STATISTICS OF PHONEMIC SYSTEMS

GUSTAV HERDAN

In matters of greater importance and of considerable difficulty, it is a good plan to go back to the classical works in the field. Reading, with this in mind, once more, Trubetzkoy's *Grundzüge der Phonologie* (Prague, 1939) I found that the way in which he contrasted his own views about the relation between phonetics and phonemics with that of Professor E. Zwirner, provided the right point from which to develop the subject on which I have been asked to report.

After defining language as a system of norms, whose realization in speech are the actual sounds of the language or, as we would say, the phoneme occurrences, Zwirner proceeds by saying (translation):

> Since such norms (phonemes) cannot be realized by the speech organs in exactly the same way twice, the transition from phonology as the science of the norms to phonetics as their realization in speech must be of a statistical nature, such that the variations of a sound are distributed around their average according to the Gaussian law (normal curve), and these averages are what correspond to the norms.[1]

With this interpretation, Trubetzkoy does not agree. He objects to it on the grounds that the phonometric average of a particular sound quality belongs again to phonetics, and for this very reason does not belong to phonology. As an instance, he adduces the different phonetic averages which can be obtained from what phonometrically is only one phoneme /k/, according to whether in German it stands before a consonant or a vowel, and then again before a stressed or unstressed vowel. According to Trubetzkoy, the phonemic opposition is beyond "Mass und Zahl".

This was more than 20 years ago, and thus at a time when the application of statistics to linguistics was hardly known – though there had been sporadic attempts at applying it to problems of style. In the light of what has since emerged about statistical linguistics as the quantification of de Saussure's conception of language as a part of semiology, and, in particular, of his fundamental dichotomy of language into "langue" and "parole", as corresponding to the fundamental distinction between statistical population and sample,[2] we must arrive at the conclusion that there is good support for Zwirner's conception, and that Trubetzkoy himself, had he seen the matter in the light of modern statistics, would most likely have found that the

[1] Quoted from Trubetzkoy's *Grundzüge der Phonologie* after E. Zwirner and K. Zwirner, *Grundfragen der Phonometrie* (Berlin, 1936).

[2] See G. Herdan, *Language as Choice and Chance* (Groningen, P. Noordhoff, 1956).

statistical hypothesis was in agreement with his own conception of the relation between phonetics and phonology.

It seems clear that what made Trubetzkoy reject Zwirner's hypothesis was that the latter identified the phonometric *average* with the corresponding phoneme. Since an average, or arithmetic mean, is something "after the event", whereas the phoneme is the norm for that event, and must therefore precede it – though not in time! –, Trubetzkoy could not reconcile himself with Zwirner's idea. Moreover, as he showed, a phoneme like /k/ in German corresponds not only to one, but to a number of phonetic averages.

Seen in the light of the theory of mathematical statistics, that difficulty does not exist. We regard, in agreement with statistical theory, the phoneme, whose realisation is an actually observed phonetic event, as a category of the *statistical population* of all phonemes in the language. By "statistical population" of phonemes, we understand the norms plus their probability of occurrence in actual speech. The set of norms plus the probabilities then corresponds to what de Saussure has called "langue", and their realisation in speech and writing to his "parole".

This means that *the phonemes retain their pre-realisation property* as the norms in the formation of the corresponding sounds in actual speech, and *are no longer identified with the phonometric averages, precisely as Trubetzkoy demanded it.*

A case like German /k/ would then be dealt with as follows. Since there is only one phoneme /k/ in the German language, all three types of realisation of that phoneme must be regarded as samples of it in "la parole". In order, however, to measure the sounds phonometrically, we must bear in mind the three sound varieties according to whether the phoneme is preceded by a consonant, by a stressed or unstressed vowel. This shows that according to our statistical conception there is no identity between phoneme and phonometric average.

The mathematical formulation of our conception of the relation between phonetics and phonology, which will occupy us now, has the advantage of exhibiting clearly the relation between the three important aspects of linguistics on the phonemic level: the *statistical distribution* of phonemes, the *entropy* of a phonemic system (in the sense of information theory) and their *linguistic distribution* in the language. It may be said without exaggeration that no other method but the mathematical is capable of clearly defining that three-fold relation.

It should be stressed that our conception does not refer to the methods used for the phonometric delimitation of particular phonemes, but to the simultaneous occurrence of the different phonemes in the language.

1. THE PHONEMIC FREQUENCY DISTRIBUTION – THE MULTINOMIAL THEOREM

The statistical universe of language at the phonemic (alphabetic) level is that of a multinomial population, that is a universe in which the probabilities of the different categories of the variable are given by the multinomial theorem. Basically, these are

universes of qualitative linguistic variables. The probabilities according to the multinomial theorem are calculated as follows.

We denote the probability of the *r* phonemes in a language by p_1, $p_2 \dots p_r$, wit the subsidiary condition that $\sum_{i=1}^{r} p_i = 1$. If a sample of *n* phonemes – *n* being a very great number – is taken at random from written or oral material of that language, and n_1, $n_2 \dots n_r$ are the numbers of phonemes falling into each of the *r* phonemic categories, and $n_1 + n_2 + \dots + n_r = n$, then, by the multinomial theorem, the probability of a particular combination of *n* phonemes is given by

$$P\,(n_1, n_2 \dots n_r) = \frac{n!}{n_1!\; n_2! \dots n^r!}\, p_1^{\;n_1} p_2^{\;n_2} \dots p^{r\,n_r} \qquad (1)$$

This is the probability that a sample of *n* phonemes from a running text with n_1, n_2 $\dots n_r$ occurrences of phonemes *1, 2, …r* resp. represents a random sample from the dictionary population of these phonemes with probabilities p_1, $p_2 \dots p_r$.[3]

It is characteristic for language that at any level or stratum of language structure there is a number of categories. If one category, say A, is considered versus all the other categories together, as non-A, the *multinomial* is reduced to the *binomial* and the probability according to (1) becomes that of finding A in specified parts of literary texts, such as lines, pages, volumes. This acquires importance for the problem of linguistic sampling from literary texts.

The smaller the linguistic units, that is the more the linguistic units are removed from, or independent from, meaning, the more will their frequency distribution (iu the statistical sense) conform with the probabilities calculated according to the multinomial theorem.

2. THE MULTINOMIAL THEOREM AND THE ENTROPY

Let p_1, $p_2 \dots p_r$ be the probabilities of *r* categories of a linguistic form in the language, and n_1, $n_2 \dots n_r$ the frequencies with which the categories occur in a sample of *n* units, and $n_1 + n_2 \dots + n_r = n$.

By the multinomial theorem,

$$(p_1 + p_2 + \dots + p_r)^n = \frac{n!}{n_1!\; n_2! \dots n^r!}\, p_1^{\;n_1} p_2^{\;n_2} \dots p^{r\,n_r}$$

The probability of a particular distribution is then given by

$$P = \frac{n!}{n_1!\; n_2! \dots n_r}\, p_1^{\;n_1} p_2^{\;n_2} \dots p^{r\,n_r}$$

Denoting the relative frequencies n_1/n in the sample by p_i', we have

$$\begin{aligned} \log P &= -n\Sigma\, p_i' \log p_i' + n\Sigma\, p_i' \log p_i \\ &= -n\Sigma\, p_i' \log p_i' + n\Sigma \frac{p_i'}{p_i}\, p_i \log p_i \end{aligned} \qquad (2)$$

[3] H. Levi and L. Roth, *Elements of Probability* (Oxford, 1936).

The first term on the right of formula (2) appears to be the *entropy* for n observations, and dividing by n we obtain Shannon's well-known formula for the entropy of a single observation. *We have thus derived the formula for the entropy according to information theory from the multinomial theorem*, as the term of P which does not contain the population values of p_i.

Apart from the theoretical interest, this affords the possibility of new computational methods for P according to the multinomial theorem by using tables of information values of p.

The second term on the right is of a very similar form, *but it contains the population values* p_i. It will occupy us in § 3.

3. LIKELIHOOD AS THE QUANTITATIVE EXPRESSION OF DISTRIBUTION IN THE LINGUISTIC SENSE

We have shown, so far, that what is called the probability of a phonemic frequency distribution is composed of two parts, one of which was recognized as the entropy. It now remains to give the linguistic interpretation of the other part. Our contention is that it represents the quantification of the concept of *distribution* as linguists understand the term.

To make the matter easier to follow, we replace formula (1) for multinomial probability by that for binomial probability, to which the formula reduces when, instead of all phonemes in a language, we observe one particular phoneme, say /e/, against all others lumped together as non-/e/. The multlnomial law goes over into the binomial probability law

$$^nC_r\, p^r\, (1-p)^{n-r}$$

where n is the size of the sample and r the number of members possessing the required quality. Choosing /e/ as the particular phoneme, formula (3) gives the probability that in a sample of certain size n, we shall find /e/ r times, and non-/e/ n-r times, if the probability of the phoneme /e/ in the population is p. Observe now that p, the population probability of the phoneme, does not occur in the first part ^nC_r, which in the multinomial law we have recognized to represent the entropy, but only in the second part. The expression $p^r\,(1-p)^{n-r}$ has been called by R. A. Fisher the "*likelihood*": it is not in itself a probability, as we have seen, but can be used as an instrument for selecting the most likely population from among a given class, simply by being maximised, which is a well defined mathematical procedure.

Likelihood, in the sense in which the term is used in mathematical statistics is, however, a quantitative concept. In order to understand the relation between likelihood and linguistic distribution, we may compare the mathematical presentation of language structure with a small-scale map describing language structure as a whole. Although a particular phonological opposition is a qualitative feature, *the summation of oppositions of the same type gives the likelihood as a quantitative characteristic of the language*. The probability patterns arrived at by contrasting the share which different types of phonological oppositions have in the phonemic system of a language are the likelihood patterns underlying the frequency distribution of phonemes.

We may, with Trubetzkoy, take it for granted that the phonemic system of a language consists of a vast network of phonological oppositions. If a phoneme count of dictionary material has been made, the resulting distribution (in the statistical sense) of phonemes as parts of words is the result of the accumulated patterns of linguistic distribution in the language.

Insofar as we are here dealing with an accumulation or sum of phonological oppositions, we have before us again a numerical quantity built upon the probabilities of such phoneme types as we have taken into consideration.

These basic probability patterns we call the likelihood. Since it is derived from the network of phonological oppositions in the language (dictionary), it must be clearly distinguished from the phoneme distribution in running texts, which can be regarded as random samples from the dictionary count as the population.

In conclusion, it may be well to emphasize the advantage procured by the statistical study of phonemics. I believe that most linguists will agree that, so far, they were not quite clear about the relation of distribution in the linguistic sense, probability or distribution in the statistical sense, and information theory. And, unhappily, this is one of the cases in which the relations become less clear the more they are talked about. This points to mere language not being the right instrument for understanding these important relations. A more powerful logical tool is required, and I believe to have shown that it is found in mathematical statistics. Using the theory of mathematical statistics and, in particular, starting with the multinominal theorem as the most appropriate statistical presentation of the co-existence of phonemes in the language, it was possible to show that the statistical distribution law of phonemes or their *probability* in the language has two parts: the *entropy*, representing the number of combinations of the basic arrangement of phonemes, and the likelihood which can be regarded as the summation of phonological oppositions in the language when suitably classified, or as the quantification of *linguistic distribution.*

University of Bristol