Askar Dzhubanov - Bakhitdzhan Khasanov

COMPUTATIONAL DESCRIPTION OF THE KAZAKH LANGUAGE

The subsequent development of the languages and their comprehensive study in connection with social progress is a natural process in our multinational socialist country. Such development and interrelation (or interaction) of languages can be observed in Kazakhstan, where the representatives of more than one hundred and twenty nations and nationalities live and the languages are developed in conditions of complete equality of rights. Some of these languages have already become the object of linguistical investigations. For example, the Kazakh linguists at present investigate the problems of the Kazakh, the Russian and the Uigur languages and their interaction with many other languages in our republic. Hence at the International conference on computational linguistics we should like to speak about our experiments of computational studying of the modern Kazakh language.

The first experiments of statistical study of the Kazakh language have been made by professor H. K. Dzhubanov at the beginning of the thirties. Then such professors as S. K. Kenesbayev, N. T. Sauranbayev, M. B. Balakaiev, S. A. Amanzholov and others made attempts to use a statistical method in their researches. Nevertheless it can be said that statistical methods have been accidentally used by Kazakh linguists. At present there is a group of " Statistical investigation and automation " and the laboratory of experimental phonetics.

The Institute regularly holds republican and all-union conferences on statistical and informative study of Turkish languages (1969), and also on the automatic recognition of acoustic images (1972), a number of scientific works have already been published by our scientific workers.[1] One of them is Dzhubanov's investigation, its theme " The statistical investigation of a Kazakh text with the use of a computer ", (the use of the material of Auezov's epic *Abai zholy* – Abai's path).[2]

---

[1] K. B. Bektayev, *Language Statistics (1957-1972)*, Alma Ata, 1972.

[2] M. O. Auezov, "*Abai*", *A novel*, Foreign languages Publishing House, Moscow, 1957.

First of all it was necessary to work out methods on the automatic construction of frequency lists of linguistic units with the use of a Turkish text where a computer is used for the first time. Apparatus coding text information, introductory and discharge equipment of modern computers are founded on Latin and Russian drawing. Therefore for the convenience of coding of the texts and the universality of machine programs for putting into shape the linguistic material of these languages, the unit transcriptions of identical phonemes of different Turkish languages are adopted in this work.

The investigation was made on the harmony of statistic distributions of classes of words of the Kazakh language by the following theoretical laws: normal, logarithmic-normal, distributions of Puasson and Sharlie in two types (*A* and *B*). It was established that the choice of a standard (amount of series, a number of series and the amount of choice) influences the character of distribution of the researching linguistic unit in this connection.

One can say that the language of the epic was statistically described completely.

The epic has 465,966 words (expressions): I book has 105,788, II book – 124,398, III book – 112,727, IV book – 123,053 – words (or expressions).

One hundred highly frequent wordforms have been commented in the investigation. It's interesting to remark that different conjunctions (*da, de*), particles (*ma, me*) are used more often. Many of them helped the author to create individual expressions. Then he commented auxiliary verbs (from the stems *e* and *de, otir, tur, gur, gatir*) and modal words; then the pronouns (*men, sen*) and the adverbs (*endi, kazir*), the adjectives, nouns, pronouns.

The frequency of equally used wordforms has been defined, the total number of them is 110. The correlation of highly frequent words in the epic in its whole and in its parts taken separately has been defined too.

Kazakh linguists are intending to analyse qualitatively all parts of speech of the Kazakh language using works of leading Kazakh poets and writers.

For instance, statistics for the lexicomorphological structure of the adjectives were studied. They are used 38,531 times in the epic which comprises 8.3 % of the whole wordbuilding. Various wordforms of the adjectives can be met 4,773 times without accounting for repetition, that is 7.7 % of all wordforms in the epic amounting to 61,824 words.

These percentage correlations express vividly the equality of distribution of words – adjectives in accordance with a word – list and text of the epic.

Statistic linguists of the Institute started studying statistically the complete works of M. Auezov consisting of 12 volumes. It is planned to investigate works of other leading Kazakh writers in the future.

The experiment and method worked out in the investigations of the Kazakh language have practical meaning in the statistic-linguistic study of other languages of the nations of Kazakhstan.