A MODEL AND AN HYPOTHESIS FOR LANGUAGE STRUCTURE*

VICTOR H. YNGVE

Department of Modern Languages and Research Laboratory of Electronics,
Massachusetts Institute of Technology

(*Read November 13, 1959*)

One of the outstanding characteristics of language is its wealth of complexity, particularly on the level of sentence structure. The complexity seems to be divided into two parts—a part with an obvious communicative or signaling function, and a much larger part with little or no apparent function.

As an example of obviously functional features, we can point to systems for subordination and role marking, like case endings, prepositions, subordinating conjunctions, or word order as in *John saw Mary* vs. *Mary saw John.* But even here, although the communicative function of such devices is clear, it is not obvious why a given language should use so many different types, particularly in view of the fact that it has proved to be possible to devise languages for mathematics or symbolic logic which use only one uniform method of subordination and role marking in much the same way as English uses *after* to subordinate either a phrase or a clause.

As an example of the apparently nonfunctional syntactic complications, there is the prevalence of discontinuous constituents as typified by the expressions *a more priceless (possession) than jewels* or *too heavy (a box) to lift.* Not only do these complications seem to be nonfunctional, but they seem so anomalous that they tax our descriptive methods. In linguistic analysis we usually try to go as far as possible on the basis of continuous constituents and only admit the more complex kind of analysis when it is forced upon us.

In this paper, a simple and easily mechanized model for sentence production is set up. On the basis of the behavior of the model, and the assumptions underlying it, an hypothesis is advanced which leads to a number of specific predictions concerning the types of syntactic structures to be expected in language. The structure of English is then examined in the light of these predictions, and it is shown that the predicted structures account for much of the apparently nonfunctional complexity.

The model arose out of research directed toward the mechanical translation of languages. An adequate translating machine must have at least three parts: a part that receives the incoming text and analyzes it, providing an explicit representation of its structure; a part that selects an appropriate structure in the other language; and a third part that actually produces the output text.[1] To accomplish this last task, we need a device that can produce grammatical sentences in English, assuming that there is available as input an indication of just which sentences to produce. Since such a device must work according to the grammatical rules of English, we immediately face the problem of coping adequately with the wealth of its syntactic complexity.

A conception of language structure involving, in some form or other, a phrase-structure hierarchy, or immediate constituent organization, has been used extensively[2] in spite of its shortcomings. It has the advantage of a certain simplicity and elegance, and it provides a framework for the description of many of the significant features of language structure. But inasmuch as

* This work was supported in part by the National Science Foundation, and in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U. S. Navy (Office of Naval Research).

The author is indebted to a number of his friends, who read earlier versions of this paper and with whom he had the opportunity of discussing these matters. For constructive criticisms, comments, and suggestions, he would particularly like to thank J. R. Applegate, A. N. Chomsky, D. A. Dinneen, M. Halle, E. S. Klima, W. S. McCulloch, G. H. Matthews, and J. Viertel.

[1] Yngve, V. H., A framework for mechanical translation, *Mechanical Translation* **4**: 59–65, 1957.

[2] See, for example: Bloomfield, Leonard, *Language,* 184 ff., New York, Henry Holt and Company, 1954. Wells, Rulon S., Immediate constituents, *Language* **23**: 81–117, 1947. Chomsky, Noam, *Syntactic structures,* The Hague, Mouton and Co., 1957. Nida, Eugene A., *A synopsis of English syntax,* Norman, Oklahoma, Summer Institute of Linguistics of the University of Oklahoma, 1960.

it is a descriptive framework, it is static. As it is usually conceived, it is not a mechanism or model of sentence production. We can, however, adopt it as the conceptual basis for a model and try to deal with the shortcomings when they arise.

Since the model will be used to make predictions about language structure, which will then be compared with observation, it may be possible later to make inferences about the validity of the assumptions underlying the model. For this reason an attempt will be made to identify and state explicitly the essential assumptions.

(1) The first assumption is that a phrase-structure or immediate constituent framework can be used as the basis for a model of sentence production and that any shortcomings can be overcome.

In contrast to a descriptive framework, a model for sentence production is involved in an essential way with the element of time. The sentences of a language are uttered in a time sequence one at a time, as are the words of each sentence. But there is no natural or grammatical limit to the length of a sentence. No matter how long a sentence is given, it is possible to construct a longer one by adding, for example, a dependent clause somewhere. If there is no natural limit to the length of a sentence, it is unreasonable to assume that sentences are formed in the mind of the speaker in their full detail before he starts to utter them. In fact, there is evidence to the contrary. There are many examples in which a person starts a sentence and has to stop before he has finished and start again because he has not completely thought out the whole sentence.

(2) The second assumption is that the model should share with the human speaker of the language the property that words are produced one at a time in the proper time sequence; that is, in left-to-right order according to conventional English orthography.

THE MODEL

The model consists of a grammar and a mechanism. The grammar contains the rules of the particular language that is being produced. The mechanism, on the other hand, is quite general and will work with the grammar of any language. It is merely a device that applies rules and thus produces sentences.

The grammar consists of a finite, unordered set of constituent structure rules that may be of one or more of the following types:

$$\begin{aligned} A &= B + C \\ A &= B + C + D \\ A &= B + C + D + E \\ &\ \ \cdot \\ &\ \ \cdot \\ &\ \ \cdot \end{aligned}$$

There may also be rules of the type

$$A = B$$

The precise significance of the rules will become clear with the description of the mechanism that applies them. It can be stated, however, that the rules should determine the immediate constituents and their order for every construction in the language. The symbol on the left of the "equation" designates the construction, and the symbols on the right, separated by plus signs, designate the constituents of the construction. These constituents may either be words, or they may be symbols for other constructions.

It is now possible to amplify the first assumption—that an immediate constituent framework is adequate:

(1*a*) We assume that a grammar consists of constituent-structure rules of the type given above.

(1*b*) We assume that an adequate set of constituent-structure rules for a language is finite. We can thus store the grammar in a device with a finite memory.

(1*c*) We assume that the set of constituent-structure rules is unordered. This means that no grammatical significance is attached to the order in which the rules are listed in the memory. Any order will do; an alphabetical order of listing may be convenient.

The mechanism gives precise meaning to the set of rules by providing explictly the conventions for their application. The mechanism is illustrated schematically in figure 1. It is an idealized computer and is physically realizable. It consists of four cooperating parts. There is an output

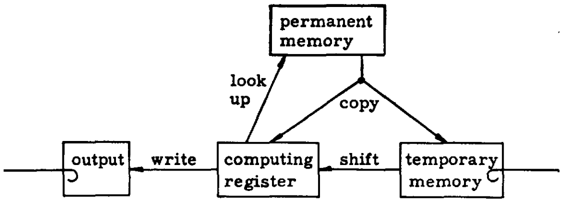


FIG. 1. How the mechanism is organized.

device that prints the output symbols one at a time in left-to-right fashion on an output tape. There is a computing register capable of holding one symbol at a time. There is a permanent memory in which the grammar rules are stored, and there is a temporary memory, in the form of a tape, on which intermediate results are stored.

The mechanism operates in such a manner that assumption 2, the left-to-right order of producing words, is satisfied. The details of this assumption can be amplified as follows:

(*2a*) The rules are applied by expansion. This means that in the mechanism, the symbol for a construction is expanded, i.e., replaced by the symbols for the constituents of that construction, as indicated in the appropriate grammar rule. The way in which this is accomplished will be given in the program below.

(*2b*) The left-most constituent of a construction is always expanded before the other constituents of that construction.

(*2c*) Constituents awaiting their turn to be expanded are stored in a temporary memory equivalent to a tape that can be rolled in and out. Only the portion closest to the roll is available to be written on or to be read from and erased.

The program that the mechanism uses is given below. It is cyclic. The steps are carried out in the order indicated. Each time a complete cycle is executed, one word or constituent symbol is written out on the output tape. A construction may or may not be expanded during each cycle.

```
I.   START. Insert $S$ (for Sentence) into the computing register.
II.  WRITE. Produce an output symbol as follows:
     A. Unroll enough output tape for one symbol, and
     B. Copy (duplicate) the symbol that is in the computing register so that it also appears on the output tape.
III. CONDITIONAL SHIFT OR STOP.
     A. If the symbol in the computing register is a word, delete it, then
        1. If there is at least one symbol in the temporary memory,
           a. copy the left-most one into the computing register and
           b. delete it from the temporary memory, and then
           c. roll in the blank tape so produced, then
           d. go on to step II.
        2. But if there are no symbols in the temporary memory, STOP (end of sentence).
     B. But if the symbol in the computing register is not a word, go on to step IV.
IV.  LOOK UP. Look up and select a grammar rule as follows:
     A. Compare the symbol in the computing register with the left sides of the grammar rules stored in the permanent memory, and note the rules where there is a match, then,
     B. Select one of the matching grammar rules. Different choices will result in different sentences.
V.   COPY. Copy the right-hand side of the selected grammar rule as follows:
     A. Delete the contents of the computing register.
     B. If the right-hand side of the selected grammar rule has more than one symbol, unroll enough temporary memory tape to accommodate the symbols in excess of one.
     C. Copy (duplicate) the symbols in the right-hand side of the selected grammar rule into the available space in the computing register and temporary tape. The first symbol goes into the computing register; the others go, in order, onto the temporary tape.
     D. Now execute step II.
```

For illustration of the various steps in constructing a sentence, we can use any convenient grammar that is in the proper form. Let us take the following simple grammar:

$$\begin{aligned} N &= \text{man}\\ N &= \text{boy}\\ NP &= T + N\\ S &= NP + VP\\ T &= \text{the}\\ T &= \text{a}\\ V &= \text{saw}\\ V &= \text{heard}\\ VP &= V + NP \end{aligned}$$

in which S stands for sentence, N stands for noun, V stands for verb, P stands for phrase, T stands for article.

In figure 2 the first line shows the result of step I: the symbol S has been inserted in the computing register. The next line shows the result after steps II, IV, and V have been taken. (Step III does not apply at this point.) The following lines show the result after step V in each succeeding cycle, and the last line shows the result after the mechanism has stopped in step III. The final output sequence of symbols appears on the output tape.

An interesting property of the output sequence of symbols is that it represents not only the terminal string (the words) but also all of the nodes of the tree of derivation and their interconnections. In other words, the output sequence represents not only a sentence, but also its constituent structure in an explicit form. This explicit form is the familiar Polish[3, 4] parenthesis-free notation in which every node is written followed immediately by the subtrees under it in their natural order. Thus the output sequence represents a constituent structure diagram or tree, as in figure 3. It is of course important to ensure, in the parenthesis-free notation, that the number of branches from a node is unambiguous.

(output tape)	(register)	(temporary)	
	S		
S	NP	VP	
S NP	T	N	VP
S NP T	the	N	VP
S NP T the	N	VP	
S NP T the N	man	VP	
S NP T the N man	VP		
S NP T the N man VP	V	NP	
S NP T the N man VP V	saw	NP	
S NP T the N man VP V saw	NP		
S NP T the N man VP V saw NP	T	N	
S NP T the N man VP V saw NP T	the	N	
S NP T the N man VP V saw NP T the	N		
S NP T the N man VP V saw NP T the N	boy		
S NP T the N man VP V saw NP T the N boy			

FIG. 2. Steps in producing a sentence.

The order of the symbols in the output sequence is the order in which the nodes of the tree enter the computing register. It can be seen from figure 3 that this is equivalent to expanding the left-hand member of every construction first, and when reaching the end of a branch, retracing to the next higher unexpanded right-hand member.

Although the number of rules in the grammar is finite, this device has the desired property that it can produce any sentence of an infinite set of sentences. This comes about because certain rules can be reapplied along the same branch during the production of a sentence. Indefinitely long sentences can result, since there is the possibility of "infinite loops" that involve continual reapplication of rules and provide sentences having clauses within clauses. Figure 4 shows the beginning of the production of such a sentence. The rule $S = NP + VP$ was applied at node 1, and it has been applied for the second time at node 13.

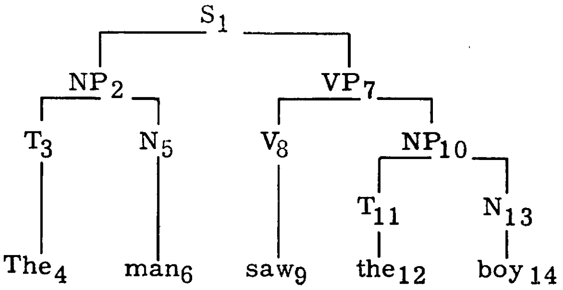


FIG. 3. The constituent-structure tree represented by the last line in figure 2. The subscripts indicate the order in which the symbols enter the computing register.

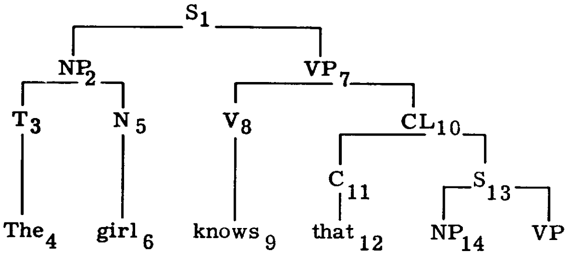


FIG. 4. Indefinitely long sentences can be produced when the grammar permits rules to be reapplied. Here the rule $S = NP + VP$ has been reapplied.

ADEQUACY OF THE MODEL

If we look upon our device as a component in a translating machine, we are interested in asking whether or not the device will be adequate; that is, whether it is capable of handling any output sentence it may be called upon to produce. We are also interested in certain practical questions as to its efficiency. But if we look upon our device as a model of language production, we are interested in the extent to which it conforms to human behavior, and whether there are any predictions from the model that can be checked against observations of language.

We can first try our device on a very simple language, the notation of algebra. A constituent

[3] Łukasiewicz, Jan, *Aristotle's syllogistic,* 77 ff., Oxford, 1957.

[4] Łukasiewicz and Tarski, Untersuchungen über den Aussagenkalkül, *Comptes Rendus des Séances de la Société des Sciences et des Lettres de Varsovie* 23, CI III: 31-32, 1930.

EQN	=	EXPR + "=" + EXPR	EXPR	=	SUM	LEFT	=	SUM'
SUM	=	EXPR + "+" + EXPR	EXPR	=	QUOT	LEFT	=	QUOT'
SUM	=	EXPR + "-" + RITE	EXPR	=	PROD	LEFT	=	PROD
PROD	=	LEFT + RITE	EXPR	=	LTRL	LEFT	=	LTRL
QUOT	=	LEFT + "/" + DNOM	RITE	=	SUM'	DNOM	=	SUM'
SUM'	=	"(" + SUM + ")"	RITE	=	QUOT	DNOM	=	QUOT'
PROD'	=	"(" + PROD + ")"	RITE	=	PROD	DNOM	=	PROD'
QUOT'	=	"(" + QUOT + ")"	RITE	=	LTRL	DNOM	=	LTRL

FIG. 5. A grammar for the notation of algebra. The symbols are interpreted as follows. *EQN:* equation; *SUM:* sum or difference; *PROD:* product; *QUOT:* quotient; *EXPR:* expression, term, or minuend; *LEFT:* left factor or numerator; *RITE:* right factor or subtrahend; *DNOM:* denominator; *LTRL:* literal number. There would be further vocabulary rules of the type $LTRL = X$, $LTRL = Y$, etc.

structure grammar is well adapted to mathematical notation. As an example, we give in figure 5 a grammar for the infinite set of well-formed algebraic equations and expressions involving addition, subtraction, multiplication, and division indicated by /. Parentheses will be inserted correctly according to the usual custom of inserting them only where they are absolutely necessary to avoid ambiguity of grouping.

It should be noted that this grammar of algebra is a linguistic description, not a mathematical one. It gives rules for producing well-formed strings of symbols but says nothing about their mathematical meaning. The rules for parentheses, instead of being stated in terms of resolution of ambiguity, are stated in terms of the kinds of nodes in the tree that are being expanded. It can be seen that a product is enclosed in parentheses if it is a denominator; a quotient is enclosed in parentheses if it is a left factor, a numerator, or a denominator; and a sum is enclosed if it is a factor, a subtrahend, a numerator, or a denominator.

If we now try to write a constituent structure grammar for English, we run into complications not found in mathematical notations. The most annoying of these complications is the prevalence of discontinuous constituents, as in the following two examples. The first is the sentence *It is true that he went.* The second is the phrase *as far as the corner.*

In the sentence *That he went is true,* we have a subject clause *that he went* and a predicate *is true.* But in the sentence *It is true that he went,* the subject is discontinuous, consisting of *It . . . that he went.* So far our device is not able to handle the added complication of discontinuous constituents. One alternative is to try to force our treatment of this case into a model with continuous constituents. We could say that the sentence *It is true that he went* consists of two clauses, an introductory predicate clause *It is true,* and a subject clause *that he went.* Our grammar will now produce the desired sentence type.

In the phrase *as far as the corner,* we have an adverb *far* and an adverbial expression of degree *as . . . as the corner,* and this adverbial expression consists of a noun phrase *the corner,* and *as . . . as.* It is possible to represent this construction, too, in terms of continuous constituents, but at the price of doing some violence to our intuitive concept of the structure. We could represent *as far as the corner* as three continuous constituents, an adverbial introducer *as,* an adverb *far,* and an adverbial expression *as the corner.*

A possible alternative to forcing a construction with discontinuous constituents into the mold of continuous constituents is to allow rules of the form $A = B + \ . \ . \ . + C$ to appear in the grammar, and to alter the mechanism and its program in such a way that whenever the computing register and temporary storage contain, for example, $|\,A\,|\ Q\ R\ S\ T,$ and the rule $A = B + \ . \ . \ . + C$ is applied, the result is $|\,B\,|\ Q\ C\ R\ S\ T.$ In order to do this, we replace step V in the program for the mechanism by the following:

V. COPY. Copy the right-hand side of the selected grammar rule as follows:
   A. Delete the contents of the computing register.
   B. If the right-hand side of the selected grammar rule contains the symbol ". . .", then
      1. copy the left-most symbol from the temporary memory into the computing register, and
      2. delete it from the temporary memory.
      3. Unroll the temporary memory tape one space and

```
        4. copy the symbol in the computing register onto this new unrolled section of tape, then
        5. delete the symbol from the computing register.
    C. But if the selected grammar rule does not contain ". . .", then
        1. If the right-hand side of the selected grammar rule has more than one symbol, unroll enough temporary memory tape to accommodate the symbols in excess of one.
    D. Copy (duplicate) the symbols in the right-hand side of the selected grammar rule (but not the ". . .") into the available space in the computing register and temporary tape.
    E. Now execute step II.
```

In this way we are able to take care of many of the discontinuous constituents in English. For example, the sentence *It is true that he went* can be produced if the following rules are included in the grammar, and we start with S.

$$\begin{aligned} ADJ &= \text{true} & V1 &= \text{is} \\ CL &= \text{that} + S & V2 &= \text{went} \\ NP1 &= \text{it} +. \ . \ .+ CL & VP1 &= V1 + ADJ \\ NP2 &= \text{he} & VP2 &= V2 \\ S &= NP1 + VP1 \\ S &= NP2 + VP2 \end{aligned}$$

Here we note that the rule $NP1 = \text{it} +. \ . \ .+ CL$ will enclose the symbol $VP1$ between the word *it* and the symbol CL. $VP1$ is eventually expanded to *is true* and CL is eventually expanded to *that he went,* or to *that it is true that he went,* and so on.

In order to produce the phrases *very far* and *as far as the corner,* we can use the following rules:

$$\begin{aligned} ADV &= ADVI + ADV2 \\ ADV1 &= \text{very} \\ ADV1 &= COMP +. \ . \ .+ NP \\ ADV2 &= \text{far} \\ COMP &= \text{as} +. \ . \ .+ \text{as} \\ N &= \text{corner} \\ NP &= T + N \\ T &= \text{the} \end{aligned}$$

We have seen that assumption 1 (that an immediate constituent framework can be used) is quite adequate as it stands for the simple notations of mathematics, but that English contains certain complications—discontinuous constituents—that are difficult to handle within our original amplifications of this assumption. We have proposed a method of dealing with these complications which seems to be successful enough to warrant an additional amplification.

(*1d*) We assume that the grammar can also contain rules of the form $B = D +. \ . \ .+ E,$ and that they will be interpreted by the mechanism in such a way that after $B + C$ has resulted from the application of the rule $A = B + C,$ the result of applying $B = D +. \ . \ .+ E$ will be $D + C + E$.

MEMORY AND DEPTH

We shall now examine the adequacy of the model further, and turn our attention to the temporary storage. From this we shall be led directly to our main hypothesis.

If the set of sentences that the grammar generates is infinite, there is the possibility that an infinite amount of temporary storage may be required. But a device with an infinite amount of temporary storage is not physically realizable, and we suspect that the human memory, too, does not have an infinite amount of storage capacity. We are thus led to the final amplification of our assumptions:

(*2d*) We assume that the temporary memory in our mechanism can store only a finite number of symbols.

A device having a mechanism with a finite amount of temporary storage and operating with a grammar consisting of a finite set of rules is a finite-state device or finite automaton. This is clear from the following considerations. We can define the contents of the computing register together with the contents of the temporary storage as the state of the device. Since the amount of temporary storage is finite and since the number of different symbols that can be introduced by a finite number of grammar rules is finite, the number of states is finite. It is also clear that the output symbol and the possible transitions to the next state are both uniquely determined by the current state. Thus our constituent structure sentence generator with a finite temporary storage is equivalent to a finite-state device.

The constituent structure aspect of our device provides a method whereby the state can be factored: Although the state of the device is determined by the full contents of the computing register and the temporary memory, it is broken up or factored into a number of separate symbols, It is only the left-most symbol, the one in the computing register, that affects the choice of transitions to the next state. The transition is in general to a nearby state; that is, all of the symbols except the one in the computing register remain unchanged, and their order relative to one another also remains unchanged. It is thus possible to take advantage of redundancies in the state diagram by representing only once in the grammar several groups or neighborhoods of states, each group having the same internal connectivity. The finite constituent structure model thus achieves an efficient use of the permanent memory.

We have seen that a constituent structure grammar seemed to be adequate for algebra, but that we had to add certain extra facilities to cope with the discontinuous constituents of language. We then assumed that our device will have a finite temporary memory. We must therefore re-examine the question of adequacy. Any restriction on the size of the temporary memory may have a disastrous effect on the proper operation of a constituent structure device. What would happen if the mechanism tried to apply a rule and no more temporary storage was available?

Such a situation would come about, for example, in the following case. Suppose that our mechanism has a finite memory limited to six symbols and that it is trying to produce the "sentence" $(AB + C)D = E$ by applying the rules of our grammar for the notation of algebra. In this case, the mechanism would not be able to apply the last rule and expand the inner product because by this time the temporary memory would already have been filled with symbols. But if we suppose, instead, that our temporary memory is not limited to six symbols, but to some larger number, we shall have the same problem again with a longer equation. It is thus clear that a finite constituent structure device cannot produce the full set of algebraic equations even though it can produce any equation from an infinite set of equations. But how about a language like English?

We should now make a distinction between the sentences that our phrase-structure device with a finite memory can actually *produce*, and the sentences that the grammar will *generate*—that is, the full set of sentences implied in the mathematical sense by the grammar. It is easy to see that in the case of algebra, the full set of sentences *generated* by the grammar could be *produced* by our device only under the simplifying assumption that it have an infinite memory. Under our assumption (2*d*) of a finite temporary memory, the full (infinite) set of sentences *generated* by the grammar contains as a subset the (infinite) set of sentences that the mechanism can actually *produce*, and some additional sentences that the device cannot produce.

We now ask three questions. Under what conditions will a finite constituent-structure device fail to operate properly because it has used up its temporary storage capacity? Is it possible to have a *well-behaved* grammar, that is, a grammar so restricted that all the sentences *generated* can actually be *produced* by a finite constituent structure device with a given temporary memory capacity? And is it possible that the grammar of English, unlike the grammar of algebra, is well behaved? In order to answer these questions we are led to investigate the relation between output sequences and the amount of temporary storage needed to produce them.

An examination of the example given in figure 2 shows that the maximum number of symbols stored in the temporary memory is two (in the third and fourth line). The maximum amount of temporary storage needed for producing a given output sequence can be calculated for any given output tree in the following way: First, number the branches of each node from 0 to n-1, where n is the number of branches from that node. Start numbering from the right as in figure 6. Then, compute the depth d of each terminal node by adding together the numbers written along all branches leading to that terminal node. This is illustrated in figure 7. The amount of temporary storage needed to construct the tree is then $D = d_{\max}$, the largest value that d takes in the tree. We call D the depth of the sentence. It is the amount of temporary memory needed to produce that sentence.

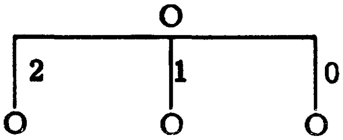


FIG. 6. Numbering of the branches from a node.

Let us now investigate what tree structures can be produced by a finite constituent-structure device. If the memory is small, say with a capacity for only three symbols, only sentences with a depth no larger than three can be produced, as in figure 8. We call a tree that branches off to the left as in figure 8 (*a*), a regressive structure because the mechanism first moves down the stem, expanding each node, and then moves back up and completes each branch. The longer a regressive structure becomes, the more temporary storage it requires. On the other hand, a structure branching off to the right, as in figure 8 (*b*), a progressive structure, can be extended indefinitely without requiring more than a minimum of temporary storage. It is only because of the possibility of indefinitely long progressive structures that our device can produce any sentence out of an infinite set of sentences.

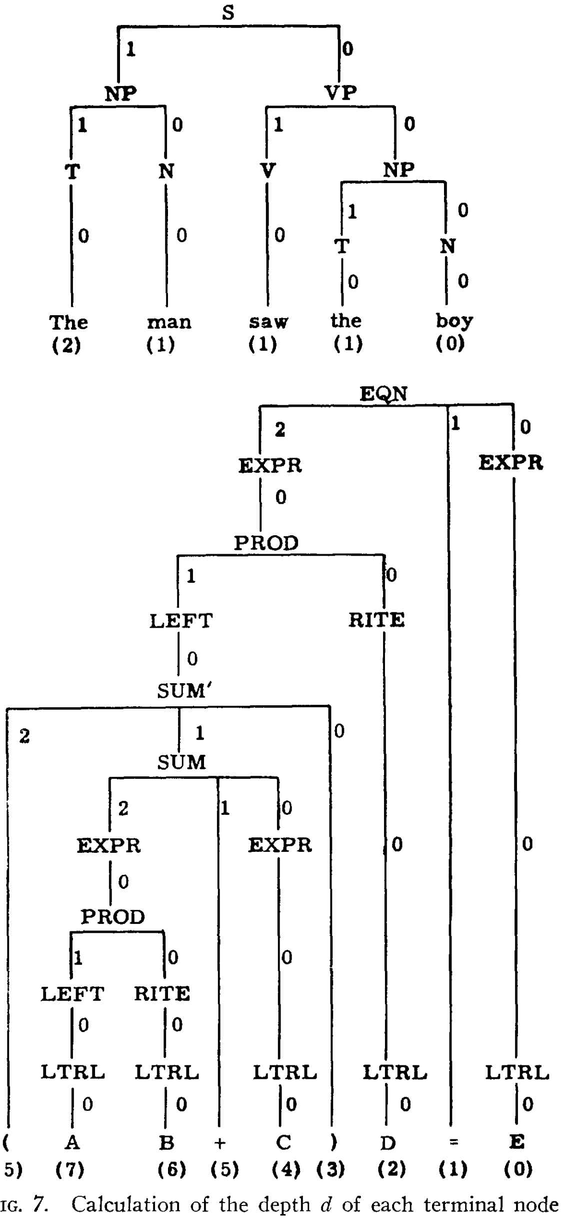


FIG. 7. Calculation of the depth d of each terminal node in a sentence. $D = d_{\max}$, the depth of the sentence, is 2 for the English sentence, and 7 for the algebraic expression.

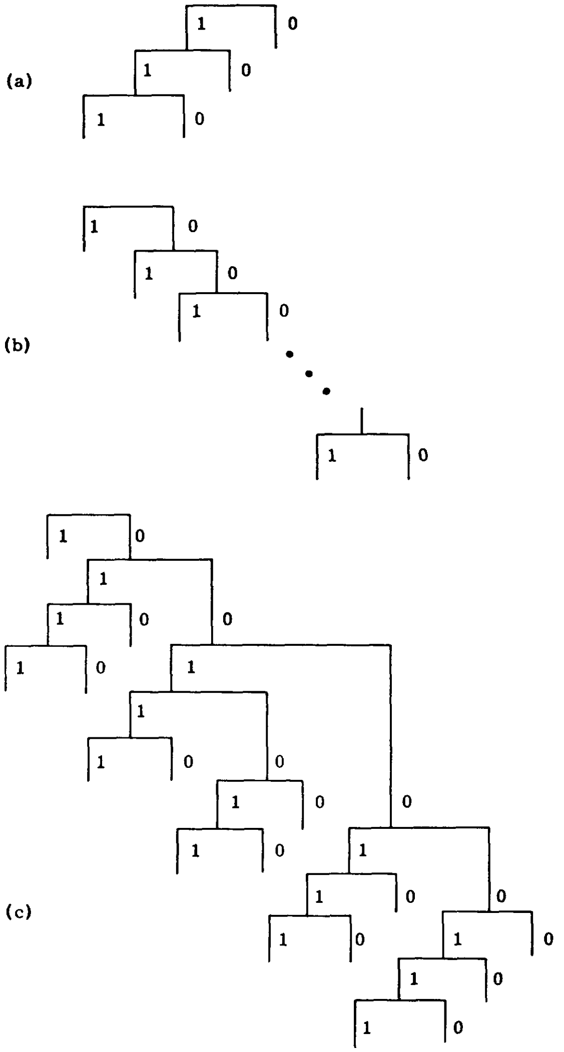


FIG. 8. (*a*) A regressive structure of depth three requiring a temporary memory for three symbols. (*b*) A progressive structure of depth one requiring a temporary memory for one symbol. (*c*) A complicated structure of depth three.

THE HYPOTHESIS

Although our device could work for a written language like algebra only under the simplifying assumption of an infinite temporary memory, we do not yet know whether it could work for a spoken language like English. It will be strictly adequate for English if it turns out that the grammar of English is strictly well behaved, that is, if the grammar of English is restricted in such a way that temporary memory can never become exhausted. But if the grammar of English is not strictly well behaved, the device might still be practically adequate if it has a temporary memory that is large enough to be practically infinite, i.e., large enough so that it would become exhausted only for an insignificant fraction of the sentences it is actually called upon to produce.

Psychologists have measured what they call the human span of immediate memory. They have found that we can comprehend at one time, remember, and repeat back approximately seven items, approximately seven random numbers or random words. Miller[5] has given an interesting discussion of this phenomenon. It seems that we can attend to only a few things at once. It is tempting to identify the temporary storage of our model in the case of spoken language with the facility that we use for immediate memory. If this identification is correct, the span of immediate recall can be used to estimate an upper limit to the maximum possible depth of English utterances. It is an upper limit because other kinds of processing besides that which our model represents may go on during speech.

A temporary memory of seven, plus or minus two, is very small. We have seen that a memory of seven would be required for our device to produce the simple equation $(AB+C)D=E$. One would expect English sentences of this depth to be frequent if there is no depth limitation. But since it is possible for a speaker to use the language fluently, with very infrequent failure of the type attributable to the premature filling of a temporary memory, we would expect that there would be an easily observable effect of the depth limitation in the grammar of English. In order to test this idea, we propose the following hypothesis and then make a number of specific predictions for comparison with observations of English structure.

a) Although all languages have a grammar based on constituent structure,

b) the sentences actually used in the spoken language have a depth that does not exceed a certain number

c) equal or nearly equal to the span of immediate memory (presently assumed to be 7 ± 2).

d) The grammars of all languages will include methods for restricting regressive constructions so that most sentences will not exceed this depth,

e) and they will include alternative constructions of lesser depth that would maintain the power of expression of the language.

f) For all languages, much of the grammatical complexity over and above the minimum needed for the signaling function can be accounted for on this basis.

g) When languages change, depth phenomena will frequently be involved, and will often play an important role.

In this hypothesis, part (*a*) is a restatement of our first assumption. Part (*b*) follows from our second assumption, that the constituent-structure tree is to be built from the top down and from left to right, coupled with our assumption of a finite temporary memory. The depth limitation will not apply to algebra, for example, because it is not a spoken language. The user has paper available for temporary storage. If (*b*) is correct, the value of the maximum depth can be determined from an examination of the grammar and sentences of a language. In (*c*), the identity of this limit with the span of immediate recall can perhaps be tested by suitably devised psychological tests. Parts (*d*) and (*e*) represent features that a well-behaved grammar might be expected to have. In order that we may be able to recognize the kinds of structures to be expected in well-behaved or almost well-behaved grammars, a number of detailed predictions will be presented. Part (*f*) asserts that depth considerations are among the most important factors in the grammar of any language. Part (*g*), a diachronic statement, would seem to follow if the previous synchronic parts of the hypothesis turn out to have any truth in them. It proposes that depth phenomena be added to the list of already known factors affecting language change. A depth factor in language change should be easily observable if it exists.

WELL-BEHAVED GRAMMARS

A constituent-structure grammar that is well behaved from the point of view of our model can be expected to contain some of the following structural characteristics.

[5] Miller, George A., Human memory and the storage of information, *I.R.E. Transaction on Information Theory* **IT-2**: 129–137, 1956.

A. METHODS FOR LIMITING REGRESSION

We would expect a well-behaved grammar to include methods for limiting regression to the maximum amount allowed by the temporary memory. It is important in languages to allow as much regression as possible in order to maintain expressive power: The greater the permissible depth, the more sentence structures are possible for sentences of any given length except short ones. Two methods of restricting regression come to mind.

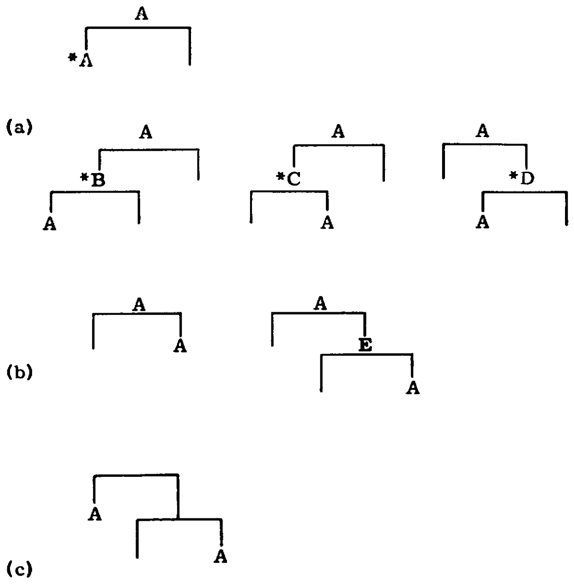


FIG. 9. Application of the method of an ungrammatical first step to prevent regression. The starred constructions in (*a*) would be ungrammatical, at least when they are constituents of *A* as indicated.

1. *An Ungrammatical First Step*

Regression can be prevented if it is ungrammatical to reapply a rule along the same series of branches unless its new node of application is connected to the old one through extreme right branches—a progressive connection. This is shown in figure 9. Here *A* is the rule that is reapplied. An ungrammatical first step would make the starred constructions at (*a*) ungrammatical. The reapplication of a rule at a depth no greater than its previous application, as at (*b*), can safely be made. Reapplications along different series of branches, as at (*c*) is also safe. Sentences unlimited in length could be expected to be formed on the pattern of (*b*).

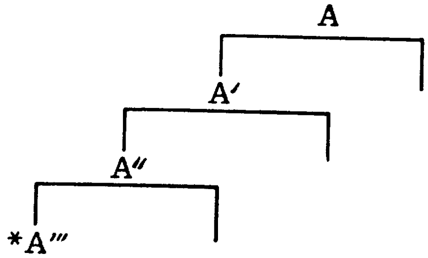


FIG. 10. Application of method of restricted relabeling to limit regression.

2. *Restricted Relabeling*

A regressive branch can be allowed to grow for a certain length and then stopped if some method is used for automatically counting or keeping track of the number of regressive steps so that the n^{th} step can be prevented. Since our simple model does not include a counter attached to the mechanism that will give an alarm like the bell on a typewriter when the temporary memory is nearly full, this counting must be handled in the organization of the grammar rules. One method of doing this is to have some grammatical feature that relabels a rule each time it is reapplied. Figure 10 shows the rule *A* applied three times, each time grammatically relabeled by adding a prime. If there were no *A''' in the language, a fourth step would be ungrammatical, and a regressive construction of too great a depth would be ruled out.

B. METHODS FOR CONSERVING DEPTH

We would expect that constructions of less depth would be preferred over equivalent constructions of greater depth.

1. *A Preference for Binary Constructions*

When a construction with three constituents is represented by a ternary rule, the left-most constituent node appears at a depth that is greater by two than that of the construction. If it is reinterpreted, however, as a binary progressive

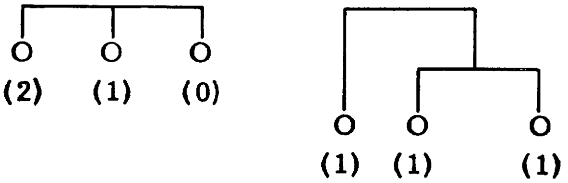


FIG. 11. A binary progressive structure is more conserving of depth than a ternary structure.

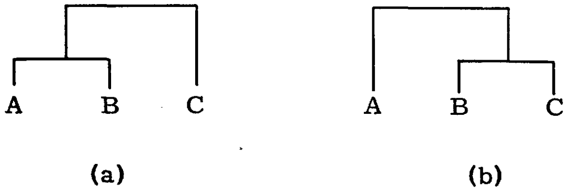


FIG. 12. Changed pattern of modification.

structure, as in figure 11, nowhere is the depth increased by more than one. Two rules are required in the grammar, however, instead of one, but this is a small price to pay for the saving in depth because the permanent memory in which the grammar is stored has a much larger capacity than the temporary memory. We would therefore expect that binary rules would predominate in a well-behaved grammar, perhaps even to the almost complete exclusion of ternary or other larger rules.

2. *Changed Pattern of Modification*

We would expect developments in a language whereby a sentence or a phrase with a regressive structure would be reinterpreted occasionally or habitually as a structure of lesser depth. We might expect, for example, that the phrase *A B C* in figure 12 (*a*), which originally had the major constituent break between *AB* and *C*, would be used occasionally or habitually in changed form as in figure 12 (*b*), where the major constituent break is now between *A* and *BC*. In this way a regressive structure of depth two is changed into a progressive structure of depth one.

3. *Word Building*

We would expect compound words to be built up from regressive structures of separate words. In this way the compound word becomes a single node, and is entered into the permanent memory as a separate lexical item, and can be used as if it had no internal structure. Thus the regressive steps involved in its internal structure are effectively circumvented. In figure 13 compound words would tend to be formed from the individual words *A, B, C, D* and *E*. *E,* which is a post-modifier of a whole regressive phrase, would tend to become a depth-reducing suffix so as to form the compound words *DE* or *ABCDE*. This would be particularly likely if *E* belonged to a small class of words, where the price in additional lexical items would be small. Similarly, *A,* which is at the other end of a regressive structure, would tend to become a depth-reducing prefix, particularly if it were a member of a small class of words modifying *B*. On the other hand, *P,* although like *E* it modifies a whole phrase, would not tend to become a prefix because this would increase the vocabulary without gaining a reduction in depth. Similarly, *T* would not tend to become a suffix, although it modifies its neighbor as does *A*.

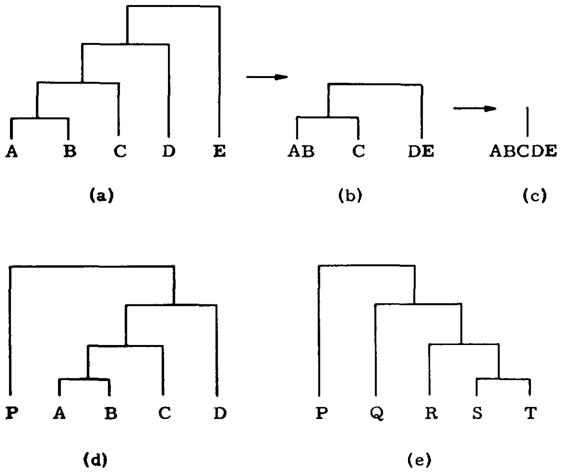


FIG. 13. Regressive structures leading to the building of compound words from the individual words *A, B, C, D,* and *E,* but not from *P, Q, R, S,* or *T.*

C. METHODS FOR MAINTAINING THE POWER OF EXPRESSION

We would expect methods to be developed which would allow phrases involving regression to be postponed to a point of application of smaller depth.

1. *Structure Reversal*

One way of postponing a regressive phrase is to transform the phrase of which it is a part so that the regressive phrase is placed farthest to the right. Figure 14 gives an example. *A* is

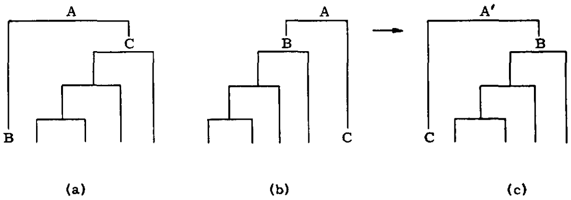


FIG. 14. Structure reversal can reduce depth by postponing a regressive phrase.

composed of constituents *B* and *C*. When *B* is extensively modified, particularly by a regressive structure as at (*b*), *A* is transformed into *A'*, a structure synonymous with *A*, but with a reversed order of constituents. *B* is thus moved from a node of depth one to a node of depth zero. On the other hand, if it is *C* that is extensively modified, as at (*a*), the transformation does not take place. There would be a tendency for a structure like (*b*) to become ungrammatical, resulting in obligatory use of the transformed structure.

2. *Discontinuous Constituents*

When there are other structural reasons for not moving a regressive phrase to a point of lesser depth, the device of discontinuous constituents can be used to postpone it effectively: If, for example, as in figure 15 (*a*), construction *A*, made up of constituents *B* and *C* utilizes the relative order of *B* and *C* to signal their function in *A*, the structure can be transformed to that at (*b*). Here B_1 is a special place marker inserted before *C* to preserve the order signal, and B_2 is the regressive phrase, now postponed to a point of lesser depth and tied to B_1 by some grammatical device like agreement, marking it as part of the discontinuous constituent *B*. But if *B* is already composed of two constituents as in (*c*), the regressive one can be postponed, leaving the other one to carry the word-order signal as in (*d*).

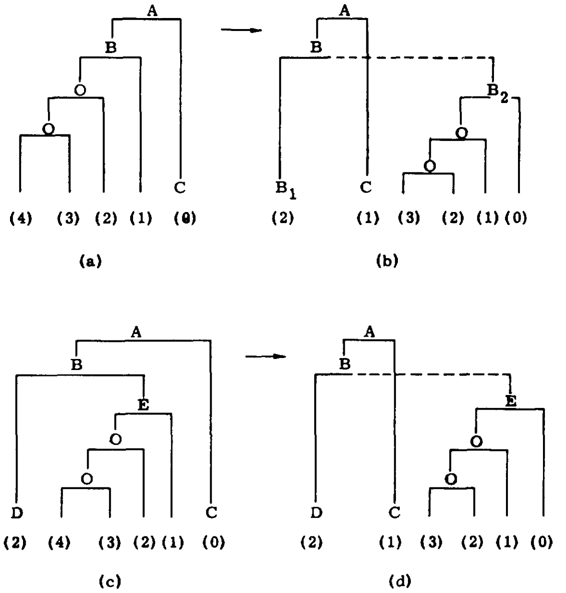


FIG. 15. A discontinuous constituent can reduce depth by postponing a regressive phrase. The numbers in parentheses indicate node depths with respect to *A*.

A LOOK AT ENGLISH

The following is a listing of a number of features of English morphology and syntax, together with suggested interpretations in light of the hypothesis. It should be kept in mind that the structures and analyses offered are tentative. A complete and consistent grammar of English has not been worked out.

In all the examples that have been looked at thus far, the constructions have either one or two constituents. English seems to have an essentially binary nature, in line with the hypothesis.

The mechanism that English uses to limit depth is a restricted relabeling scheme involving sentence, clause, noun phrase, primary attribute (adjectival), secondary attribute (adverbial), and tertiary attribute (adverbial). As an example, we can examine the sentence in figure 16.

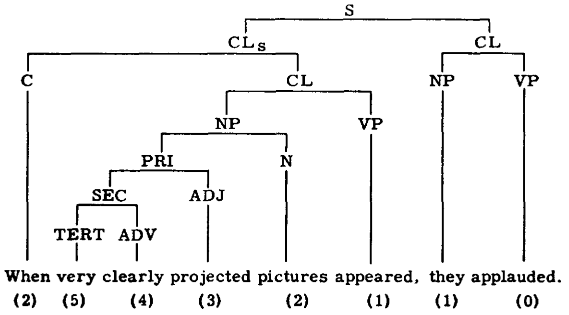


FIG. 16.

Here we have a depth of five, counted off by the different types of nodes labeled *S, CL, NP, PRI, SEC, TERT*. As will be seen later, a certain amount of variation and expansion of this basic pattern is possible. Clauses may be piled up more than two deep. It is also possible, in certain cases, to go beyond the three types of attributes. But in general, the depth limitation is rather well imposed.

It is well known that there is a syntactic distinction connected with the sequence verb, noun, adjective, adverb, that need not parallel a semantic distinction. Nor is the part of speech distinction needed logically to maintain a hierarchial pattern of modification or subordination. A single method of subordination would suffice—for example, word order as in the Łukasiewicz notation—but then there would be no grammatical limit to the amount of subordination allowed. It thus seems obvious that these distinctions in English have as their purpose the provision of a restricted relabeling scheme.

Sentences, clauses, phrases, and attributes can each be coordinated indefinitely in progressive structures (fig. 17).

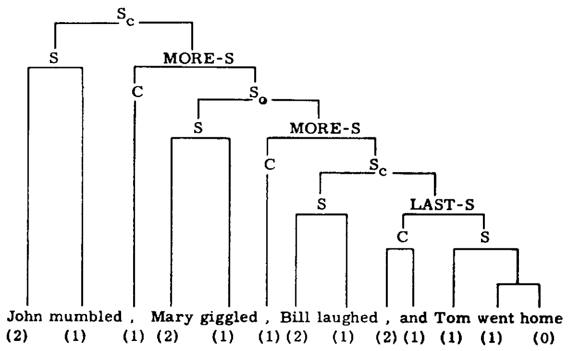


FIG. 17.

The following coordinations have a similar structure.

The men, the women, the children, the dogs, and the cats.
Red, white, blue, yellow, and black.

Whereas there is a limit to the number of regressive steps in a sentence, enforced by the relabeling scheme, there is no grammatical limit to the number of progressive steps that can be taken by means of coordination. The pattern can be extended as far as the speaker desires; certainly beyond seven steps.

Turning now to complex sentences, we observe the possibility of structure reversal without a significant change in meaning. The subordinate clause may either precede or follow the main clause.

He does it whenever they ask him.
Whenever they ask him, he does it.

As can be seen from figure 18, the first clause, whether it is the main or the subordinate clause, starts at a depth of one, whereas the second clause starts at a depth of zero. Of the two, however, only the progressive structure can be used with an indefinite number of clauses. We can have

He cried because she hit him because he called her names because she wouldn't give him any candy.

but not the regressive structure:

Because because because she wouldn't give him any candy, he called her names, she hit him, he cried.

If we wish to string the clauses together in the reverse order, we use, instead, a different subordinating conjunction that allows a progressive connection.

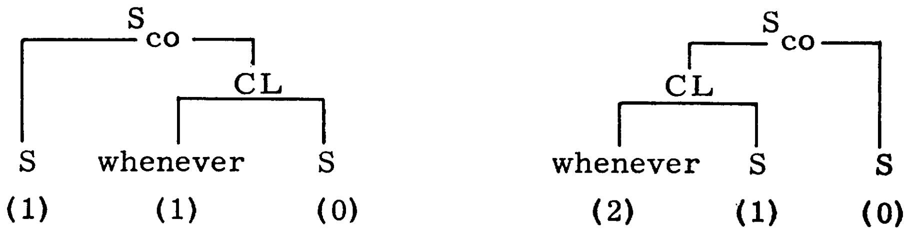


FIG. 18.

She wouldn't give him any candy, so he called her names, so she hit him, so he cried.

Looking now at the internal structure of a clause or simple sentence, we see that we do not have a first split into three constituents, a subject phrase, a verb phrase, and an object phrase, which might be suggested by logic (fig. 19). Neither do we have a split into verb, as head of the construction, and attributes, subject and object, as might also be suggested by logic (fig. 20).

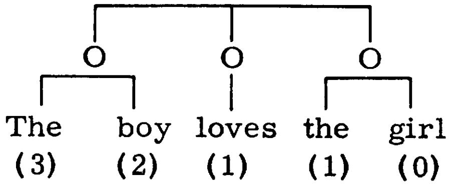


FIG. 19.

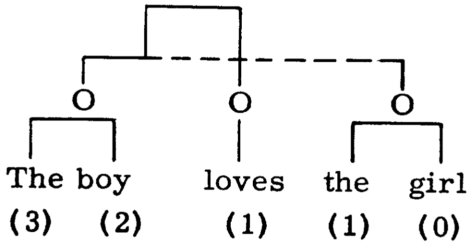


FIG. 20.

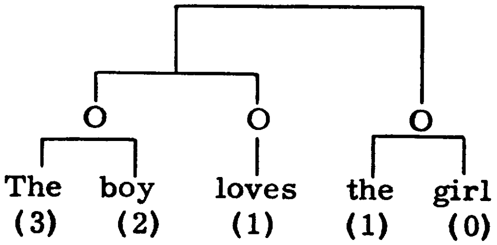


FIG. 21.

Neither do we have a split between the subject and verb together, and the object, as might be expected on the basis that there is subject-verb agreement (fig. 21). But we have, instead, a split between the subject, and the verb and object taken together (fig. 22). This pattern gives a depth of two. According to the hypothesis it would be the favored one because the others would give a depth of three.

When the sentence is put into the passive, however, we do not retain the same connection that we had in the active (fig. 23). Instead, the pattern of modification is changed, re-establishing the subject-predicate split and retaining the advantage of a depth of two (fig. 24).

When there are two constituents after the verb, it is sometimes the case that the order of these constituents can be changed without essentially changing the meaning.

He gave the child a toy.
He gave a toy to the child.
He called the girl up.
He called up the girl.

But although the resulting sentences are essentially synonymous, there is an important and frequently noted difference. If one of the constituents is a pronoun, it is almost invariably given the first position near the verb, either because the other possibility is stylistically poor, or because it is actually ungrammatical.

He gave her a toy. (preferred?)
He gave a toy to her.

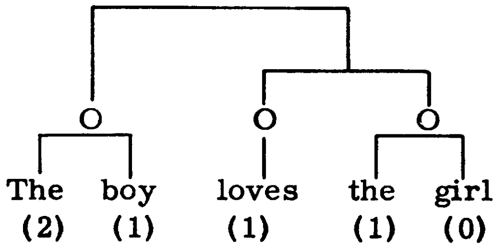


FIG. 22.

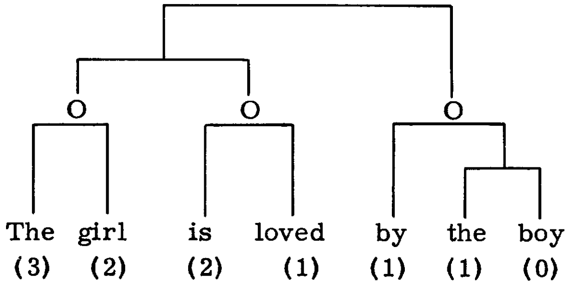


FIG. 23.

He gave it to the child.
He gave the child it. (ungrammatical)

He called her up.
He called up her. (ungrammatical)

If one of the constituents is a long and extensively modified clause, it almost invariably is placed in the last position. Grammarians like to speak of these clauses as "heavy" and it is noted that there is a tendency for heavy elements to come at the end. When they come first, the construction is said to be "top heavy," as if some vague principle of balance were involved.

(*a*) He gave the girl a box of candy he got in New York while visiting his parents for ten days around Christmas and New Year's. (preferred)
He gave the box of candy he got in New York while visiting his parents for ten days around Christmas and New Year's to the girl.

(*b*) He gave the candy to the girl that he met in New York while visiting his parents for ten days around Christmas and New Year's. (preferred)
He gave the girl that he met in New York while visiting his parents for ten days around Christmas and New Year's the candy. (ungrammatical?)

(*c*) He called up the girl that he met in New York while visiting his parents for ten days between Christmas and New Year's.
He called the girl that he met in New York while visiting his parents for ten days around Christmas and New Year's up. (ungrammatical)

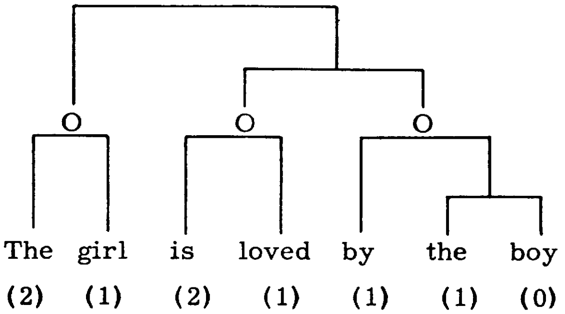


FIG. 24.

From the point of view of our hypothesis, these phenomena have an easy explanation. No matter how the predicate construction is broken down into constituents, whether we have a ternary structure, a binary progressive one, a binary regressive one, or a structure with discontinuous constituents, the middle constituent starts at a depth that is one greater than that of the last constituent. It would be expected that devices would be found in the language that would ensure that a potentially deep expression would start at the minimum depth, and therefore it would be last. The devices of placing a pronoun first or of postponing a modified expression would serve this purpose well.

It is interesting that, when a meaning difference is associated with the order of the constituents following the verb, something else happens. The "top heavy" construction that would seem to be inevitable is avoided by an additional syntactic complication.

He saw through the matter.
He saw the matter through.

If we have an extensively modified expression, the first of the preceding examples retains its order.

He saw through the matter that had caused him so much anxiety in former years when he was employed as an efficiency expert by the company.

But when the extensively modified construction should come first in order to maintain the meaning difference, we would have a "top heavy" construction:

He saw the matter that had caused him so much anxiety in former years when he was employed as an efficiency expert by the company through.

What is usually done, instead, is to make the long modified constituent into a construction with discontinuous constituents, retain the noun head in its proper position to carry the meaning distinction, and postpone all of the rest of the modified construction to a favored position of lesser depth.

He saw the matter through that had caused him so much anxiety in former years when he was employed as an efficiency expert by the company.

Another somewhat similar feature of sentence or clause organization is the placing of an interrogative pronoun in first position. In this position it cannot add to the depth of later elements in the sentence. These later elements are then left at a minimum depth ready for extensive modification if required.

Besides coordination, there are other progressive structures that allow indefinite expansion. At the clause level there are a number of types of object clauses. A few will be given as examples of the pattern.

(*a*) A "what" object clause:

He knows what should have been included in what came with what he ordered.

(*b*) A "that" object clause:

John said that Bill said that Paul said that Jim had won the prize.
I know that you know that I know that you told the secret.

(*c*) A "him doing" object clause:

I imagined him listening to the announcer reporting Bill catching Tom stealing third base.

(*d*) A "him do" object clause:

I watched him watch Mary watch the baby feed the kitten.

Although these sentences seem awkward, they all can be extended with no grammatical limitation. One can obtain less awkward sentences by combining elements from several of these types into one sentence.

Bill knows that Paul said that I imagined Mary watching the baby feed the kitten.

An object clause may itself have an object clause, and this may in turn have an object clause, and so on.

The situation with subject clauses is different. A subject clause is tied to its position before or after the verb, to maintain the meaning difference signaled by word order.

What he said is true.
Is what he said true?

In either case, the node representing the whole subject clause appears at a depth of one. Its subject clause, in turn, if it has one, appears at a depth of two. Each additional subject clause would appear at a depth of one more than does the clause that it is the subject of. Continuing without limit in this way, we would obtain a sentence of unlimited depth.

That it is obvious isn't clear.
That that it is true is obvious isn't clear.
That that that they are both isosceles is true is obvious isn't clear.

The first of these examples is certainly grammatical. The second, with its subject clause having a subject clause is very awkward, but it is possible to find an intonation pattern that fits and that can serve as a method of relabeling. It is difficult to find an appropriate intonation pattern for the third example.

A further factor in the ungrammaticalness here may be the repetition of the word *that*. If this is a factor, it is not simply the repetition of the word, but of the word with the same subordinating function, for we can have:

I believe that that, that that child said isn't quite true.

Here again we need to use an appropriate intonation pattern.

In order to preserve the expressive power of the language, a progressive alternative is available for subject clauses within subject clauses. The alternative consists of a postponement transformation that leaves a dummy *it* in position before the verb where it can carry the word order signal. The clause is then postponed until after the predicate, forming a discontinuous construction with the *it*. In this position the subject clause now finds itself at depth zero, the favored position that the object clauses had occupied, and here an indefinite expansion is possible.

It isn't clear that it is obvious.
It isn't clear that it is obvious that it is true.
It isn't clear that it is obvious that it is true that they are both isosceles.

There is a similar situation with a "what" subject clause:

What it would buy in Germany was amazing.
What what it cost in New York would buy in Germany was amazing.
What what what he wanted cost in New York would buy in Germany was amazing.

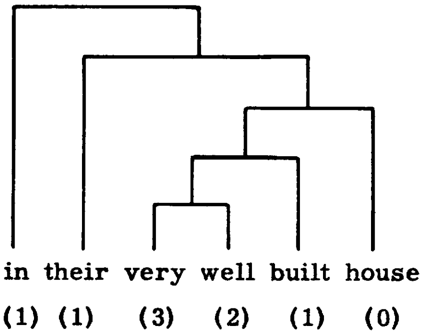


FIG. 25.

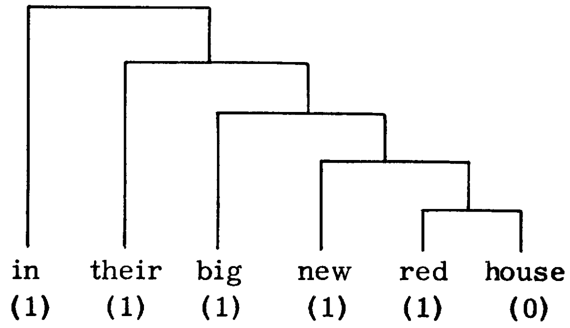


FIG. 26.

The progressive alternatives that the language offers are a little more complicated here. As a first step, the main subject clause can be postponed by a discontinuous construction as was done for the "that" clauses:

It was amazing what what what he wanted cost in New York would buy in Germany.

The second subject clause can be effectively postponed by using instead an object clause after a passive verb:

It was amazing what could be bought in Germany for what what he wanted cost in New York.

Finally, the last subject clause can be postponed by making it the object of the preposition *of* in a different type of nominalization. We achieve a fully grammatical sentence:

It was amazing what could be bought in Germany for the cost in New York of what he wanted. (". . . for the cost of what he wanted in New York" is ambiguous.)

Much more could be said about the structure of nominal clauses. We shall, however, move down one rung on the relabeling ladder and talk about noun phrases and their attributes.

The first thing to notice is that prepositions and determiners do not add to the depth of nominal expressions (fig. 25). Here we see that the preposition, the determiner, and the modified noun form a progressive structure. If there are several adjectives, they generally can form an accumulative non-coordinated pattern of modification that is progressive in contrast to the regressive relabeling scheme of noun, adjective, adverb, adverb (fig. 26). Here, the classes of adjectives are quite nebulous compared to the adverb-adjective distinction, and the number of adjectives seems not to be grammatically, but semantically, limited if it is limited at all. A counting scheme is not needed to limit the depth.

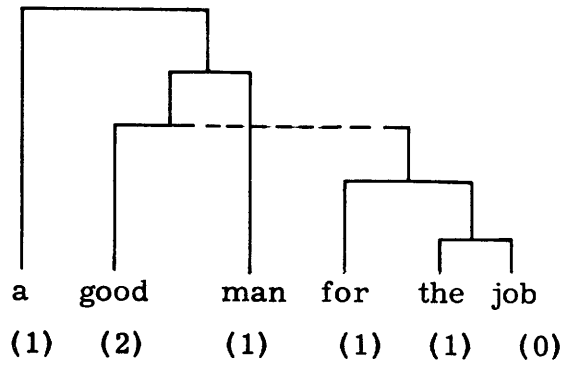


FIG. 27.

If the adjective is not a single word, but a phrase, the regular pattern is to place the adjective before the noun in its regular position, but to postpone the rest of the phrase in a discontinuous construction (fig. 27). In this way, the noun is pushed one deeper, but the phrase that can be extensively modified is pushed up to the same level as the whole noun phrase. When lines cross in our diagrams, the depth of the postponed node is reduced by one, and the depth of the other is increased by one.

The same type of construction is used if the determiner involves extensive modification (fig. 28).

The other type of adjectival modifier is the relative clause. It always is placed in the position of minimum depth after the noun. In this position it is possible to have a noun in the relative clause further modified by a relative clause, and so on, indefinitely. The classic example is, of course:

This is the dog, that worried the cat, that killed the rat, that ate the malt, that lay in the house that Jack built.

But with relative clauses built on subjects, we have much the same problem that we had with subject clauses. We have a regressive structure, and it is grammatically limited. We cannot have:

This is the malt that the rat that the cat that the dog worried killed ate.

We shall return to this case later. Now, let us turn our attention to the grammatical alternative that the language offers:

This is the malt, that was eaten by the rat, that was killed by the cat, that was worried by the dog.

By converting to the passive with a "by" phrase, an example of structure reversal, we have been able to form a progressive structure. This structure reversal is very important in English. It provides a strong reason for the existence of the form of the passive alongside the active, with which it is perhaps synonymous. As a further example of the utility of structure reversal in English, we can take the following sentences:

A pair of opposed fingers operate the said rocker lever.
The pair of opposed fingers extend from the pitman.
A crank stud oscillates the pitman.
The crank stud extends eccentrically from a shaft.
The shaft is rotatably mounted in a bracket.
A worm gear is on the shaft.
A worm pinion drives the worm gear.
The motor has a drive shaft.
The worm pinion is mounted upon the drive shaft.

By means of various types of structure reversal, it is possible to transform some of these sentences so that all of them can be put into one sentence having relative clauses in a progressive construction. We obtain the following sentence from a U. S. patent:

The said rocker lever is operated by means of a pair of opposed fingers
which extend from a pitman
that is oscillated by means of a crank stud
which extends eccentrically from a shaft
that is rotatably mounted in a bracket
and has a worm gear thereon
that is driven by a worm pinion
which is mounted upon the drive shaft
of the motor.

The main type of structure reversal used in the preceding example is the passive construction, although other types are also represented. This sentence is admittedly extreme, but without structure reversal, one would have the monstrosity shown in figure 29.

We have already seen that adverbial clauses can either precede or follow the main clause with-

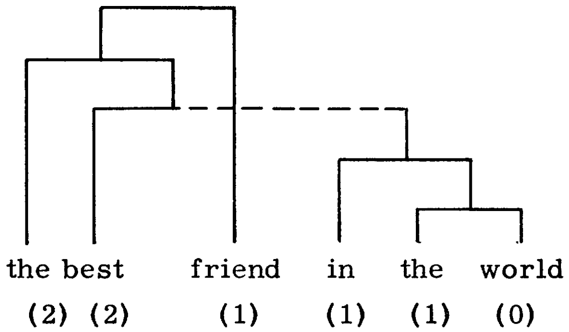


FIG. 28.

out a meaning change. It is interesting to note that in algebra, where structure reversal is not needed, we do not have:

"$A\ \theta\ B$" with the same meaning as "$B - A$," or
"$A \backslash B$" with the same meaning as "B/A."

But in English, we have:

"A subtracted from B" as well as "B minus A," and
"A divided into B" as well as "B divided by A."

Of course there are other functions for the passive. A very important one is to provide a subject when it is desirable not to mention what would have been the subject of the active:

Yield curve data for the ground state were taken with a broad-range spectrograph using either. . . .

This allows the experimenter to remain in the background. But notice what happens if this noun phrase that has been moved to a deeper position up ahead of the verb is extensively modified. A discontinuous construction is frequently used that moves the extensive modifiers back to their original position of lesser depth:

In a recent paper measurements (were presented) of the effect of alloying on the superconductive critical temperature of tin.

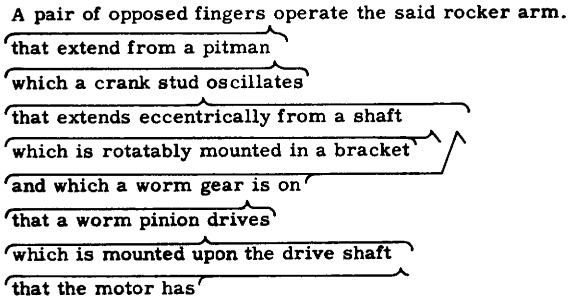
We can now return to the regressive constructions involving relative clauses.

This is the malt that the rat ate.
This is the malt that the rat that the cat killed ate.
This is the malt that the rat that the cat that the dog worried killed ate.

Each example is worse than the one before. This is the same phenomenon that we saw in the case of the subject clauses. The grammaticalness or lack of it for these examples seems to hinge on the pattern of pitch, stress, juncture, and speed. Pike[6] has noted that by using the proper patterns, we can distinguish in speech between

"$A - (B + C)$" and "$(A - B) + C$," and between "$A - [B + (C \times D)]$" and "$A - [(B + C) \times D]$" and "$[A - (B + C)] \times D$,"

but there is a definite limit to how far we can go. The usage in reading mathematics is probably merely the restrictive relabeling carried over from English, where it is used for marking inserted clauses and the like.



A pair of opposed fingers operate the said rocker arm.
that extend from a pitman
which a crank stud oscillates
that extends eccentrically from a shaft
which is rotatably mounted in a bracket
and which a worm gear is on
that a worm pinion drives
which is mounted upon the drive shaft
that the motor has

FIG. 29.

The children, that I see, turn around.
The children, that I see turn around, . . .
The children that I see, turn around.
The children that I see turn around, . . .

In these examples, sentence *vs.* noun phrase, and restrictive *vs.* nonrestrictive relative clause, are marked.

That he said it isn't true, . . . (clause)
That he said it, isn't true. (sentence)
That it is clear, is true.
That,, that it is clear, is true,, is amazing.
That what is clear is true, is amazing.

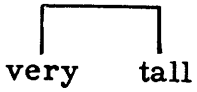
The process of relabeling, provided by patterns of pitch, stress, juncture, and speed, is aided considerably if the clauses are maximally different in form rather than all nearly the same as in the *house that Jack built* examples. The more plausible examples of regressive clause constructions seem to involve this further difference that can serve as auxiliary relabeling:

That what the poem the woman he knows wrote implies, is obscure, is obvious.

It is difficult to determine what the actual limit is for regressive clause constructions. It is probably three or four clauses, making two or three steps down. One could have three steps down and still not go over a depth of seven if the clauses themselves did not involve deep noun phrases. English seems to take an ambivalent position on the amount of regression taken with clauses. In the first place, there is an advantage in expressive power in allowing as much clause regression as possible, but with each additional step down there is additional danger that a noun phrase in one of the clauses will go over the depth limit. The following sentence has a depth of eight:

If what going to a clearly not very adequately staffed school really means is little appreciated, we should be concerned.

[6] Pike, K. L., *The intonation of American English*, 72, University of Michigan Press, 1945.



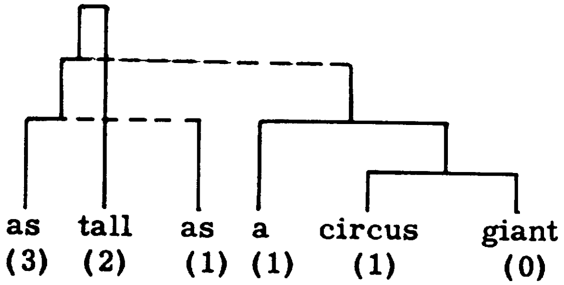


FIG. 30.

It is very difficult to find sentences of this or greater depth that would be plausible as casual utterances, or would be immediately understandable to the hearer. But the sentence above can be rephrased so that it has a depth of two or three and a "length" of about eighteen progressive steps.

We should be concerned if there is little appreciation of what it really means to go to a school that clearly isn't very adequately staffed.

We can now go down to the lower steps in the relabeling scheme—the secondary and tertiary modifiers or adverbs. We find again, as we did with the adjectives, that an extensively modified expression leads to some evasive tactic that causes the extended modifier to bob up to a lesser depth, usually up to the original depth of the noun phrase itself.

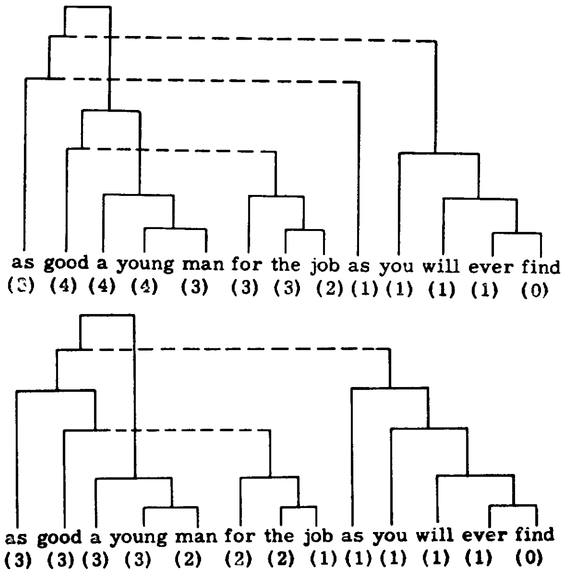

as good a young man for the job as you will ever find
(3) (4) (4) (4) (3) (3) (3) (2) (1) (1) (1) (1) (0)
as good a young man for the job as you will ever find
(3) (3) (3) (3) (2) (2) (2) (1) (1) (1) (1) (1) (0)

FIG. 31.

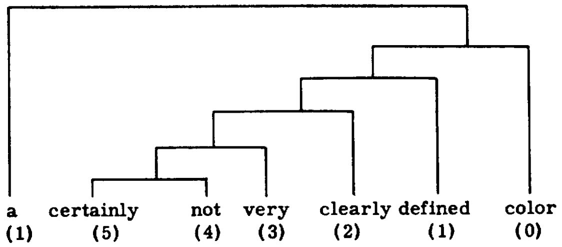


FIG. 32.

A very tall man
A taller man
(A) taller (man) than a circus giant (discontinuous)

or

A man taller than a circus giant (structure reversal)

but not

A taller than a circus giant man

Also we have:

A good job
A good enough job
A job good enough to pass inspection (structure reversal)

or

(A) good enough (job) to pass inspection (discontinuous)

but not

A good enough to pass inspection job

Structures get rather complicated down here at the adverb level as the language strives to keep itself above water (fig. 30). Then we have:

a very tall young man
a young man as tall as a circus giant
as tall a young man as a circus giant

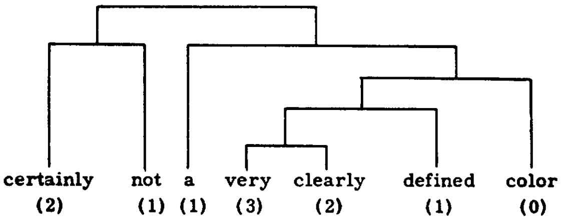


FIG. 33.

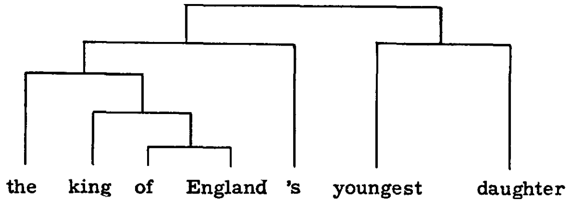


FIG. 34.

but, of course, not

an as tall as a circus giant young man

In the sentence

He is as good a young man for the job as you will ever find.

we have a number of complications for conserving depth—so many that it is quite difficult to be sure of the proper analysis. The constituent structures shown in figure 31 can be produced by the mechanism.

We have seen that the determiner forms one constituent, with the rest of the noun phrase forming another (fig. 32).

There is a strong tendency for the deepest modifiers from the regressive structure to be moved over so that they will modify the whole noun phrase, even if the meaning is thereby somewhat changed. Here they find themselves at a lesser depth (fig. 33).

The determiner may sometimes be a possessive (fig. 34).

Besides the well-known problem about the role of the *'s*, which seems to be at the same time affixed to a word and to a phrase, there is another difficulty. The structure might be regressive (fig. 35). If the analysis is correct, the structure has a depth of nine.

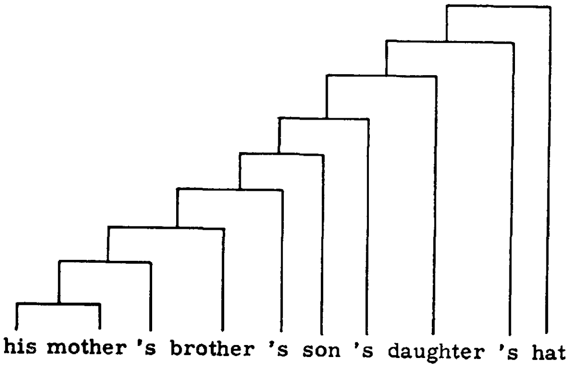


FIG. 35.

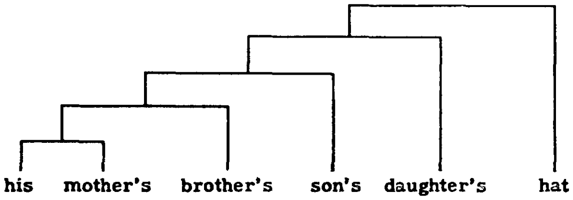


FIG. 36.

If we consider the *'s* as part of the word, so that the whole unit is brought out from a permanent memory already formed, the depth is reduced to five (fig. 36). According to this conception, we should also have figure 37, where we have set up a relabeled noun phrase, *NP's* which can then function as a determiner. There is a saving in depth of one unit.

In the case of *his mother's brother's son's daughter's hat,* we have the more convenient if not as explicit *his uncle's granddaughter's hat.* We would run to a depth of eight if we had *John's father's father's father's father's father's father's father came over from England.* Although there is no grammatical limit to such a string, and in this respect the grammar of English is not well behaved, it is difficult to conceive of a person uttering such a sentence without resorting to some auxiliary device, like counting on his fingers, or breaking up the string into groups of three by intonation patterns. In this case we may have compound words: *John's father's-father's-father's father's-father's-father's father.* If this is true, the whole sentence would have a depth of four.

The *'s* genitive in English does not have as wide a use as the *of* genitive. Perhaps the more extensive use of the *of* genitive in English is a result of its progressive nature.

The fact that the two genitive markers in English differ, the one that occurs in final position being a suffix, and the one that occurs in initial position being a separate word is predicted by the

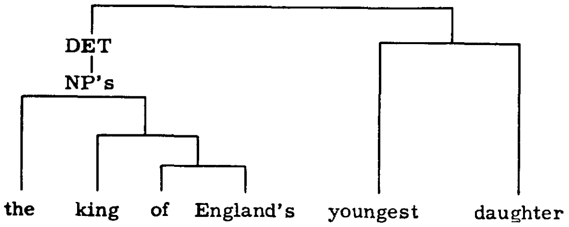


FIG. 37.

hypothesis. A somewhat similar situation holds for pairs like:

to swim	swimming
to construct	construction

In fact, English has a large class of suffixes like *-tion, -ment, -ize, -en, -ness* and so on, that have as their function the changing of verbs to nouns, nouns to adjectives, etc., and a number of particles that are words, not prefixes, with the analogous function of converting nouns to noun phrases, noun phrases to prepositional phrases, sentences to clauses and the like. Derivational particles that occur initially will not add to the depth, but those that occur finally will, unless they disappear as part of a word. This is clear from the following. We have figure 38

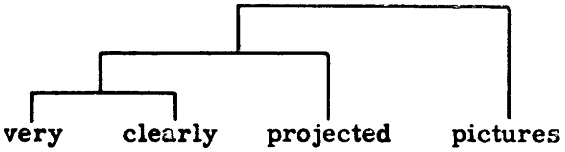


FIG. 38.

and not figure 39

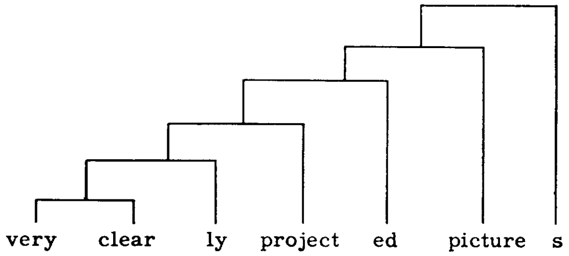


FIG. 39.

even though the *s* might be thought of as making the whole noun phrase plural, the *ed* as making the verb *very clearly project* into a participle, the *ly* as making the adjective *very clear* into an adverbial.

As regards prefixes, English follows the predictions of the hypothesis in the few really productive prefixes that it has. We have pairs like:

renegotiate:	negotiate again
reconstruct:	construct again
co-occur:	occur together

One advantage of word building in a language is that morphemes can be put together in a productive way to express new things if this is done at a fairly small depth. Then, if the concept is referred to frequently, it can be put into the permanent memory as a separate word ready to be brought out and used with a minimum depth requirement.

SUMMARY AND CONCLUSION

A model of sentence production that can easily be mechanized has been set up. On the basis of the model and certain simple assumptions relating to the order in which the constituents of the sentence are expanded, and on the basis of an assumption that the temporary memory of the device is limited, an hypothesis of a depth limitation in language has been proposed. This hypothesis leads to a number of specific predictions about the types of syntactic structures to be found in languages.

A cursory examination of the structure of English discloses that, indeed, much of the complexity of its syntax can be explained on the basis of the hypothesis. The following idiosyncrasies of English structure can be easily understood in the light of the hypothesis: the hierarchy of sentence, clause, noun phrase, adjective, and adverb; the different behavior of subject and object clauses; the phrase structure of the active and the passive with a *by* phrase; the reversal of order of direct and indirect object; the shifting of the position of the separable verb particle; the function of the anticipatory *it;* the first position of the interrogative pronoun; the position of adjectives before the noun and relative clauses after the noun; the discontinuous nature of adjectival and adverbial phrases; the position of certain adverbs before the article; the fact that when the genitive marker follows its noun phrase, it is an affix (*'s*), and when it precedes, it is a separate word (*of*); and that derivational affixes are suffixes, and prepositions, articles, and conjunctions are separate words.

From the point of view advanced here, the grammar of English seems to be very nearly well behaved. It appears that the syntax of English is not an endless catalogue of whimsical complications, although there are some relics of the past. Neither does English appear to be an abstract formal system, on a par with certain elegant mathematical notations. Instead, it is a particularly well-engineered instrument of communication, with many ingenious innovations to adapt it to the capabilities of its users and to circumvent as much as possible the limitations of the human memory.

It remains to be seen how well the hypothesis applies to other languages. Of course other languages are not the same as English: They will not have the same features as those that we could explain in English on the basis of the hypothesis. They may even have just the opposite features to some of those explained in English, postpositions instead of prepositions for example. But it is the over-all structure of the syntax of the language that should be judged in the light of the hypothesis, since it is usually the cooperation of a number of features that keeps any given sentence from becoming too deep. All languages should have the "left-to-right" asymmetry that is predicted. This means that there could not be a language that has the same structure as another in all respects except that its structures go from right to left instead of from left to right. Such a language would be unworkable because depth would not be limited.

If the hypothesis turns out to be essentially true, we would have the following picture of language production. Human speech is produced by a finite-state device and by an essentially left-to-right process. There exists a temporary or working memory that can contain about seven items. This is used to make possible a factoring of the state, which results in great economies in the use of the permanent memory. This factoring of the state is represented in our model by a constituent structure organization of the grammar. The actual process of sentence production corresponds in our model to applying rules by successive expansion from the top down. This process of applying constituent structure rules, if unrestrained, would lead to sentences requiring more than the available temporary memory capacity. On this account, grammars contain various restrictions and devices like postponement transformations to render them effectively well-behaved. Thus the grammar is brought back within the capabilities of the finite-state device. We cannot conclude that the particular organization of the temporary memory in our model—last item in, first item out—necessarily has any basis in the structure of the brain, for this organization is the result of the particular way in which we have represented the grammar in terms of constituent-structure rules. In fact, even our model modifies the "last in—first out" rule in order to handle discontinuous constituents. There is good evidence in the structure of English, however, that the temporary memory can hold not many more than seven items, and there is also good evidence that the speaker is not ordinarily aware of the number of units he has stored in his temporary memory at any one time: The grammar does this counting for him.

The current state of the hypothesis, after a brief and necessarily preliminary examination of English, is as follows. Part (*a*) stated the assumption that languages have a grammar based on constituent structure. We have found nothing in English that would cast doubt on this assumption if the concept of constituent structure includes the possibility of discontinuous constituents. Part (*b*) stated that the sentences that occur in a language do not exceed a certain depth. Evidence has been presented that tends to substantiate this assumption in the case of English. It appears that the maximum depth of naturally occurring English sentences is in the vicinity of seven or not far above seven. A better determination of the exact upper limit will require a more extended study of English syntax. Part (*c*), that the depth is equal or nearly equal to the span of immediate memory also seems to be substantiated for English. It is not known, however, whether the memory involved in sentence production is the same memory that is involved in the psychological tests of the span of immediate memory. This psychological question is yet to be investigated. Parts (*d*), (*e*), and (*f*), which state that all languages have devices to limit depth, devices to circumvent this limitation, and that these devices represent much of the syntactic complexity of language, seem to hold for English.

Part (*g*), which states that depth phenomena play an important role in language change, cannot be examined until we have grammars or appropriate grammatical sketches representing different periods in the history of a language. The results of the synchronic study of English, however, have already revealed a point that may be of significance for language change. We have noticed that stylistic factors frequently seemed to be involved when deep structures were not actually ungrammatical. An example of this was found in certain of the sentences involving a direct and an indirect object. Since stylistic preferences generally seemed to favor sentences of lesser depth, it is probable that much of our feeling of awkwardness in sentences is associated either directly or indirectly with depth. When it is associated directly with depth, the deeper version of a sentence is more awkward. Our feeling of awkwardness is sometimes also connected with a particular syntactic feature like repetition of a word or construction. It is clear how this could come about. It could be connected with our arguments about an ungrammatical first step. There would thus be a structural reason connected to depth for our feeling that repetition is generally to be avoided. This feeling would then apply in all cases of repetition, and would account for the fact that although repeated object clauses involving repetition of the same kind of object clause construction, for example, seem stylistically poor but grammatical, they are much more acceptable if they involve the cascading of different kinds of object clauses. When feelings of stylistic elegance or awkwardness are associated with alternative constructions that are nearly synonymous, we have the essence of a mechanism for language change that might operate in the following way.

Whenever speakers embark on grammatical sentences that exceed the depth limit, they become trapped and have to stop and start over. If this becomes a frequent occurrence, speakers will try to avoid the constructions that got them into trouble. Many of these constructions are useful in other, less troublesome, sentences, so in general they will not be given up. However, in time, alternative structures develop. At first there is a feeling of stylistic inelegance for the deep structures and a preference for the others. But gradually the grammar of the language will change in such a way that the stylistically poor constructions become ungrammatical, and the alternatives become the rule. Depending upon the circumstances under which this happens, we will have an agglutinative force, a force for change in word order, a force for reinterpreting a construction with a new constituent structure, and so on. These forces are very specific in that they operate only under certain particular circumstances in which depth is involved.