International Statistical Review (1988), **56**, 3, pp. 279–301. Printed in Great Britain

Empirico-Statistical Methods in Ordering Narrative Texts

A.T. Fomenko

Faculty of Mathematics and Mechanics, Moscow State University, Moscow, 119899, USSR

Summary

A new method of processing statistically quantitative textual information of a narrative character is introduced. The method makes it possible to discover the chronologically correct order of separate textual fragments and recognize duplicates. The method can be used to recognize dependent and independent texts among large collections of texts. The concept of the text itself can be treated differently (in physics, linguistics, history, and so on).

Key words: Chronological order; Decks of cards; Dependent texts; Frequency damping principle; Frequency duplicating principle; Histogram; Historical texts; Matrix of the name; Random variables; Statistical duplicate recognition.

1 Introduction

The problem of recognizing dependences (and dependent texts) arises in many branches of applied statistics, linguistics, physics, genetics, etc. As applied to source research for example, of considerable interest is the discovery of dependent texts arising from a common primary source or original (possibly not surviving). On the other hand, it is useful to have an idea which texts are to be called independent or based on substantially different sources and archive data. Meanwhile, the concept of the text itself can be treated very differently. We can consider a sequence of symbols, signals, codes (of some nature), e.g. genetic codes in DNA chains, as a text. The general problem of search for 'dependent texts' lies in finding 'similar' portions in a given sequence, i.e. textual fragments 'duplicating' each other. Today there are many methods for finding dependences of this sort. We suggest certain new empirico-statistical procedures which can prove useful both in analyzing narrative texts (such as annals or chronicles) and in studying biological codes to find so-called homologous fragments, etc.

Suppose there were originally several decks of cards, identical in composition and (unknown) order P_0. Assume that the cards were then put in one large deck K, and shuffled, obtaining a new order P_1. Suppose that 'traces' of the initial order P_0 are retained in K, that is the shuffling is 'incomplete', and that the number of the original decks (and their volumes) is unknown, only assuming it to be considerably less than the volumes. How can we learn for a certain P_0 whether or not the deck K with order P_1 was obtained by the same method, and what was the initial order P_0?

The natural approach is to search for similar pieces in K. The more similar pieces are found, the more confidently we can assert that some or other piece preserves the influence of P_0. Thus, we can attempt to restore P_0 piecewise. Besides investigating the mutual disposition of similar pieces in K, we can determine whether or not the order P_1 is obtained, on the basis of inserting several decks with order P_0 somehow shifted relative to each other, as is always done in shuffling, and also find the shift values. We should,