Chapter 1

METHODS FOR THE STATISTICAL ANALYSIS OF NARRATIVE TEXTS

1. The Maximum Correlation Principle for Historical Chronicles and Its Verification by Distribution Functions. Analysis of Russian Chronicles

The coefficient $d(X,Y)$ permitting us to estimate quantitatively the stability of the maximum principle relative to original data perturbations was introduced in Part 1, Ch. 2, § 4. We give here a short description of it, and recall the basic related concepts. Let $t_1(X),\dots,t_p(X)$ be the years in the time interval (A,B) in which the volume graph for a narrative text X exhibits splashes (= spikes or peaks). Suppose that the vector $T(X) = \big(t_1(X),\dots,t_p(X)\big)$ is related to the "authentic event" vector $T(A,B)$, where (A,B) is the period described in the text. If there are two texts X and Y, then the simplest relations among them are described by the diagram

$$T(X) \longleftarrow T(A,B) \longrightarrow T(Y) \quad \text{and} \quad T(A,B) \longrightarrow T(X) \longrightarrow T(Y).$$

As the proximity measure for $T(X)$ and $T(Y)$, we can use the following, viz.,

$$R(X,Y) = \sum_{i=1}^{p} \min \left|t_i(X) - t_j(Y)\right| + \sum_{j=1}^{q} \min \left|t_j(Y) - t_i(X)\right|.$$

For brevity, $R(X,Y)$ will be called in the following the *distance* between X and Y. I offered V. V. Fedorov from the All-Union Institute of Systems Research, Moscow, to verify the maximum correlation principle by the ordinary statistical methods. In 1981, Fedorov suggested the above function, which turned out to be convenient for computerization and was carried out with the participation of I. S. Shiganov. The meaning of $R(X,Y)$ is extremely simple. We fix a certain maximum for a text X, and find the nearest one of another text Y. We then calculate the distance in years between them, and sum these distances for all maxima of the first text. Interchanging the texts, we repeat the procedure. $R(X,Y)$ is obtained by summing up the results. I have performed the experiments without any further symmetrization of distance so defined, i.e., considering the first and second sum separately, thus being able to construct the non-symmetrical distance matrix.