XVI. MECHANICAL TRANSLATION

Dr. V. H. Yngve

A concise description of the structure of a language is something that is being sought more and more by linguists. If such descriptions for various languages were available in a fairly complete form the problem of devising mechanical translation routines would be a great deal simpler.

The statistics of language has been investigated by many people. In particular, much work has been done in counting the frequency of different words as they appear in carefully selected samples of language.

An investigation has been made of the possibility that statistical information on the occurrence of words might be useful in finding a description of language. If a large sample of language is obtained, and certain words are selected throughout the text, one can investigate the distribution in the lengths of the gaps that appear between selected words. A gap is measured between two selected words by counting the number of words that occurs between the two words. The two selected words may be called the initial word and the final word, though neither is counted as part of the gap. The smallest possible gap is zero when the two selected words are adjacent.

If the initial and final words are selected at random throughout the text, the distribution of gap lengths follows an exponential

$$p(n) = f \exp(-kn) \text{ with } k = -\ln(1-f)$$

where p(n) is the probability of a gap of length n, and f is the frequency of the selected words in the text, each selected word being used once as an initial word and once as a final word.

If the initial word is selected at random, but the final word is the first occurrence of a particular given word, the distribution of gap lengths will again be an exponential.

If, on the other hand, both the initial and the final word are not selected at random, but each is a particular given word, the observed distribution of gap lengths will depart from the exponential. Generally, if the initial word and the final word are alike, gaps of zero will be less frequent than expected because of the fact that English rarely doubles a word. In this case we may say that like words tend to repel. Certain words selected as initials and certain other words selected as finals show an attraction. For example, the words "of" and "the" have more gaps of zero length than would be expected from random behavior.

The structure of language can be considered as a deviation from randomness; thus it is that the comparison of gap distributions actually observed with the distributions expected on a random basis can lead to information, obtained entirely by the use of statistical techniques, on the structure of language. A more complete discussion of the possibilities and difficulties of the method is being prepared for journal publication.