[From: Allen Kent (ed.) *Information retrieval and machine translation* (New York: Interscience, 1959)]

CHAPTER 30

Steps Toward Grammatical Recognition*

H. HIŻ

University of Pennsylvania, Philadelphia, Pennsylvania

I. GOALS OF A GRAMMATICAL ANALYSIS

Grammatical analysis of a language may aim at various results. The methods to be used in the analysis may differ considerably, depending on the choice of goals. One may wish to set up a grammar which will produce all the sentences of a given language without producing strings of words which are not sentences.** A grammar that generates exactly the totality of the language does not necessarily answer effectively the question whether a string of words is a sentence or not. And such a grammar is just like a mathematical system that is complete but not decidable. A grammar may also aim at an algorithm for distinguishing sentences from strings that are not sentences.† Generative grammars, both of nondecidable and of algorithmic character may try to exhibit interesting and natural features of the sentence structure. Different kinds of grammatical problems arise when one wants to analyze existing given sentences. Here one may assume that all sentences are given (or are in the process of being given), and grammar has only to say what the structure of each coming sentence is.†† Recognition of sentence structure may in practice be interwoven with the generation of sentences. The one may substantially help the other. But for several aspects of systematic study it is useful to separate recognition problems from generation problems for the time being. This paper deals only with recognition grammar, and assumes that in one way or another there is a decision concerning the sentencehood of each word string. For example one may interrogate native informants.

*This paper was written while the author was engaged in a research project at The University of Pennsylvania, sponsored by The National Science Foundation.

**Attempts of this sort are outlined, e.g., in Chomsky (1) and (2).

†That program is stated, e.g., at the head of Lambek (8).

††This is a more usual linguistic stand. See, e.g., Harris (5).

II. SENTENCE STRUCTURES

It is a fundamental assumption of a grammatical recognition analysis that every sentence has a structure, i.e., a set of relations by which single words and some strings of words in the sentence are related to the entire sentence, and that sentences can be systematically compared as to their structures. Whether these structures are effectively, algorithmically discoverable for each sentence is not definitely known. It may also be that there are many interesting grammatical systems, each of which assigns to a sentence a different structure. To achieve a system of this kind several steps may be taken. This paper does not pretend that the methods suggested below are the only path for recognition analysis. However, other proposals should at least approximate the difficulties which the procedures presented here are tailormade to overcome. Nor should it be assumed that the methods must be applied in the order of their presentation. In practice, one of the methods helps the other and it is not quite correct to speak about levels of linguistic analysis as if there were a clear hierarchy among them ((2) pp. 24–25).

III. GROUPING

A sentence is a string composed of words. One must distinguish between a string and a sequence. A string is a linear array of symbols considered in such a way that the string x of any successive parts of string y, where x includes all the symbols of y, is identical with the string y. A string forms a concatenation algebra.* Thus, the string of the sentences of this paper is identical with the string of words of this paper. The first structure we impose on a sentence is a grouping of the words into some substrings of words The sentences

(3.a) John seems to be healthy
(3.b) John swims to be healthy

as pure strings differ only by one word. But they differ considerably in that to be healthy in the latter stands much more together than to be healthy in the former sentence. This can be seen and tested by several procedures (omitting to be healthy in (3.b) results in a sentence, omitting it in (3.a) leads to a non-sentence; slight difference in intonation; the insertion of so as is allowed in (3.b) but not in (3.a); etc.). To discover what groupings are appropriate for what sentences may be considered one of the main goals of grammars. In the quoted examples, the difference in grouping is that after swims in (3.b) there is a station dividing John swims

*In the sense of Tarski (9), pp. 172–174; or in a sense restricted to allowable concatenations within a language; see Hiż (7), p. 217.

from to be healthy, whereas in (3.a) there is no such station. Seems to be should rather be read together. The matter of grouping is here not determined solely by the kind of verb involved. It is not correct that seems and to in this order, are always grouped together. In

(3.c) This seems to be a confusion

the grouping is similar to that of (3.a). But in

(3.d) This seems to me a confusion

one wants to pick up to me as a separate unit, so that one is inclined to put stations just before it and just after it. When we impose stations in a sentence, we consider the sentence to be not only a string of words but also a sequence of expressions between the stations. We often impose stations within expressions that are already between stations. Thus, in

(3.e) *He is going to his home*

we first impose a station just after going, and then we want a second degree station after to, so that to his home will be grouped together, as will also his home within that substring. In this way we obtain a hierarchy of substrings within a sentence. Grouping by no means exhausts the problems of a grammar of a sentence, but it is a factor which appears in many other grammatical methods, and thus it may be of interest to study it in isolation from other factors.

IV. PARENTHESES

In spoken English, intonation indicates some, though not all of the groupings intended. Note that there is no intonation difference between (3.c) and (3.d). In writing, there are many punctuation devices. Here we shall use only one kind, viz., left and right parenthesis. Left hand parenthesis and right hand parenthesis will be abbreviated by lhp and rhp respectively.

(4.a) α is a parenthetical expression if and only if (1) every proper substring of α that starts where α does has more lhps than rhps, (2) every proper substring of α that ends where α does has more rhps than lhps, (3) the interior between any lhp and rhp in α is always a nonempty expression which does not satisfy both conditions (1) and (2).

(4.b) We say that stations are imposed on a sentence if and only if there is a sequence of mutually exclusive parenthetical expressions that exhausts the entire sentence.

(4.c) A grouping is imposed on a sentence if and only if stations are imposed on the sentence such that every word is the interior of a parenthetical expression and the entire sentence is a parenthetical expression.

The intention here is that every parenthetical expression in a sentence with imposed grouping plays a grammatical role in the sentence, whereas strings that are not parenthetical expressions do not have such roles assigned. Thus (3.c) and (3.d) become with their groupings

(4.d) ((This) ((seems) ((to) (be))) ((a) (confusion)))
(4.e) ((This) (seems) ((to) (me)) ((a) (confusion)))

We shall often omit some of the parentheses if this does not lead to a confusion. The grouping performed e.g., in (4.d) or in (4.e) follows the intuition of native speakers, but since this intuition is not very sharp, there may be alternative groupings that are both possible and persuasive. The indefinite article was here grouped with its following noun. This grouping is more persuasive in the case of the definite article, as in

(4.f) (Harrisburg) (is) (((the) (capital)) ((of) (Pennsylvania)))

In some cases the indefinite article can persuasively be grouped with the verb, as in

(4.g) John (is a) man

This second grouping is taken into account by set theory, when a single $\underline{\varepsilon}$ is put for $\underline{\text{is a}}$.

V. ALGEBRA OF GROUPING

Grouping is an important, though perhaps rudimentary and inaccurate, grammatical analysis. We can compare sentences as to their groupings, and for sentences with the same grouping, we can describe parenthetical similarities between various parts of them. E.g. the sentence (4.f) has the same parenthetical structure as

(5.a) (Today) (George) (((has) (invited)) ((the) (guests)))

And we may say that the grouping of

(5.b) ((George) (Smith)) (likes) (((Scotch) ((or) ((Canadian) (whisky)))) ((on) ((the) (rocks))))

is included in the grouping of (5.a). The parenthetical roles of $\underline{\text{today}}$ in (5.a) and of $\underline{\text{George Smith}}$ in (5.b) is similar. One can build an algebra of groupings and it would already be a grammar. As elements of such an algebra, we may take all sequences composed of three symbols: L, R, and . (L for lhp, R for rhp, . for an expression that is not parenthetical), such that they satisfy conditions (1), (2), and (3) of (4.a). We also assume that L, R and . are all distinct and that L . R is an element of the algebra. Then one may have a k-place operation $\Pi(\alpha_1,\ldots,\alpha_k)$ such that if $\alpha_1,\ldots,\alpha_k$ are elements, then $\Pi(\alpha_1,\ldots,\alpha_k) = L^\frown\alpha_1^\frown\ldots^\frown\alpha_k^\frown R$, where $k \geq 2$. We

do not allow k = 1, for then we would obtain as parenthetical expressions strings of the kind:

(5.c) ((.))

It is exactly to avoid useless parentheses that we have imposed condition (3) in (4.a). We can prove that this algebra forms an upper semi-lattice under the operation Π.

VI. INSUFFICIENCY OF GROUPING

Grouping is not sufficient to establish the grammatical roles of various expressions in the sentence. Two expressions may occupy parenthetically the same places in their respective sentences, which may themselves have the same groupings, and it would still be odd to assign to the two expressions the same grammatical role. As a matter of fact, the connection between groupings and the totality of roles an expression plays in the sentence is many sided. To see the insufficiency of grouping, one should realize that expressions that are parenthetically similar in their respective sentences are not necessarily mutually replaceable.* For example, of Pennsylvania in (4.f) and the guests in (5.a) are mutually non-replaceable. Moreover, these two expressions do not satisfy other tests for similar grammatical roles. Thus of Pennsylvania is omittable in (4.f) the result being a sentence, whereas the guests is not omittable in (5.a). On the other hand grouping is too restrictive for some purposes. One wants to assign a similar role—in some respect—to the guests in (5.a) and

(6.a) The guests arrived

though their parenthetical occurrences do not justify it.

VII. REGROUPING

One may consider not only a single grouping of a sentence but also several admissible groupings of the same sentence. Thus, a class of groupings will be associated with a sentence. This may lead to a deeper analysis than provided by single grouping. One of the differences between (3.e) and

(7.a) He is going with a girl

is that the latter, but not the former, is susceptible of the following grouping

(7.b) (He) ((is) (going with)) (a girl)

---

*About various substitutivity criteria see Chomsky (3), pp. 203-204.

These considerations are supported by transformational analysis. For (7.a) can be transformed into

(7.c) This is the girl he is going with.

But (3.e) is not subject of a similar transformation. Also the variety of admissible groupings seems closely related to considerations of co-occurrence. Going with his home does not occur, whereas going with his girl does occur. To use the terminology of Harris (6), in the construction VPN the N-co-occurrence of going with includes a girl, but presumably not his home.† Generally, if in a construction $\alpha\beta\gamma\delta$ the γ-co-occurrence of β_1 differs substantially from the γ-co-occurrence of β_2 then the construction admits of two groupings, one $\alpha(\beta\gamma\delta)$. the other $\alpha\beta(\gamma\delta)$.

VIII. DISCOVERING GROUPINGS

It may be objected that grouping as described in Sections III-VII is not computable. No algorithm was provided for finding the groupings associated with a sentence. Whether there is such an algorithm is not known. Rather one may hope that there will be at least a partial algorithm leading to proper groupings if one also considers other aspects of grammatical analysis. But before an algorithm is discovered for a problem, it is advisable to state and study the problem independently of its mechanical decidability. Some of the groupings are discoverable by empirical methods from native speakers. We can ask them to put parentheses and encourage some alternative groupings of the same sentence. We may study the intonation patterns. We may ask whether a given place in a sentence admits of an intrusion, etc.* In discovering groupings, categorization of parenthetical expressions, phraze structure analysis, transformations, perhaps even translations may prove to be useful. But to all of them, in turn, grouping is a requirement. Such interdependence of various approaches to the same data is well known in science. In such situations we may try to accept some parts of one analysis and proceed with another, till we can return to the first with the additional apparatus of the second line of thinking; a permanently changing view is a built-in characteristic of scientific method. Still there is a possibility that grouping is not completely discoverable by purely mechanical procedures. This would not discredit grouping, but rather mechanical procedures.

IX. GRAMMATICAL CATEGORIES

Parenthetical expressions may be classified into grammatical categories. It is customary to distinguish nouns (N), verbs (V),

---

†Reference (6), pp. 285-286.

*Several tests for rudimentary grouping are listed in Reference (4), Chapter VII.

adjectives (A), adverbs (D), articles (T), prepositions (P), conjunctions (C), numerals (L) etc. These classifications are done on the dictionary level, and pertain always to single words. In addition, grammar speaks about noun-phrases ($\bar{N}$, like a very beautifully woven and painted orange-colored silk scarf) verb-phrases ($\bar{V}$, like may have not been able to persist), adverbial phrases ($\bar{A}$, like very beautifully woven and painted) etc. The grammatical categories correspond more closely to the classification of phrases than to the dictionary marks. To define grammatical categories for a language, we accept some categories as known ab initio. These primitive categories may still be obtainable by some procedures. For example, the second word in a two-word sentence is a $\bar{V}$. It is convenient, of course, to include among the primitive categories that of a sentence ($\bar{S}$). It is a difficult problem what categories should be listed as primitive for a given language and how to recognize them. Other grammatical categories are then defined as the results of combining in one way or another, those grammatical categories which are already known. Thus one may propose a recursive definition of the notion of a grammatical category. The first part of it specifies primitive categories. The second part of the definition may be of the following form:

(9.a) If in a sentence $S, z_1, \ldots, z_{n-1}, z_{n+1}, \ldots, z_k$ are parenthetical expressions of the categories $\beta_1, \ldots, \beta_{n-1}, \beta_{n+1}, \ldots, \beta_k$ respectively and, in S, $z_1 \frown \ldots \frown z_{n-1} \frown x \frown z_{n+1} \frown \ldots \frown z_k$ is of the category α, and x is a parenthetical expression, then x is of the category $F(\alpha; \beta_1, \ldots, \beta_{n-1}, —, \beta_{n+1}, \ldots, \beta_k)$.

Several points have to be stressed. First of all a category is here understood as a triple relation between the category formed by the functor* and its arguments, the sequence of the left argument categories and the sequence of the right argument categories. Secondly, if categories α and $\beta_1, \ldots, \beta_{n-1}, \beta_{n+1}, \ldots, \beta_k$ are known to belong to the entire construction, and various z_i respectively, then the category of x is computable. We may know the categories either by primitive assignment of some primitive categories to some expressions, or by previous applications of (9.a). Then, the analysis of (9.a) is limited to one single sentence. An expression which is a copy of x in another sentence may be of a different category in that other sentence. Also, the same expression in S may be assigned to two different categories if we start computing from other elements. To illustrate, consider the sentence (4.d) with the following primitive assignments (we write the category by the opening parenthesis):

(9.b) (This (seems (to be)) (a confusion))
$\bar{S}$ $\bar{N}$ $\bar{V}$ $\bar{N}$ $\bar{N}$

*The term functor in this sense was coined by T. Kotarbinski.

Now (9.a) leads to the following conclusions:

Seems to be is an $F(\overline{S};\overline{N},\_,\overline{N})$
seems is an $F(F(\overline{S};\overline{N},\_,\overline{N});\_,\overline{V})$
a is an $F(\overline{N};\_,\overline{N})$.

From the assignments (9.b) we cannot find the categories of to and of be. Consider now the same sentence with a different primitive assignment:

(9.c) (This (seems (to be)) (a confusion))
$\overline{S}$ $\overline{N}$ $\overline{V}$ $\overline{P}$ $\overline{N}T$

Here we may conclude by (9.a) that

seems to be is an $F(\overline{S};\overline{N},\_,\overline{N})$
to be is an $F(F(\overline{S};\overline{N},\_,\overline{N});\overline{V},\_)$
be is an $F(F(F(\overline{S};\overline{N},\_,\overline{N});\overline{V},\_);P,\_)$
confusion is an $F(\overline{N};T,\_)$.

Again with the assignment

(9.d) (This (seems (to be)) (a confusion))
$\overline{S}$ $\overline{N}$ $\overline{V}$ $\overline{T}$

we similarly conclude that

a confusion is an $F(\overline{S};\overline{N},\overline{V},\_)$
confusion is an $F(F(\overline{S};\overline{N},\overline{V},\_);T,\_)$.

X. CLASSES OF CATEGORIES

The approach discussed in section 9 presents some difficulties. Three of them will be shown here (in the present section and in xi and xii). The analysis that shows the relationship between a functor and its arguments, though elegant, does not answer some important problems. In

(10.a) Almost all the deans are incompetent

deans, the deans, all the deans, and almost all the deans are noun-phrases but only the last one forms the noun-phrase which is the subject of this sentence. To say that deans is here $\overline{N}$ is not yet very instructive. To say that deans is an $F(\overline{N};D,Q,T,\_)$ does not exhaust the story. For deans in

(10.b) Just almost all the deans are incompetent

is $F(\overline{N};D,D,Q,T,\_)$ and in

(10.c) All deans are incompetent

it is $F(\overline{N};Q,\_)$. There is however, a common feature in the roles of deans in (10.a), (10.b), and (10.c). This common feature is that it

closes a noun-phrase. One is perhaps led to introduce a class of categories which are closers of noun-phrases. And similarly, a class of openers of noun-phrases. Then we can speak about maximal noun-phrases in a sentence. In an analogous way, one can form a class of categories that are openers of a verb-phrase and a class of categories that are closers of a verb-phrase. This leads to the following extension of (9.a).

(10.d) If in a sentence $S, z_1, \ldots, z_{n-1}, z_{n+1}, \ldots, z_k$ are parenthetical expressions of the categories $\beta_1, \ldots, \beta_{n-1}, \beta_{n+1}, \ldots, \beta_k$ respectively and, in $S, w_1 \frown \ldots \frown w_t \frown z_1 \ldots \frown z_{n-1} \frown x \frown z_{n+1} \frown \ldots \frown z_k \frown u_1 \frown \ldots \frown u_r$ is of the category α, and x is a parenthetical expression, then x is of a category that is a member of the class $F(\alpha, \ldots, \beta_1, \ldots, \beta_{n-1}, —, \beta_{n+1}, \ldots, \beta_k, \ldots)$.

E.g. an article is always $F(\bar{N}; \ldots, \_, \ldots)$, though we may encounter a variety of categories to the left and a variety of categories to the right. Thus, an article is not necessarily a closer or an opener of a noun-phrase, but is certainly inside a noun-phrase. This information is of interest on its own. The more information one has indicating that a given word is an opener, a closer, or inside a particular kind of a phrase, independently of what other expressions may enter the phrase, the more easily will a formal or mechanical analysis of the text be formed.

XI. ENVIRONMENTS

To define noun-phrase closers we will take one more step. We may admit that an expression is assigned to a category (or to a class of categories) not only on the strength of what expression the functor forms and out of what expressions it forms it, but also *on the strength of what occurs in the environment—independently* of whether the arguments show any formal relation to the environment. Thus, e.g. word which is classified as a noun, a numeral, a quantifier, or a pronoun is a noun-phrase closer provided that it does not occur just before a word which is also a noun, a numeral, a quantifier or a pronoun. Thus

(11.a) x is $F(\bar{N}; \ldots, \_)$ if and only if x is N or L or Q or R and x_{+1} is none of these. Once you have a noun-phrase closer you can examine whether or not the preceding word is within the noun-phrase.

(11.b) x is $F(\bar{N}; \ldots, \_, F(\bar{N}; \ldots, \_))$ if and only if x_{+1} is $F(\bar{N}; \ldots, \_)$ and x is N,L,A,G,S,Cl,Q,T,B or H

(G = ...ing; S = ...ed or ...en; Cl = and, or; B = adjectival pronoun, H = that).

In a similar way we can characterize:

(11.c) x is $F(\bar{N}; \ldots, \_, A, F(\bar{N}; \ldots, \_))$ if and only if
x_{+1} is A and

x_{+2} is $F(\overline{N};\ldots,\_)$ and
x is D,L,A,G,S,Cl,Q,T,B or H.

(11.d) x is $F(\overline{N};\ldots,\_,D,A,F(\overline{N};\ldots,\_))$ if and only if
x_{+1} is D and
x_{+2} is A and
x_{+3} is $F(\overline{N};\ldots,\_)$ and
x is L,A,G,S,Cl,Q,T,B, or H.

(11.e) x is $F(\overline{N};\ldots,\_,Cl,A,F(\overline{N};\ldots,\_))$ if and only if
x_{+1} is Cl and
x_{+2} is A and
x_{+3} is $F(\overline{N};\ldots,\_)$ and
x is A,G or S.

(11.f) x is $F(\overline{N};\_,F(\overline{N};\ldots,\_))$ if
x_{+1} is $F(\overline{N};\ldots,\_)$ and
x is A and
x_{-1} is Cl and
x_{-2} is not (A,G,S).

(11.g) x is $F(\overline{N};\ldots,\_,L,F(\overline{N};\ldots,\_))$ if and only if
x_{+1} is L and
x_{+2} is $F(\overline{N};\ldots,\_1$ and
x is L,Cl,Q,T,B or H.

(11.h) x is $F(\overline{N};\_,F(\overline{N};\ldots,\_))$ if
x_{+1} is $F(\overline{N};\ldots,\_)$ and
x is L and
x_{-1} is Cl.

etc.

Such characterizations may form a basis for a program of mechanical grammatical analysis and, as a matter of fact, were used in this role by a linguistic research group at The University of Pensylvania. To use one more illustration, consider the adjectival phrases. We may propose a routine which recognizes adjectival phrases by scanning the sentence from right to left. A tree for recognition of $\overline{A}$ is shown in Fig. 1. The squares contain instructions. Instead of this tree we can write the following six formulas:

(11.i) x is $F(\overline{A};\ldots,\_)$ if and only if x is A and x_{+1} is not A.
(11.j) x is $F(\overline{A};\ldots,\_,F(\overline{A};\ldots,\_))$ if and only if x_{+1} is $F(\overline{A};\ldots,\_)$ and either x is C and x_{-1} is A, or x is D and x_{-1} is C or D, or x is A.
(11.k) x is $F(\overline{A};\ldots,\_,C,F(\overline{A};\ldots,\_))$ if and only if x is A, x_{+1} is C, and x_{+2} is $F(\overline{A};\ldots,\_)$.
(11.l) x is $F(\overline{A};\ldots,\_,D,F(\overline{A};\ldots,))$ if and only if x_{+1} is D, x_{+2} is $F(\overline{A};\ldots,\_)$ and either x is C and x_{-1} is D, or x is D.

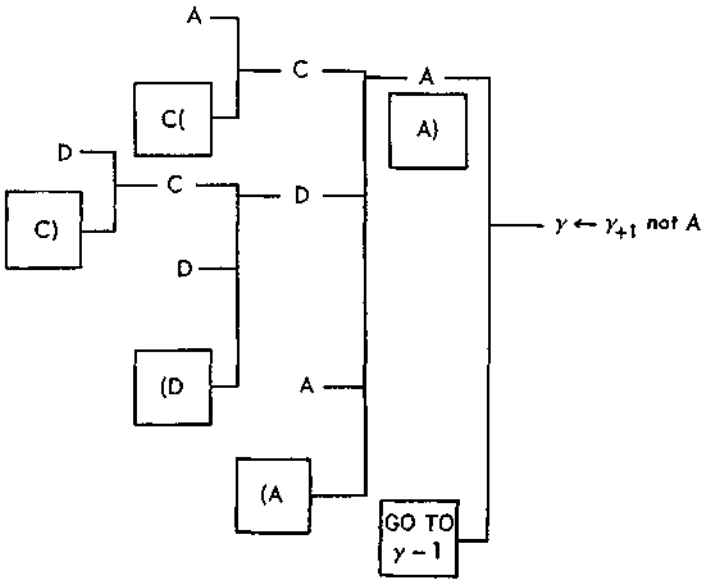

Fig. 1. Steps Toward Grammatical Recognition.

(11.m) x is $F(\bar{A};\ldots,\_,C,D,F(\bar{A};\ldots,\_))$ if and only if x is D, x_{+1} is C, x_{+2} is D, and x_{+3} is $F(\bar{A};\ldots,\_)$.
(11.n) x is $F(\bar{A};\_,\ldots)$ if and only if either x is A, x_{-1} is C and x_{-2} is not A, or x is D, x_{+1} is A, x_{-1} is C, and x_{-2} is not D, or x is D, x_{+1} is A, x_{+2} is neither C nor D, or x is A and x_{-1} is neither C nor D nor A.

XII. DOUBLE CLASSIFICATION

A substantial complication occurs with subordinate clauses.

(12.a) To a man who was poor John gave money.
Who in this case plays a double role; as a clause opener and as a noun. One is tempted to give to it a classification which will allow a man to be "cancelled" twice: once with the first $\bar{N}$ of $F(\bar{S};P,\bar{N},\bar{N},\_,\bar{N})$ for gave, the second time with $\bar{N}$ in $F(\bar{S};\bar{N}\_,\bar{A})$ for was. This solution would seen attractive when $\bar{A}$ must agree in gender and in number with the first $\bar{N}$. A clause opener like who, which, what, where, when, why, whose, etc. all occur in front of not complete sentences and call for a double classification.*

*Harris (6) (p. 303) divides them into two morphemes; the wh-element and -o, -ich, -at (these are $\bar{N}$), -ose ($\bar{A}$), -y ($P\bar{N}$), etc.

XIII. OTHER METHODS

To close the discussion let us stress that the presented methods do not exhaust the fruitful methods of grammatical analysis. Transformational analysis, e.g., was not mentioned here. And let us recall that the problem of primitive categories remains open and demands completely new methods; to establish some categories ab initio we cannot treat the categories as relations in the same way as it was done above.

REFERENCES

1. Chomsky Noam, Three Models for the Description of Language, "IRE Transactions on Information Theory," Vol. IT-2, No. 3, pp. 113-124.
2. Chomsky, Noam, "Syntactic Structures," 'S-Gravenhage, 1957.
3. Chomsky, Noam, "Review of Joseph H. Greenberg's Essays in Linguistics," Chicago, 1957; "Word," 15 (1959), pp. 202-218.
4. Chomsky, Noam, "The Logical Structure of Linguistic Theory," Mimeographed, 1955.
5. Harris, Zellig S., "Methods in Structural Linguistics," Chicago, 1951.
6. Harris, Zellig S., Co-occurrence and Transformation in Linguistic Structure, "Language," 33 (1957), pp. 283-340.
7. Hiż, H., Types and Environments, "Philosophy of Science" 24 (1957), pp. 215-220.
8. Lambek, Joachim, The Mathematics of Sentence Structure, "American Mathematical Monthly," 65 (1958), pp. 154-170.
9. Tarski, Alfred, "Logic, Semantics, Metamathematics," Oxford, 1956.