XXV. LINGUISTICS*

Prof. R. Jakobson
Prof. A. N. Chomsky
Prof. M. Halle
Prof. L. M. Kampf
Prof. A. L. Lipson
Dr. G. H. Matthews
Dr. Paula Menyuk
A. R. Carlson
J. A. Fodor
C. Fraser
Barbara C. Hall
J. J. Katz
S. J. Keyser
D. T. Langendoen
T. M. Lightner
Krystyna Pomorska
P. M. Postal
J. Reitzes
D. E. Walker

RESEARCH OBJECTIVES

This group sees as its central task the development of a general theory of language. The theory will attempt to integrate all that is known about language and to reveal the lawful interrelations among the structural properties of different languages as well as of the separate aspects of a given language, such as its syntax, morphology, and phonology. The search for linguistic universals and the development of a comprehensive typology of languages are primary research objectives.

Work now in progress deals with specific problems in phonology, morphology, syntax, language learning and language disturbances, linguistic change, semantics, as well as with the logical foundations of the general theory of language. The development of the theory influences the various special studies and, at the same time, is influenced by the results of these studies. Several of the studies are parts of complete linguistic descriptions of particular languages (English, Russian, Siouan) that are now in preparation.

Since many of the problems of language lie in the area in which several disciplines overlap, an adequate and exhaustive treatment of language demands close cooperation of linguistics with other sciences. The inquiry into the structural principles of human language suggests a comparison of these principles with those of other sign systems, which, in turn, leads naturally to the elaboration of a general theory of signs, semiotics. Here linguistics touches upon problems that have been studied by modern logic. Other problems of interest to logicians – and also to mathematicians – are touched upon in the studies devoted to the formal features of a general theory of language. The study of language in its poetic function brings linguistics into contact with the theory and history of literature. The social function of language cannot be properly illuminated without the help of anthropologists and sociologists. The problems that are common to linguistics and the theory of communication, the psychology of language, the acoustics and physiology of speech, and the study of language disturbances are too well known to need further comment here. The exploration of these interdisciplinary problems, a major objective of this group, will be of benefit not only to linguistics; it is certain to provide workers in the other fields with stimulating insight and new methods of attack, as well as to suggest to them new problems for investigation and fruitful reformulations of questions that have been asked for a long time.

R. Jakobson, A. N. Chomsky, M. Halle

A. ON THE LIMITATIONS OF CONTEXT-FREE PHRASE-STRUCTURE DESCRIPTION

1. Introduction

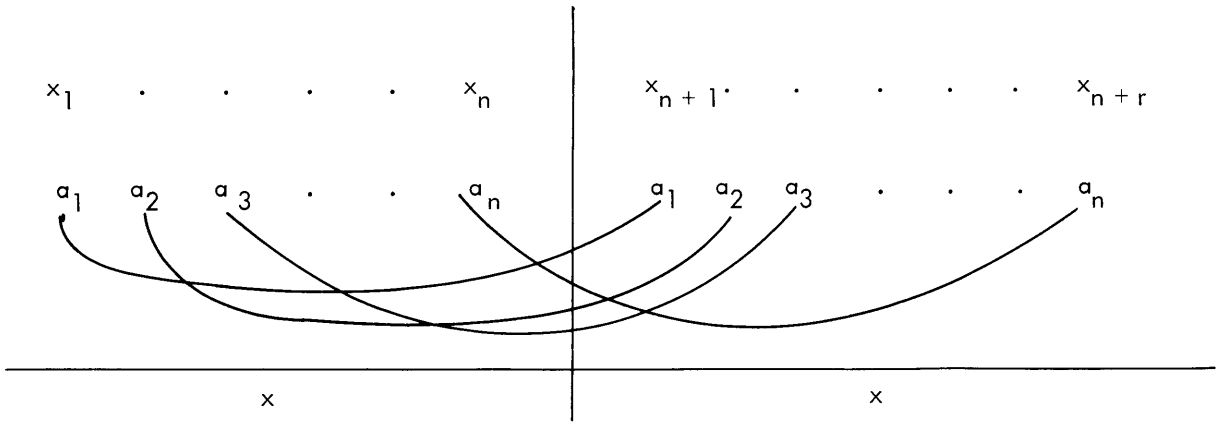
Although the term "phrase-structure grammar" appears widely in linguistic literature, it is, considered from the formal point of view, ambiguous. In fact, multiply so.

*This work was supported in part by the National Science Foundation (Grant G-7364 and Grant G-13903).

Chomsky[1] has considered two types of phrase-structure systems. One of these consists of productions of the form 'a ⟶ B', where a is a single symbol and B is a non-null string distinct from a. Since each of these is to be interpreted as the instruction "rewrite a as B," this type of system consists of rules that are free of context restrictions. That is, the rewriting of a is not determined in any way by the surrounding symbols of the string in which a is embedded. A finite set of rules of this form is called by Chomsky a "context-free phrase-structure grammar," and languages that can be generated by such grammars are called "context-free languages."

Mohawk is one of the five living Northern Iroquoian languages and is spoken by several thousand people in Canada and New York State. The chief purpose of this report is to show that Mohawk is not a context-free language. That is, it is impossible to construct a finite set of context-free rules that will enumerate all, and only, Mohawk sentences. It has been proved by Chomsky[1] that the language consisting of all and only the strings $/a^n b^m a^n b^m ccc/$ is not a context-free language, and a similar, although more complicated, proof yields the same result for the general case, that is, for the language /XX/, where X is a variable over strings of arbitrary length. All of the sentences of such a language have a dependency structure like that shown in Diagram 1, in which x_{n+i} must equal x_i for all i.

DIAGRAM 1



We shall demonstrate that Mohawk lies outside the bounds of context-free description by illustrating that it contains, as a subpart, an infinite set of sentences with the formal properties of the language /XX/.

2. Demonstration

A simple Mohawk sentence may consist of a subject noun, a verb, and an object noun in that order. A noun consists of a noun prefix, a noun stem, and in some cases a noun

suffix. A verb consists minimally of a pronominal prefix, a base, and a suffix. Hence we have sentences like

E1 ka|ksa?a| |ka|nuhwe?|s| |ne-| |ka|nuhs|a?| 'the girl likes the house'
1 2 3 4 5 6 7 8 9

Here,

1 = noun prefix
2 = noun stem 'child'
3 = pronominal prefix
4 = base = verb stem 'to like'
5 = suffix = serial aspect
6 = article
7 = noun prefix
8 = noun stem 'house'
9 = noun suffix

It should be noted that the base constituent of the verb in E1 is simply a verb stem. However the base constituent may also consist of an incorporation marker plus a verb stem. That is, one of the rules of the grammar is

Base → (inc) + Verb Stem

in which the parentheses indicate an optional element. And there is an interesting transformational rule that incorporates the noun stem of the object of a sentence like E1 into the verb by substituting the noun stem for the incorporation marker. In these cases the base constituent then has the structure: Noun Stem + Verb Stem. Thus there are sentences like

E2 kaksa?a kanuhs∕nuhwe?s 'the girl likes the house'

The crucial fact about incorporation from the point of view of this report is that under certain conditions, which need not concern us here, the incorporation rule also leaves the original noun behind so that we find sentences like

E3 kaksa?a kanuhs∕nuhwe?s kikɔ kanuhsa? 'the girl likes this house'

In such sentences there is a strict dependency of exactly the type shown in Diagram 1 between the incorporated noun stem and the noun stem of the object noun. Thus, although there are sentences such as

E4 kaksa?a kanuhwe?s ne- ka?sreht 'the girl likes the car'

E5 kaksa?a ka?srehtanuhwe?s 'the girl likes the car'

E6 kaksa?a ka?srehta∕nuhwe?s kikɔ ka?sreht 'the girl likes this car'

there are no sentences such as

E7 * kaksa?a kanuhs∕nuhwe?s kikɔ ka?sreht

or

E8 * kaksa?a ka?srehta∕nuhwe?s kikɔ kanuhsa?

There is then a subset of sentences in Mohawk which contain constituents whose

structure is representable by Diagram 1. Sentences can contain both incorporated noun stems and external noun objects only if the noun stem of the object is the same as the incorporated noun stem. Hence, in order to show that Mohawk is not a context-free language, it is only necessary to show that the number of noun stems that can occur in such sentences is unlimited. But this is in fact the case.

There is a nominalization transformation that applies to the base constituent of a verb and that adds a nominalizing morpheme (tsra/hsra) to produce a noun stem. Thus from the verb base nuhwe? one can produce the noun stem nuhwe?tsra that occurs in the noun kanuhwe?tsra? 'the liking'. But this nominalization rule applies as well to base constituents that contain incorporated noun stems. Hence from the verb ka?srehtʌnuhwe?s in E5 we can produce the noun stem ?srehtʌnuhwe?tsra that occurs in the noun ka?srehtʌnuhwe?tsra? 'the liking of the car'. But now complex noun stems produced by this nominalization rule are themselves capable of being incorporated in verbs, hence capable of renominalization, of reincorporation, and so on. Thus the process of noun-stem formation is recursive and there is no bound on the length of noun stems, and no limit to the number of noun stems.

We can illustrate this process of noun-stem formation by building up a rather long noun stem. We omit here the subject noun.

1. katsɔries ne- kanuhsa? $\xrightarrow[\text{incorporation}]{}$ kanuhsʌtsɔries
   'she finds the house'
   $\xrightarrow[\text{nominalization}]{}$ kanuhsʌtsɔrihsra? 'the finding of the house'
2. kanuhwe?s ne- kanuhsʌtsɔrihsra? $\xrightarrow[\text{incorporation}]{}$ kanuhsʌtsɔrihsranuhwe?s
   'she likes the finding of the house'
   $\xrightarrow[\text{nominalization}]{}$ kanuhsʌtsɔrihsranuhwe?tsra?
   'the liking of the finding of the house'
3. yao?taksɔ? ne- kanuhsʌtsɔrihsranuhwe?tsra? $\xrightarrow{\text{incorporation}}$
   yaonuhsʌtsɔrihsranuhwe?tsraksɔ?
   'the liking of the finding of the house is evil'
   $\xrightarrow[\text{nominalization}]{}$ kanuhsʌtsɔrihsranuhwe?tsraksɔhsra?
   'the evil of the liking of the finding of the house'
4. kaharatats ne- kanuhsʌtsɔrihsranuhwe?tsraksɔhsra? $\xrightarrow{\text{incorporation}}$
   kanuhsʌtsɔrihsranuhwe?tsraksɔhsrakaratats[a]
   'she praises the evil of the liking of the finding of the house' $\xrightarrow{\text{nominalization}}$
   kanuhsʌtsɔrihsranuhwe?tsraksɔhsrakarattsra?
   'the praising of the evil of the liking of the finding of the house'

This last noun can occur perfectly well in a sentence with an incorporated noun stem identical to its own. Thus

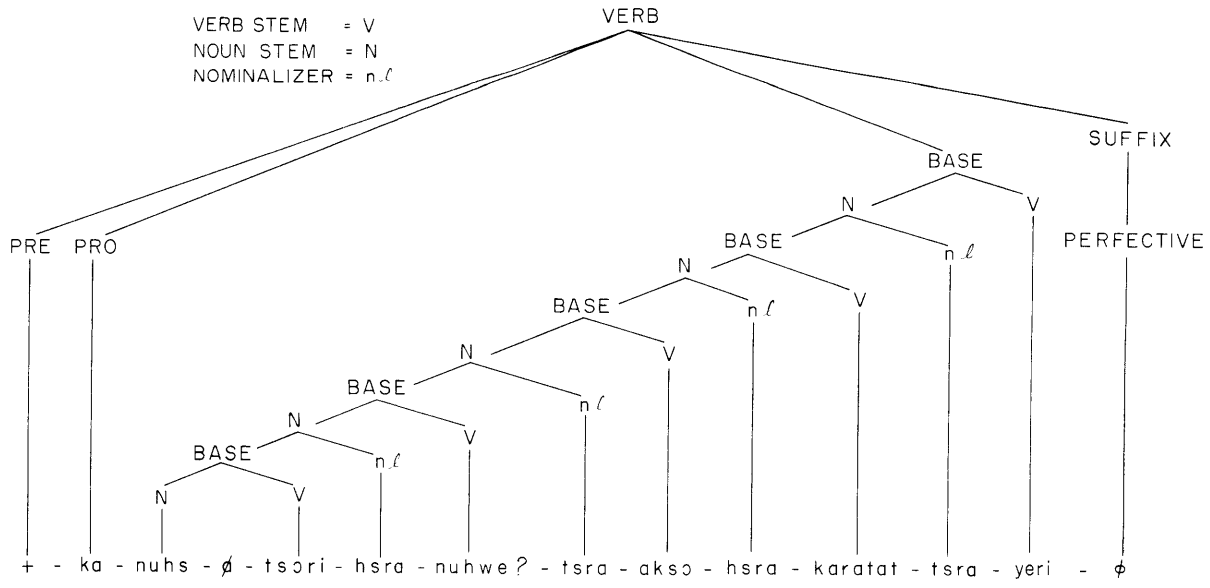
E9 tkanuhsʌtsɔrihsranuhwe?tsraksɔhsrakarattsrayeri kikɔ
kanuhsʌtsɔrihsranuhwe?tsraksɔhsrakarattsra?
'this praising of the evil of the liking of the finding of the house is right'

We therefore see that the subset of sentences in Mohawk which requires an identity of constituents is in fact infinite and that the set of constituents that must be identical (noun stems) is infinite. But, since these are the crucial formal properties of the language /XX/, it follows that

THEOREM: Mohawk is not a context-free language[b]

Incidentally, the process of noun-stem formation which we have considered has another interesting consequence. It has been suggested by Yngve,[3,4] on the basis of the assumption of finiteness of memory plus a certain model of speech production, that a general feature of human language will be the assignment of phrase structure to sentences in such a way as to guarantee that the branching of phrase-structure trees to the left does not exceed some small finite number of successive embeddings, probably seven. But if we now look at Diagram 2, which presents the structure of the verb in E9, we see that the processes of noun-stem formation and base formation are left branching and these processes are recursive. Therefore there is no bound whatever on the branching of trees to the left in Mohawk. Since the assumption of finiteness of memory is unassailable, it follows that the model of speech production which yields the prediction of a general restriction on left branching cannot be correct.

DIAGRAM 2



3. Significance

We can now ask: What is the importance of the fact that Mohawk is not a context-free language? First, as pointed out by Chomsky,[5] context-free grammars appear to be the natural formalization of the type of grammar which would result from the methods of immediate constituent analysis[c] as these have been understood in contemporary American descriptive linguistics.[d] Insofar as this is true, we have a proof that these methods, which depend chiefly on substitutability operations, cannot yield correct grammars for natural languages. Second, we have increased to a significant degree the strength that is provably necessary for any general theory of linguistic structure. That is, by considering only the weakest of all of the possible formal requirements on grammars – that they enumerate the correct set of terminal strings[e] – we have shown some phrase-structure systems to be inadequate for the generation of natural languages. Results of this type will be of quite limited interest, however, until they can be extended to more flexible systems. Context-free grammars are of little linguistic interest, since, on the basis of the contingencies of actual linguistic description, it is known that context restrictions are required in rules.

Chomsky has also considered another type of phrase-structure system, Type 1 grammars. These permit context restrictions and are undoubtedly strong enough to enumerate the sentences of natural languages. However, this is of no real linguistic interest, either, since Chomsky[5] has shown that such grammars cannot ensure the correct assignment of constituent structure as can context-free grammars. It is possible to produce permutations of elements within Type 1 grammars and no clear linguistic meaning can be given to the structures that result from permutations within phrase-structure systems. That is, the relation "is a" that we wish to hold between, say, "man" and "Noun" is not correctly reconstructed by Type 1 grammars. Therefore neither context-free grammars that are provably too weak for natural languages nor Type 1 grammars that provably assign incorrect structure is an adequate formalization for natural language grammars. Parikh, however, has shown that there exists a phrase-structure system intermediate in strength[9] between context-free grammars and Type 1 grammars. Let us call these "Type 1B grammars." These permit the use of contexts but do not permit permutations of elements. Type 1B grammars thus seem to provide an effective and natural formalization of the notions of constituent structure, thereby permitting the needed flexibility of context restrictions but prohibiting permutations that are incompatible with the desired linguistic interpretation of structure.

The really interesting question at this point is to find out whether or not Type 1B grammars are provably too weak for natural languages. Chomsky was able to show that Type 1 grammars can generate the language /XX/, which is beyond context-free description, by using the fact that Type 1 grammars can incorporate permutations. But

Type 1B grammars do not have this feature, and therefore it seems very likely, on the basis of exactly the sort of construction considered in section 2, that natural languages lie outside the range of Type 1B description. If this is so, we shall have moved very far toward demonstrating, in terms of formal considerations of generative power, that which has already been shown in terms of considerations of simplicity and explanatory power, namely, that natural languages require transformational devices that make use of variables and powerful formal operations (deletion, permutation) that are impossible in phrase-structure systems.

P. M. Postal

Footnotes

[a] The stem "to praise" has the shape haratat with no incorporated noun stem, but it has the shape karatat when a noun stem is incorporated.

[b] A similar claim has been made in effect by Bar-Hillel and Shamir[2] for English. But their remarks are based on constructions with "respectively," a rather peripheral aspect of English. The Mohawk construction that we have considered, on the other hand, is not peripheral at all. The rules of incorporation and nominalization are central features of Mohawk syntax.

[c] This claim is supported by the fact that context-free grammars and the other formalization of the notions of constituent structure, namely, the so-called categorical grammars of Bar-Hillel and others,[6,7] have been proved to be equivalent in generative power by Bar-Hillel, Gaifman, and Shamir.[7]

[d] As developed, for example, by Wells.[8]

[e] A more reasonable and more stringent requirement is that the right set of structures be generated. If we insist on this, then phrase-structure grammars in general appear inadequate, because of so-called co-ordinate constructions.

References

1. N. Chomsky, On certain formal properties of grammars, Information and Control 2, 137-167 (1959).

2. Y. Bar-Hillel and E. Shamir, Finite-state languages: Formal representations and adequacy problems, Bull. Research Council of Israel 8f, 155-166 (1960).

3. V. H. Yngve, A Model and an Hypothesis for Language Structure, Technical Report 369, Research Laboratory of Electronics, M.I.T., August 15, 1960.

4. V. H. Yngve, The depth hypothesis, Structure of Language and Its Mathematical Aspects, edited by R. Jakobson (American Mathematical Society, Providence, R. I., 1961), pp. 130-139.

5. N. Chomsky, On the notion 'rule of grammar', Structure of Language and Its Mathematical Aspects, edited by R. Jakobson, op. cit., pp. 6-24.

6. Y. Bar-Hillel, A quasi-arithmetical notation for syntactic description, Language 29, 47-58 (1953).

7. Y. Bar-Hillel, C. Gaifman, and E. Shamir, On categorical and phrase-structure grammars, Bull. Research Council of Israel 9f, 1-16 (1960).

8. R. Wells, Immediate constituents, Language 23, 81-117 (1947).

9. R. J. Parikh, Language-generating devices, Quarterly Progress Report No. 60, Research Laboratory of Electronics, M.I.T., January 15, 1961, pp. 199-212.