THE MATHEMATICAL THEORY OF COMMUNICATION AND THE NATURAL LANGUAGE SYSTEM

D. VUYSJE

Language is an instrument of communication, a system of signs which is the bearer of messages being transmitted from a sender to a receiver. The transmission of a "word" may be regarded as a special case of a more general phenomenon, the transmission of signs.

From this point of view, the differentiation and the parcellability of the language-sign are pre-linguistic characteristics, which are inherent to every category of informational signs. De Saussure has already pointed out this phenomenon.

On this semiological level of communication the sender and the receiver function as an engine: the linguistic structures are to be regarded as coming up to the requirements of the transmission and are subjected to the statistical laws that determine the information of every sign system.

In a list in which the words (considered as units of form and not of signification) are ranged according to their frequency of occurrence, the order of precedence is inversely proportionate to the frequency of occurrence ($fr = c$; f = frequency, r = order, c = constant). Zipf, who has investigated this phenomenon, has denoted the relation referred to as the principle of the least effort.

However, this semiological conception disregards

(a) the elementary relation phonetic form-frequency;

(b) the assignment of signification.

For the word is not only a sign, but also the bearer of a concept.

The transmission, and also the coding and the decoding of a sign require time and energy according to the nature and the form of the sign. The "cost" is proportional to the number of phonemes out of which the word is composed, to the length of the word.

The quantity of information of a sign is proportionate to the negative logarithm of its probability, in other words, the quantity of information increases as the probability decreases.

A pack of cards may furnish an example of this thesis. The king and the five contain an equal quantity of information (the probability of a king and of a five is equal), although according to the rules of the game (the position of the party, the place of the card) the two informations have different signification of value, like a glass of wine or a glass of water, which may contain equal contents, but whose contents have different value.

In the case of linguistic signs, independent of their meaning, the quantity of information is expressed by the logarithm of their probability calculated on the basis of the frequency of the word in the language under consideration. The quantity of information is greater as the occurrence of the word is less frequent.

The relation cost-quantity of information can be expressed as follows: the number of phonemes is inversely proportionate to the probability; in other words, the quantity of information is proportionate to the "cost".

Just as the number of phonemes, of which a word is composed, can be taken a sa measure for the "cost" of the expression, the quantity of information – calculated on the basis of its relative frequency of occurrence – enables us to express the informational output of a word. The greater the effort the sending out and the reception of a sign requires from the sender and the receiver, the higher the cost, the more information to be transmitted it will contain.

It is obvious that the linguist will be particularly interested in

(a) the study of the morpho-semiological structure of the sign-system;

(b) the study of the morpho-semantical structure of the same.

With regard to the morpho-semiological aspects of the natural language system, we shall have to study the structure and the output of the system. These two characteristics of the system can be measured, and one can examine to what extent the morpho-semiological structure and the morpho-semiological output are sacrificed to the other requirements of the communication $\frac{\log r}{k} = c$. (r = order; k = number of phonemes).

In a list in which the words of a language are classified on the basis of their increasing length, the number of phonemes (or syllables) of a word is proportionate to the logarithm of its order of succession. It is obvious that the most economical system of a language is the system which will utilize all possible signs which are composed of all combinations of the fundamental elements, therefore of the 40 or 50 phonemes of which the language disposes.

But language is not a theoretical construction. It is linked to physical and psychological conditions, which are detrimental to the economy of the forms: some of them are difficult to pronounce, some others have a small acoustical output, some others again are unpolished and not aesthetical.

Seen from the point of view of the information theoretist, each sign of a system can be used most efficiently with such a frequency, that its information content (or the negative logarithm of its probability) is proportionate to its cost (the number of elementary signs of which it is composed).

The structures of the system of signs can now be determined; it results from the nature of the discontinuous subject-matter and of the laws of the numbers. The fact that the length of the signs increases with the logarithm of their order is a real datum that does not depend on the choice of the sign-user. The sender, however, who disposes of these signs may use them more or less economically.

The linguist cannot neglect the relation of information and signification. He can

scarcely realize the fact that the phonetic form of the words determines their frequency directly. For if the words "man" or "dog" are used more often than the words "cousin" or "greyhound", the frequency of these signs is not determined by their form, but by their content. There is an interrelation between the sign and the concept, meaning that the shortest signs are linked to the most frequent concepts.

Amsterdam