XV. MECHANICAL TRANSLATION*

Dr. V. H. Yngve

RESEARCH OBJECTIVES

The basic objective of our program is to devise techniques by which languages can be translated by machine. There are two aspects of the program: the linguistic and the mechanical. The two must progress side by side because there is considerable influence on the methods of each by the results of the other. In addition, a balance must be maintained between short-range programs which aim at the immediately practicable, and longer range research which may add to the basic understanding needed for future progress.

A. SYNTACTIC CATEGORIES

One of the major problems in linguistics is to devise objective methods for analyzing and describing language. Language can be considered as an ordered sequence of symbols. Not all possible sequences of symbols are allowed in a language; most are forbidden by the constraints imposed by rules of usage. A description of a language is in essence a list of the symbols used in the language and a statement of the rules for combining or ordering the symbols. If the symbols are taken as words, the list is a dictionary and the statement of the rules is a grammar. The list of words in a language can be determined by fairly objective techniques, but it is difficult to find objective methods for discovering the rules of grammar and syntax.

Our most recent effort at devising such objective methods consists of comparing a sample of the language with certain features of a simple statistical model of language in such a way that the constraints of grammar and syntax show up as deviations from randomness.

The statistical model adopted assumes that the probability of finding a given word at a given position in the text depends only on the normalized frequency of this word, not on its position in the text. Experimentally, of course, the probability of a given word appearing at a given point in the text is strongly influenced by the occurrence of other words in its environment. Thus, deviations from the model are expected.

In this experiment, the influence of only the six most frequent English words upon each other was measured, since the sample text was short (9490 words). Figure XV-1 shows 8 of the 21 histograms obtained experimentally. Each histogram represents the number of times a given word appeared at various positions before and after another given word located at the center position. If the word at the center position has no influence on the probability of occurrence of the given word, one would expect the average number of occurrences to be equal to the center horizontal line, which is the product of the frequencies of the two words divided by the total number of words in the text. On a statistical basis, one would expect about one-sixth of the histogram bars

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

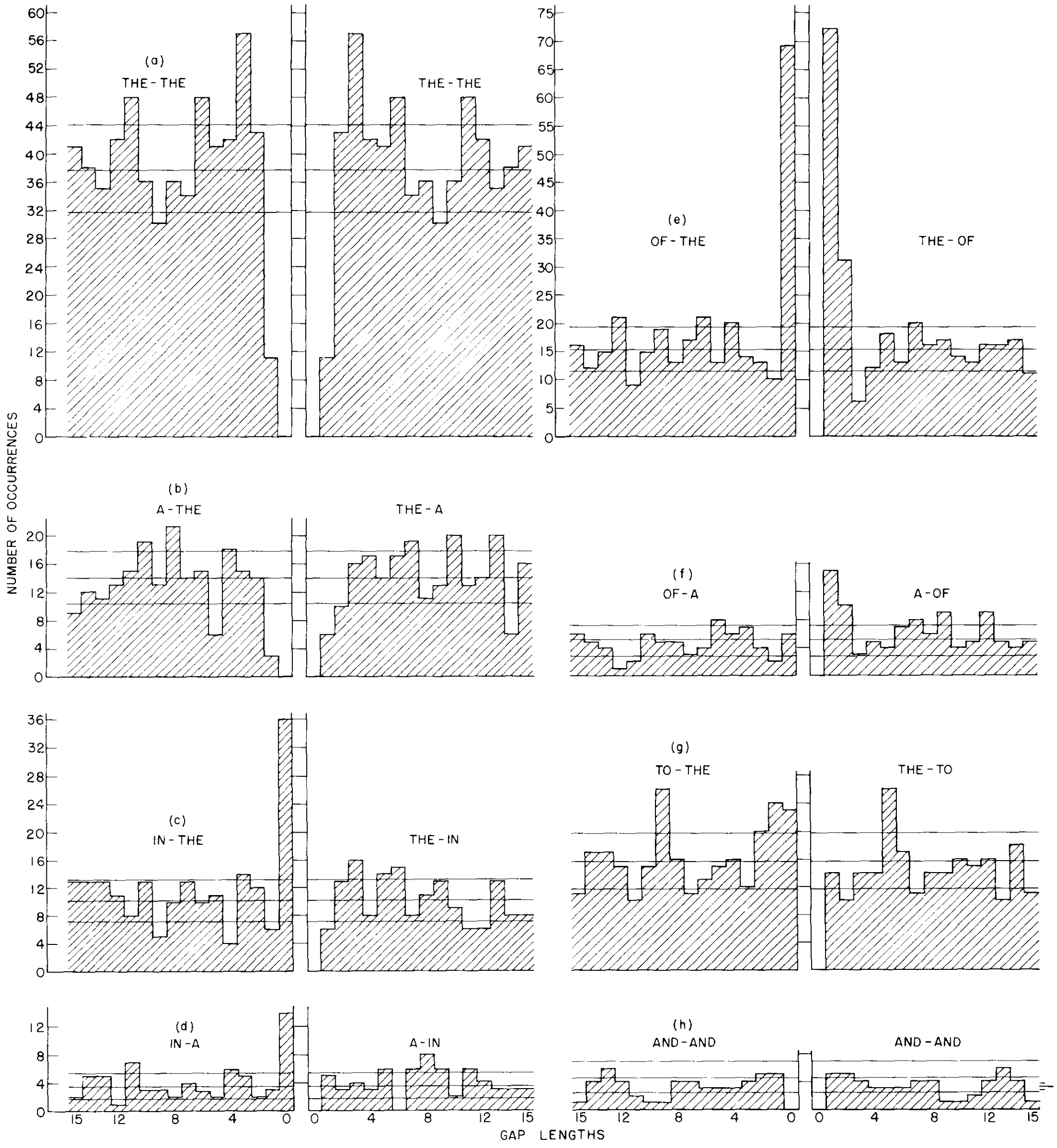

Fig. XV-1

The frequency of occurrence of certain words in the neighborhood of another word. For example, in (b), the number of occurrences of the word "a" is plotted as a function of the number of intervening words to the left or right of "the," which is at the center position.

to be above the upper horizontal line and one-sixth to be below the lower horizontal line.

Every statistically significant observed deviation from randomness has been correlated with known facts about English grammar and syntax. For example, (a) and (h) illustrate the fact that the same word tends not to be repeated at close intervals. These two histograms are, of course, symmetrical because the two words considered are the same; (b) also shows a symmetry which is the result of the grammatical similarity of "a" and "the"; (c), (e), and (g) show the high peak immediately to the left of the center because these words can all be used as prepositions. The wider peak on (g) is caused by the competition of infinitive constructions with the prepositional construction. The large peak at the right of the center of (e) represents the genitive construction.

The width of the observed peaks and valleys gives an idea of the length of the sequences of words in the various constructions. For example, in (a) it is clear that constructions with articles involve two or three words and that such constructions might tend to repeat every four or five words, while constructions with "of" are rather more compact. In this connection it is interesting to note in (h) that the observed probability of "and" in the vicinity of another "and" is lower than would be expected, even out to a separation of 15 words. This is because one of the uses of "and" is to coordinate long clauses.

The method appears to have advantages in that it gives a different viewpoint toward the grammar of well-known languages and thus may help our understanding of language. It may also be useful as another technique at the disposal of the linguist who is trying to describe a little-known language.