguistic refinement, i.e., more detailed restrictions.

One may ask whether it is possible to write such a string grammar without any restrictions at all, i.e., to express the grammatical dependencies (restrictions) in the syntactic structures themselves. In the resulting restrictionless grammar, any elements which are related by a grammatical dependency will be elements of the same elementary string. No grammatical relations, other than those given by the simple rule of string combination, obtain between two strings of a sentence. The result of this paper is to demonstrate that such a restrictionless grammar can be written [4].

In order to obtain a restrictionless form of a string grammar of English, we take as a point of departure the grammar used by the computer program for string decomposition of sentences, developed at the University of Pennsylvania [2,3]. This grammar is somewhat more detailed than the sketch of an English string grammar in [1]. A summary of the form of the computer grammar is presented below in section 2. In section 3 we show how the restrictions can be eliminated from the grammar.

An example of a typical output obtained for a short sentence from a text of a medical abstract is shown in Figs. 1 and 2. The decomposition of the sentence into a sequence of nested strings is indicated in the output by the numbering of the strings. As indicated in line 1., the sentence consists of the two assertion centers in lines 2. and 4., conjoined by and. The line 3.

contains a sentence adjunct (thus) on the assertion center as a whole. The assertion center 2. is of the form N V A : Spikes would be effective. The noun spikes has a left adjunct (such enhanced) in line 5., as indicated by the appearance of 5. to the left of spikes. The object effective has a left adjunct (more) in line 6. and a right adjunct in line 7. In the same way, each of the elements of the adjunct strings may have its own left and right adjuncts. Line 10. contains an assertion center in which the subject and the modal verb (would) have been zeroed. This zeroing is indicated in the output by printing the zeroed element in parentheses.

The difference between the two analyses in Figs. 1 and 2 lies in the decomposition of the sequence in initiating synaptic action. In the first analysis (Fig. 1), this sequence is taken as a P N right adjunct on effective, where initiating synaptic is a left adjunct (on action) of the form of a repeated adjective (parallel to escaping toxic in the sequence in escaping toxic gases). In the second analysis (Fig. 2), this same sequence is taken as a P Ving right adjunct of effective, where initiating is the Ving, and synaptic action is the object of initiating.

2. The Computer String Grammar.

In representing the string grammar in the computer, a generalized grammar string is used[5] which is defined as

(1) $Y = Y_1 / Y_2 / \ldots / Y_n$

where

(2) $Y_i = Y_{i1} Y_{i2} \ldots Y_{im}$

and

(3) $Y_{ij} = Y'$ where Y' is a grammar string like Y.

This system of nested grammar strings terminates when one of the grammar strings is equal to an atomic string (one of the word-category symbols). The Y_i are called the options of Y, and each option Y_i consists of the elements Y_{ij}.

Not every option of a grammar string Y will be well-formed each time the sentence analysis program finds an instance of Y in the sentence being analyzed. Associated with each option Y_i is a series of zero or more tests, called restrictions. If R_i is the set of tests associated with Y_i then the grammar string Y can be written:

(4) $Y = R_1Y_1 / R_2Y_2 / \ldots / R_nY_n$

A restriction is a test (which will be described below) so written that if it does not give a positive result its attached option may not be chosen.

All of the restrictions in the grammar fall into two types:

Type A: The restrictions of type A enable one to avoid defining many similar related sets of grammar strings. The options of the grammar string Y have been chosen so that Y represents a group of strings which have related

linguistic properties. This allows the grammar to be written very compactly, and each grammar string can be formulated as best suits the linguistic data. However, when a grammar string Y appears as a Y'_{ij} of some other string Y', some of the options of Y may lead to non-wellformed sequences. In order to retain the group of options of Y and yet not allow non-wellformed sequences wherever options of Y which would have that effect are used, we attach a restriction of type A to those options of Y.

For example, let Y be

(5) $Y = Y_1 \ / \ R^a \ Y_2 \ / \ . \ . \ .$

where

(6) Y_1 = which Σ V (e.g., which he chose)

and

Y_2 = what Σ V (e.g., what he chose)

Then Y can appear in the subject Σ of the linguistic center string C1:

(7) C1 = Σ V Ω

This yields Which he chose was important; What he chose was important.

As it is defined here, Y can also be used to represent the wh-clauses in the right adjuncts of the noun:

(8) $Y' = r_N = \ . \ . \ . \ / \ Y \ / \ . \ .$

but in r_N only the which option of Y gives wellformed sequences:

∃ the book which he chose

∄ the book what he chose

Hence a restriction R^a is attached to the what option of Y (eq. 5) whose effect is to prevent that option from being used in r_N.

Type B: With some given set of rather broadly defined major categories (noun, verb, adjective, etc.) it is always possible to express more detailed linguistic relations by defining sub-categories of the major categories. These relations then appear as constraints on how the sub-categories may appear together in the grammar strings Y.

If some element Y_{ij} of Y_i is an atomic string (hence a word-category symbol) representing some major category, say C, then R^b may exclude the sub-category C_j as value of Y_{ij} if some other element Y_{ik} of Y_i has the value C_k. Y_{ik} may also be a grammar string, in which case R^b may exclude a particular option of Y_{ik} when Y_{ij} has value C_j.

The restrictions R^b may be classified into three kinds:

(a) Between elements of some string Y_i where the Y_{ij} correspond to elements of a linguistic string.

For example,

A noun in the sub-category singular cannot appear with a verb in the sub-category plural. $\not\beta$ The man agree.

Only a certain sub-category of adjective can appear in the sentence adjunct P A : in general, in particular, $\not\beta$ in happy.

(b) Between a Y_{ij} and a Y_{ik} where Y_{ij} corresponds to an element of a linguistic string and Y_{ik} corresponds to a set of adjuncts of that element. For example,

In r_N, the string to V Ω cannot adjoin a noun of sub-category N_2 (proper names): the man to do the job $\not\beta$ John to do the job.

Only a certain adjective sub-category (e.g., present, available) can appear in r_N without any left or right adjunct of its own: the people present ; ∄ the people happy.

(c) Between Y_{ij} and Y_{ik} , where one corresponds to an element of a linguistic string and the other corresponds to an adjunct set which can repeat itself; i.e., which allows 2 or more adjuncts on the same linguistic element. These restrictions enable one to express the ordering among adjuncts in some adjunct sets. For example,

Q (quantifier) and A (adjective) are both in the set ℓ_N , the left adjuncts of the noun. However, Q can precede A but A cannot precede Q when both are adjuncts of the same N in a sentence: ∃ Q A N e.g., five green books , but ∄ A Q N e.g., green five books.

The string grammar defined by eqs. 1-3, together with the atomic strings (word-category symbols) have the form of a BNF definition. The system with eq. 4, however, departs from a BNF definition in two important respects:

(a) it contains restrictions (tests) on the options of a definition;

(b) the atomic strings (word-categories) of the grammar have sub-classifications.

With the elimination of the restrictions, the computer grammar will again have the form of a BNF definition.

3. Elimination of the Restrictions

The restrictionless string grammar is obtained from the grammar described above by the methods of (A) and (B) below. Initially (in this paper), conjunctional strings have not been included in the restrictionless grammar. We estimate that the addition of conjunctional strings will increase the size of the restrictionless grammar by a factor of about 5.

(A) The linguistic strings represented in the computer grammar are reformulated in accordance with the following requirement. Given any utterance of a language containing A . . . B . . . , where a grammatical dependency obtains between A and B , the elementary strings of a restrictionless string grammar are defined so that A and B appear together in the same linguistic string, and any iterable sequence between A and B is an adjunct of that string. Iterable sequences of the type seemed to begin to in It seemed to begin to surprise him that we worked seriously , or is said to be known to in It is said to be known to surprise him that we worked seriously are analyzed as adjuncts. If we place such sequences among the left adjuncts of the verb, ℓ_V , then the sentences above can be put in the form

(9) It ℓ_V surprise him that we worked seriously

ℓ_V = seemed to begin to ; is said to be known to ; etc.

However, when the adjunct ℓ_V takes on the value zero (as can all adjuncts, by definition), then (9) above becomes the non-grammatical sequence It surprise him that we worked seriously. This happens because the first verb of ℓ_V (seemed or is) carries the tense morpheme, and the latter disappears when $\ell_V = 0$. We separate the tense morpheme from the verb, and place it in the center string as one of the required elements.

(10) $C1 = \Sigma\ t\ \ell_V\ V\ \Omega$; $\quad t = 0\ /\ -s\ /\ -ed\ /$ will, can, ...

This formulation of the assertion center string C1 (10), in which the tense morpheme is an independent element and iterable sequences are taken as adjuncts, is necessary in order to preserve, for example, the dependence between the particle it and the succeeding sequence surprises him that we worked seriously: $\not$ The book surprises him that we worked seriously. In the grammar which includes restrictions, this formulation is not necessary because this dependence can be checked by a restriction.

(B) Turning to the computer form of the grammar, all the restrictions of the grammar are eliminated either by defining new grammar strings (for the elimination of the restrictions R^a) or by replacing the general word-categories by the particular subclasses of those categories which are required by the restriction (to eliminate R^b). The application of this procedure increases the number of strings in the grammar, of course.

The restrictions R^a can be eliminated in the following manner. Suppose the option Y_i of Y has a restriction R^a on it which prevents it from being chosen in Y' (Y is a Y'_{ij} of Y'). Then define a new grammar string Y* which

contains all the options of Y but Y_i :

(15) $Y^* = Y_1 / Y_2 / \ldots / Y_{i-1} / Y_{i+1} / \ldots / Y_n$

Then the new grammar string Y^* replaces Y in Y'. Thus, in the example of R^a on p. 5, the string Y^* = which Σ t f_V V / ... (in the modified treatment of tense and iterable sequences) would replace Y in r_N.

The restrictions R^b are eliminated in a different way, according to the types described on p. 6.

(a) New strings must be written in which only the wellformed sequences of subcategories appear. In the example of subject-verb agreement, the original Y_i (Y_i = Cl) must be replaced by two options:

$$Cl = N\ t\ V\ \Omega \rightarrow N_s\ t\ V_s\ \Omega\ /\ N_p\ t\ V_p\ \Omega$$

where N_s and N_p are singular and plural nouns, V_s and V_p singular and plural verbs.

(b) If an element of a particular subcategory, say A_i, can take only a subset of the adjuncts r_A, then a new adjunct string r_{Ai} is defined. It contains those options of r_A which can appear only with A_i plus all the options of r_A which are common to all the sub-categories of A. When this has been done for all A_i having some particular behavior with respect to r_A, all the remaining sub-categories of A will have a common adjunct string r_a :

$$A\ r_A \rightarrow A_1 r_{A1} / A_3 r_{A3} / \ldots / A_2 r_a / A_4 r_a / \ldots$$

As many new sets r_{Ai} must be defined as there were special sub-categories of A. A similar argument holds for f_A and other adjunct sets which depend on A.

(c) A new element corresponding to the adjunct set must be defined in which the adjuncts appear correctly ordered with respect to each other, and each one must be able to take on the value zero.

This procedure for eliminating restrictions is also the algorithm for introducing further grammatical refinements into the restrictionless grammar. Such a general procedure can be formulated because of an essential property of a string grammar: In terms of linguistic (elementary) strings, all restrictions are either a) between elements of a string, or b) between an element of the string and its adjunct, or c) between related adjuncts of the same string. Further, there is no problem with discontinuous elements in a string grammar: all elements which depend in some way on each other grammatically appear in the same string or in strings which are contiguous by adjunction.

The cost of the elimination of all restrictions in this way is about an order of magnitude increase in the number of strings of the grammar. Instead of about 200 strings of the computer grammar, the grammar presented here has about 2000 strings. It is interesting that the increase in the size of the grammar is not greater than roughly one order of magnitude. This suggests that there may be practical applications for such a grammar, e.g. in a program designed to carry out all analyses of a sentence in real time. Also, since the restrictionless grammar is equivalent to a B.N.F. grammar of English, it may prove useful in adding English-language features to programming languages which are written in B.N.F.

Figure 1

SENTENCE NEUR-18 . SUCH ENHANCED SPIKES WOULD BE MORE EFFECTIVE IN INITIATING SYNAPTIC ACTION AND THUS BE RESPONSIBLE FOR THE OBSERVED POST-TETANIC POTENTIATION .

PARSE 01

1. SENTENCE = INTRODUCER CENTER AND END MARK
2. AND 3. 4. .

2. C1 ASSERTION = * SUBJECT * VERB * OBJECT RV *
5. SPIKES WOULD BE 6. EFFECTIVE 7.

3. ADVERB = ADVERB
THUS

4. CONJUNCTION = CENTER
10.

5. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
SUCH ENHANCED

6. ADVERB = ADVERB
MORE

7. P N = LP PREPOSITION N
IN 11. ACTION

10. C1 ASSERTION = * SUBJECT * VERB * OBJECT RV *
(5. SPIKES) (WOULD) BE RESPONSIBLE 12.

11. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
INITIATING SYNAPTIC

12. P N = LP PREPOSITION N
FOR 13. POTENTIATION

13. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
THE OBSERVED POST-TETANIC

Figure 2

SENTENCE NEUR-18 . SUCH ENHANCED SPIKES WOULD BE MORE EFFECTIVE IN INITIATING SYNAPTIC ACTION AND THUS BE RESPONSIBLE FOR THE OBSERVED POST-TETANIC POTENTIATION .

PARSE 02

1. SENTENCE = INTRODUCER CENTER AND END MARK
2. AND 3. 4. .

2. CI ASSERTION = * SUBJECT * VERB * OBJECT RV *
5. SPIKES WOULD BE 6. EFFECTIVE 7.

3. ADVERB = ADVERB
THUS

4. CONJUNCTION = CENTER
10.

5. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
SUCH ENHANCED

6. ADVERB = ADVERB
MORE

7. P NS VING(OF) O = PREPOSITION SN
IN INITIATING 11. ACTION

10. CI ASSERTION = * SUBJECT * VERB * OBJECT RV *
(5. SPIKES) (WOULD) BE RESPONSIBLE 12.

11. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
SYNAPTIC

12. P N = LP PREPOSITION N
FOR 13. POTENTIATION

13. LN = ARTICLE QUANTIFIER ADJECTIVE TYPE-NS NOUN
THE OBSERVED POST-TETANIC

NO MORE PARSES

REFERENCES

1. Harris, Z. S., String Analysis of Sentence Structure, Papers on Formal Linguistics, No. 1, Mouton and Co., The Hague, 1962.

2. Sager, N., Salkoff, M., Morris, J., and Raze, C., Report on the String Analysis Programs, Department of Linguistics, University of Pennsylvania, March 1966.

3. Sager, N., "Syntactic Analysis of Natural Language", Advances in Computers (Alt, F. and Rubinoff, M., eds.), vol. 8, pp. 153-188. Academic Press, New York, 1967.

4. This problem was suggested by Professor J. Schwartz of the Courant Institute of Mathematical Sciences, New York University.

5. The option Y_j here corresponds to the linguistic string Y of the previous section. The symbol / separates the options of a string definition.