XIV. MECHANICAL TRANSLATION*

Prof. V. H. Yngve
Prof. A. N. Chomsky
Dr. J. R. Applegate
Dr. B. Shefts
Eva Maria Ritter

A. THE RANGE OF ADEQUACY OF VARIOUS TYPES OF GRAMMARS

By a language we mean a set (finite or infinite) of strings (i.e., sequences of symbols), each finite in length and constructed from a finite set of symbols. By a grammar of the language L we mean a device of some sort for generating all and only the strings belonging to L. The principal concern of linguistic theory is to develop a precise and general characterization of linguistic structure and to specify the form of grammars in such a way that the underlying structure of each generated string can be recovered from grammars of the proper type and simple and "revealing" grammars can be provided for natural languages. No matter how we develop linguistic theory, we shall surely require that any grammar associated with the theory be finite in length, hence that the set of these grammars be countable. Therefore, any proposed linguistic theory will fail to apply to uncountably many languages. In the last analysis, the adequacy of any linguistic theory is determined by the empirical consequences of applying it to actual natural languages, but we can gain a certain insight into the relative adequacy of various conceptions of linguistic structure by studying the languages to which these conceptions (and the grammars associated with them) literally cannot apply.

The most direct way of describing a language – in a sense it corresponds to the minimum linguistic theory – is by means of a finite-state grammar of the following type. We have a finite set of states labelled $S_0, \ldots, S_n$. Transition from state S_i to state S_j is marked by production of one of the symbols a_{ijk} $(1 \leq k \leq q_{ij})$; but for certain pairs (S_i, S_j) there may be no transition from S_i to S_j. A string Z belongs to the described language if and only if it is produced by a sequence of states $S_{b_1}, \ldots, S_{b_m}$, where $b_1 = b_m = 0$, and where for $1 \leq i < m$, there is transition between S_{b_i} and $S_{b_{i+1}}$. If a language L is described by some such grammar, we say that L is a finite-state language.

Customarily, syntactic description of natural languages is presented in terms of notions of phrase structure and immediate constituent analysis. When these notions are constructed carefully (at least under one interpretation), it can be shown that the system of phrase structure underlying a given language can be reconstructed from grammars of the following form and that every language describable in terms of phrase structure can be described by these grammars. We have a finite set Σ of initial strings $\Sigma_1, \ldots, \Sigma_n$ (one will suffice, in fact) and a finite set F of instruction formulas $X_1 \rightarrow Y_1$, $\ldots, X_m \rightarrow Y_m$, where X_i and Y_i are strings, Y_i being formed from X_i by replacement of a single symbol of X_i by some string of symbols. We say that a string Z' is directly

*This work was supported in part by the National Science Foundation.

derivable from a string Z in terms of F if Z = aXb, Z' = aYb, and X → Y is an instruction formula of F. Given Σ and F, a derivation of the string Z is a sequence of strings with some $\Sigma_i \epsilon \Sigma$ as its initial term, Z as its final term, and the i^{th} term directly derivable from the $i\text{-}1^{st}$ term for each i > 1. In this case, Z is said to be derivable from $[\Sigma, F]$. A set of strings that is derivable from some system $[\Sigma, F]$ we call a derivable language.

Given a system $[\Sigma, F]$, there may be certain derivable strings from which no further strings are derivable. Any string of this kind we call a terminal string with respect to $[\Sigma, F]$. Any set of strings that is a terminal set for some system $[\Sigma, F]$, we call a terminal language.

In the interesting cases, there will be a set S of symbols with these properties: (a) every terminal string is made up of symbols of S; (b) if α is in S then there are no strings a, b, X, Y, β such that X = aαb, Y = aβb, and X → Y is an instruction formula of F. In this case, the terminal strings with respect to $[\Sigma, F]$ can be taken as the language described by the grammar, and the derivable nonterminal strings give information about the phrase structure of the language. Given a derivation of a terminal string Z, we can reconstruct the phrase structure of Z. (Actually, there are certain other conditions that $[\Sigma, F]$ must meet if it is to determine a system of phrase structure, but we shall not elaborate here.)

The sets of finite-state languages, derivable languages, and terminal languages are related in the following way:

Theorem 1. (i) Every derivable language is a terminal language, but not conversely.

(ii) Every finite-state language is a terminal language, but not conversely.

(iii) There are derivable, nonfinite-state languages, and finite-state, nonderivable languages.

To avoid certain inconsequential considerations, we take a language L to be finite-state, derivable, or terminal, if and only if the language L_1 is finite-state, derivable, or terminal, respectively, where L_1 is formed from L by replacing each string X of L by JXJ, where J is a symbol that is never developed in an instruction formula of F (i.e., $J \epsilon S$, as S is defined above).

It can be shown that grammars of the form $[\Sigma, F]$ can be greatly simplified if we order the instruction formulas and define certain of them as obligatory, and if we then require that in constructing a derivation we must proceed through the sequence of instruction formulas, applying each obligatory one and perhaps certain nonobligatory ones, and beginning again at the beginning of the sequence after reaching the end. This condition imposes no restriction on derivability. A natural minimum requirement for grammars would seem to be that the grammar must generate a certain fairly large set of utterances in a reasonable amount of time so that we can actually investigate the

generative adequacy of the grammar. Suppose we define a proper linear grammar as a system $[\Sigma, Q]$, where Q is a sequence of instruction formulas $X_i \rightarrow Y_i$ $(i \leqslant m)$, and for each X_i $(i \leqslant m)$ there is some j such that $X_i \rightarrow Y_j$ is an obligatory instruction formula. This guarantees that every time we run through the grammar each nonterminated derivation must be advanced at least one step; that is, we cannot indefinitely run through the grammar vacuously. To increase the generative power of these grammars, we permit each rule to be applied an indefinite number of times to a given string whenever the rule in question is selected for application.

We say that a set of derivations is properly producible if it is produced by some proper linear grammar. A set of derivations is said to be producible if it is produced by some system $[\Sigma, F]$, where F is a set of instruction formulas, as before. The two types of producibility are incomparable; that is, we have

Theorem 2. There are producible sets of derivations that are not properly producible, and properly producible sets that are not producible.

The import of Theorem 1 is that description in terms of phrase structure is essentially more powerful than description without higher levels. It may be, in fact, that every natural language can be regarded as a finite state language – hence a terminal language. However, when we actually attempt to construct grammars of the specified kinds for natural languages, we find that this description, though perhaps possible, is so complex that it is practically useless. Investigating the situation more closely, we find that some of the complexity is due to the presence of a large but finite number of conditions on utterances (e.g., parallelism of constructions) which, if extended to an infinite set, would take these languages literally out of the range of grammatical description of some specified kind. Investigation of the actual descriptive limits of certain types of grammars can thus offer hints about the usefulness of these grammars even within their limits. Thus the extra generative power of systems of the form $[\Sigma, F]$ over finite-state grammars is reflected in the considerable gain in simplicity of description achieved by moving to such higher levels as phrase structure. Similarly, we note that derivable languages can also be constructed in terms of a rather limited type of finite-state process. Given a system $[\Sigma, F]$, we consider a process with the states $S_o, \ldots, S_q$, with transition from S_i to S_j marked by production of the string A_{ij}. Let $A_{oi} = \Sigma_i$ for $i \leqslant n$ (where $\Sigma = \{\Sigma_1, \ldots, \Sigma_n\}$). Suppose that F contains the instruction formulas $X_i \rightarrow Y_i$ $(1 \leqslant i \leqslant m)$. Then the state of the process when the string A has just been produced is determined by that subset of $\{X_1, \ldots, X_m\}$ whose members are substrings of A, and we move to the next state by replacing X_i by Y_i in A for some X_i in this subset. With the production of a terminal string, we return to S_o. The set of strings A_{ij} derived in this way is a derivable language, and every derivable language is given by some process of this kind. The extra power of proper linear grammars (as stated in Theorem 2) results from the fact that they transcend this limitation and take into account the

"history of derivation" of a string (not merely the shape of the string itself) in carrying through a derivation. This limitation proves to have extremely undesirable consequences in the actual construction of grammars. However, as matters now stand, the extra power of proper linear grammars is actually a disadvantage, since a nonproducible set of derivations does not correspond to a system of phrase structure. This suggests that essentially new conceptions of linguistic structure are necessary, along with more extensive methods for generating sentences from given sentences and for taking into account the history of derivation (constituent structure) of the given sentences.

The fact that not all producible sets of derivations are properly producible opens up new possibilities for testing the adequacy of description in terms of phrase structure, since the requirement of proper linearity seems a reasonable one for grammars. To demonstrate this requires too lengthy a presentation; but it seems to be true that the simplest set of derivations for English sentences is not properly producible, although the simplest set of derivations for a certain independently interesting subset of sentences is properly producible. This fact adds further weight to the suggestion that a new, higher level of linguistic structure is required so that those sentences that are not derived by means of the proper linear grammar can be constructed from those that are so derived by means of transformations that take into account the constituent structure of the terminal strings.

N. Chomsky