XIII. MECHANICAL TRANSLATION*

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

A. ON THE LIMITS OF FINITE-STATE DESCRIPTION

In the Quarterly Progress Report of April 15, 1956, two types of grammars were described formally: finite-state grammars with no independent memory that produce sentences word by word, and $[\Sigma, F]$ grammars which can be represented as slightly less elementary finite-state processes and which impose phrase structure on the generated sentences rather than produce them from "left-to-right." A theorem was stated to the effect that every language describable in terms of a finite-state grammar (every finite-state language) is describable in terms of a system of phrase structure (is a terminal language) but not conversely. The natural question to raise is whether or not there are existent languages that fall outside the range of finite-state description, but within the range of phrase-structure grammars. Further investigation has shown that certain syntactic properties of English exclude it from the set of finite-state languages, but not from the set of terminal languages.

Suppose that A represents the alphabet of language L, and that $S = a_1 {}^\frown a_2 {}^\frown \ldots {}^\frown a_n$ ($a_i \in A$) is a sentence of L.

Definition 1. S has an (i, j)-dependency with respect to L if and only if

(i) $1 \leq i < j \leq n$

(ii) there are $b_i, b_j \in A$ so chosen that S_1 is not a sentence of L and S_2 is a sentence of L, where S_1 is formed by replacing the i^{th} symbol (a_i) of S by b_i, and S_2 is formed by replacing the j^{th} symbol (a_j) of S_1 by b_j.

Definition 2. $D = \{(\alpha_1, \beta_1), \ldots, (\alpha_m, \beta_m)\}$ is a dependency set for S in L if and only if

(i) for $1 \leq i \leq m$, S has an (α_i, β_i)-dependency with respect to L

(ii) for each i, j, $\alpha_i < \beta_j$

(iii) for $i \neq j$, $\alpha_i \neq \alpha_j$ and $\beta_i \neq \beta_j$.

If S contains an m-termed dependency set, then at least 2^m states are necessary in the finite-state grammar that generates the language L that contains S. Hence, a necessary condition on finite-state languages is that there must be a finite upper limit to the size of their dependency sets. With this condition in mind, we can easily construct many nonfinite-state languages. For example, let L_1 be the language containing the "sentences" aa, bb, abba, baab, aabbaa ..., and, in general, all "mirror image" sentences consisting of a string X of a's and b's followed by X read from back to front, and only these. Then, for any m, we can find a dependency set $D_m = \{(1, 2m), (2, 2m-1), \ldots, (m, m+1)\}$, so that L_1 is not a finite-state language.

*This work was supported in part by the National Science Foundation; and in part by a grant from the Eastman Kodak Company.

Turning now to the English language, we find that there are infinite sets of sentences with just the mirror-image properties of L_1. For example, let $S_1, S_2, S_3, \ldots,$ be declarative sentences. Then the following are all English sentences:

(1) (i) If S_1, then S_2.

(ii) Either S_3, or S_4.

(iii) The man who said that S_5, is arriving today.

These sentences have dependencies between "if" and "then," "either" and "or," "man" and "is." But we can choose S_1, S_3, and S_5 in (1) as (1i), (1ii), or (1iii) themselves. Proceeding to construct sentences in this way, we arrive at sentences with dependency sets of more than any fixed number of terms, just as in the case of L_1. English is therefore not a finite-state language.

Note that L_1 is a terminal language. It has the $[\Sigma, F]$ grammar with $\Sigma = \{Z\}$ and $F = \{Z \rightarrow aZa, Z \rightarrow bZb, Z \rightarrow aa, Z \rightarrow bb\}$. Hence, the argument that we have just given does not show that English is not a terminal language, since the sentences we have discussed could be given a $[\Sigma, F]$ grammar in the same way as L_1. The question of the literal possibility or impossibility of a phrase-structure description of English therefore remains open, even though there is considerable evidence that more powerful methods are required if English is to be described effectively.

A. N. Chomsky