F. Knowles

THE QUANTITATIVE SYNTAGMATIC ANALYSIS OF THE RUSSIAN AND POLISH PHONOLOGICAL SYSTEMS

The description of a language's phonological system is not complete without statistically reliable data pertaining to the varied syntagmatic relations occurring within that system. A measure of functional load is needed to elucidate the efficiency, viewed from a cybernetic angle, of the particular set of phonemes. This approach opens up the way for various information-theoretic concepts such as entropy, relative entropy, information content and redundancy. These values are derived, with respect to the Russian and Polish phoneme inventories, ranging from zero to the third order. It is useful to consider the efficiency of various methods of information-theoretic coding procedures for phonemic strings. Huffman coding can achieve not only a high degree of efficiency but also a relatively trouble-free unique decoding routine.

In the syntagmatic analysis of phonological systems, junctures of varying types must be taken into account and must be " retrievable " for statistical purposes. The presence of count details relating to juncture allow an acceptable " transfer modulus " to be derived which can be a useful parameter in the comparison of a language's phonemic and graphemic systems.

The question of sample size is important in any statistical survey, not least in linguistics. Pilot tests have to be run in order to yield a general impression of a given phonemic distribution. Nearly all phonemic inventories show a Lexian or hyperbinomial distribution. Very often these distributions can be plotted logarithmically to give a reasonable fit to a theoretical exponential curve, derived from considerations similar to those postulated by Zipf in lexical statistics. Most phoneme systems contain rare phonemes – in Russian /š,/ is a case of such a phoneme – which present some difficulty, when probability intervals are to be estimated from relative frequencies. Using a 95 % confidence level and a 10 % error margin the sample size needed to satisfy these conditions with regard to the rarest phonemes tends to become very large – often of the order of several hundred thousand – whereas a sample of 20,000 phonemes is often sufficient to " pinpoint " phonemes of the highest frequency. A compromise decision regarding sample size has to be taken in which it appears reasonable to allow no more than 10 % of the phoneme inventory – that is, between four and five phonemes, normally – to exceed the stated confidence criteria.

One area of great interest to the phonologist is the extent of binary relations occurring within phoneme systems. What is needed is a numerical assessment of relational features such as completeness, transitivity, reflexivity, symmetry. These values are computed according to the methods proposed by Hovary and Paper. The results these methods give are partly statistical and partly Boolean but taken together it can be said that Hovary-Paper indices are important parameters of phonological systems.

It can also be shown that the distribution of phonemes over arbitrary " blocks " of text is of a Poissonian character. The division of a large corpus into many small blocks can also yield information about text homogeneity.

The sampling methods used to construct a large data-base are also important. If a corpus can be broken down into a small number, say twenty, of sub-corpora, a further test of homogeneity is available. The sampling distribution over these sub-units can also supply information needed to construct " population " probability intervals.

In much of computational linguistics it is as important to comment on computational methodology as on the linguistic results obtained. Mention must therefore be made of the specialised hardware and software that was necessary for implementing the above Russian-Polish contrastive phonemic analysis. The computer used was a second-generation English Electric KDF9, having 32K of core store and appropriate backing store in the form of discs and magnetic tapes. An extension of ALGOL 60 was used as the chief programming language. A series of input and output routines allowed special characters to be printed either by overprinting or by using two-line output. A number of specially designed procedures permitted the manipulation of characters, strings and the addressing of specified parts of machine-words.

The basic sorting process used was a tree-sort which, however, had to be adapted for a number of different uses. The programs used were:

a) sort of individual phonemes, giving frequency, relative frequency, standard deviation, functional load, entropy, redundancy, confidence limits, sample size, and Huffman codes.

b) sort of phoneme pairs (digrams) giving appropriate frequencies, matrix representation, second-order information-theoretic values, full range of Hovary-Paper indices.

c) sort of phoneme triples (trigrams), giving list of clusters and third-order information-theoretic values.

d) sort of individual text files, giving full details of phoneme distribution in " numerical " as well as alphabetic order.

e) sort of phonemes over " blocks ", to fit theoretical Poisson distribution and to assess homogeneity.

f) comparison of phonological systems, giving a measure of typological " distance ".