Advances in Economics, Business and Management Research, volume 114
First International Volga Region Conference on Economics, Humanities and Sports (FICEHS 19)

Statistics of Numerals in the Text: Development of a New Method of Stylometry

Andrei V. Zenkov
Ural State University of Economics
Yekaterinburg, Russia
zenkow@mail.ru

***Abstract*—Two approaches to the statistical analysis of texts are suggested, both based on the study of numerals occurrence in a coherent literary texts. The first approach is related to the analysis of the frequency distribution of various first significant digits of numerals occurring in the text. The fre-quencies of occurrence of the digit 1, as well as, to a lesser extent, the digits 2 and 3, are usually a characteristic author's style feature, consistently manifested in all (sufficiently long) literary texts of any author. This approach is convenient for quick testing whether a group of texts has common au-thorship: the latter is dubious if the frequency distributions are sufficiently different.**

The second approach is the extension of the first one and requires the study of the frequency distribution of the numerals themselves (not their first significant digits). The approach yields non-trivial information about the author's style peculiarities and is suited for the advanced study of au-thorial texts.

The proposed approaches are illustrated by examples of computer analysis of the literary works by L. Dobychin and A. Platonov.

Keywords—stylometry; attribution of texts; text processing; numerals; first significant digit.

I. Introduction

The problems of this study relate to stylometry (the statistical study of texts to search for individual features of the author's style – in particular, for the at-tribution of texts). Traditionally, the length of sentences, the length of words, the frequency of use of function words and certain significant parts of speech, and even the frequency of letter combinations are analyzed for this. Unfortunately, different methods do not always lead to consistent conclusions.

II. Materials and Methods

In our works [1, 2, 3], a new approach, which consists in studying the fre-quency distribution of the first significant digits of the numerals used by the au-thors in their texts, is developed. The first significant digits 1, 2 (and, to a lesser extent, 3) usually have frequencies that are stable for all sufficiently large texts by this author. These are characteristic, statistically stable properties of the author's style. We associate them with psychological features that, regardless of the will of the author, unconsciously affect his/her texts. Consequently, there is a reasonable suspicion that the texts have different authorship, with significant differences in these frequencies for these two texts. Visual observation of differences is sup-ported by statistical fitting criteria.

To date, our methodology has been applied to Russian, Czech, and English-language literary texts. Using computers, quantitative and ordinal numbers were revealed in the texts. The specificity of a literary text is the predominance of ver-bal expression of numerals over digital one. In the first case, numerals (in differ-ent word forms) were first converted to a digital record, so, for example, for the numeral "four hundred and seventy-sixth" (476), only the first significant digit 4 was taken into account. In order to identify the author's use of numerals, idioms, and collocations that randomly contained numerals (Russian «семи пядей во лбу» (literally "seven spans in in the forehead", English analog "as wise as Sol-omon"), "to the four winds"), as well as bullet characters: 1), 2), 3), pagination, etc. were previously removed from the text.

In this study, we move from analyzing the statistics of the first significant numbers of numerals to analyzing the use of the numerals themselves in the au-thor's texts. The first of the two approaches can be considered a convolution of the second one. Each approach has its own advantages and disadvantages.

III. Results and Discussion

Counting the first significant digits makes sense only with respect to the sig-nificant digits 1, 2, and perhaps 3, since the occurrence of subsequent digits is subject to too strong fluctuations even in the texts of one author (see Fig. 1 be-low). Thus, only a small part of the statistical information on the numerals con-tained in the text is available for analysis. In addition, there is a problem with texts in languages in which the numeral "one" is formally indistinguishable from the indefinite article (although this is overcome by the transition to an intermedi-ary language that does not know such a problem). On the other hand, the infor-mation here is presented in a generalized form, which allows one to average spe-cific particular features of individual works of the author.

An analysis of the use of the numerals themselves (and not the first signifi-cant digits) gives richer information about the author's features of the text and, to a large extent, lacks the lack of indistinguishability of the numeral "one" and the

indefinite article. However, the analysis of statistics of numerals is technically more complicated.

We give a comparative example of the application of the original and ad-vanced analysis methods.

The literary texts of L.I. Dobychin and A.P. Platonov are distinguished by a sharp stylistic originality; ones find common literary sources and analogues in foreign literature [4, 5].

Fig. 1 shows the frequency distribution of the first significant digits of the numerals in the most voluminous works by Dobychin and Platonov.

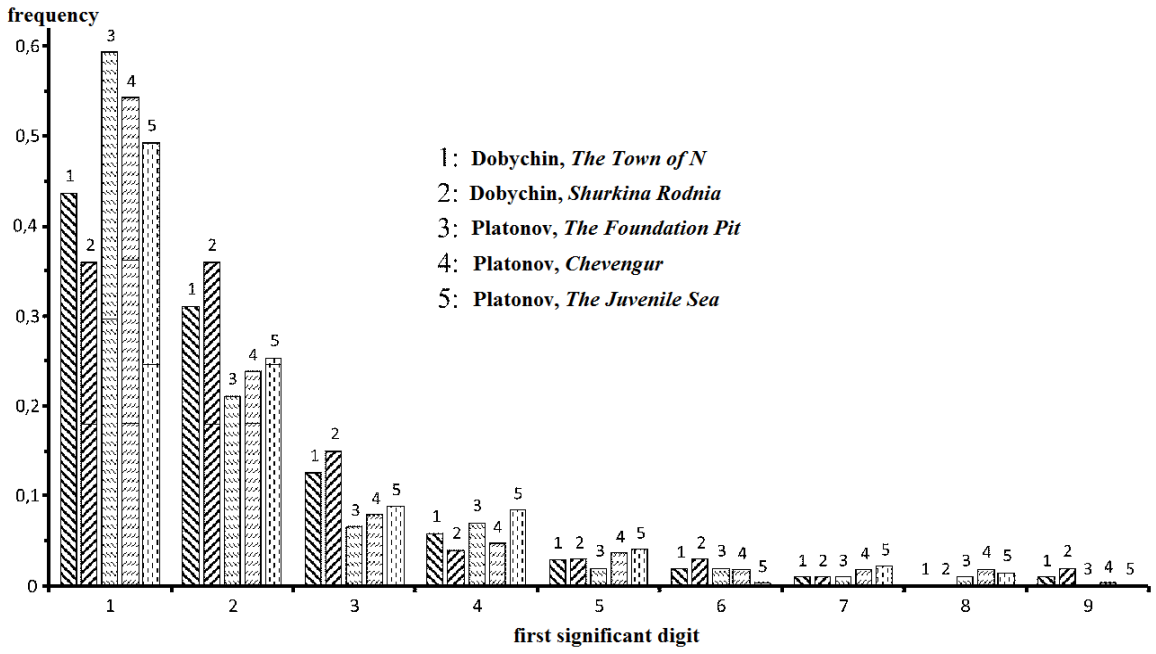

Fig. 1 Distribution of relative occurrence frequencies of the first significant digits of numerals in the texts by L. Dobychin and A. Platonov

The first significant figures 1, 2, 3 are characterized by a sharp difference in occurrence in the texts of Dobychin, on the one hand, and Platonov, – on the oth-er. This visual difference is reinforced by Pearson's statistical test . So, an analy-sis of the distribution of the first significant digits indicates undoubted style dif-ferences in the texts of the two authors. The method is convenient for quickly checking whether a certain group of texts belongs to one author: in the case of significant differences in the statistical distributions, single authorship is doubt-ful.

The results of applying the advanced statistical method that analyzes the oc-currence of the numerals themselves are incomparably richer. Figure 2 shows the frequencies of numerals from the range [0, 100] in the same texts by Dobychin and Platonov. Frequencies are adjusted taking into account the fact that texts dif-fer in volume. There are some results:

1. Platonov in his literary texts more likely use numerals than Dobychin.

2. Platonov less often resorts to rounding of numerals (10, 20, 30, ...), which, in conjunction with paragraph 1, can indirectly indicate a greater tendency to detail.

3. The numeral "one" (in different word forms) is the undisputed leader among the numerals found in Platonov's texts . But in the texts of Dobychin, the numeral "one" is inferior in frequency to the numeral "two"!

4. We note the psychologically understandable rarefaction of a number of numerals and a decrease in their occurrence as they increase, as well as a noticeable local

maximum at the numeral “one hundred,” which, of course, plays the role of an indefinitely large number.

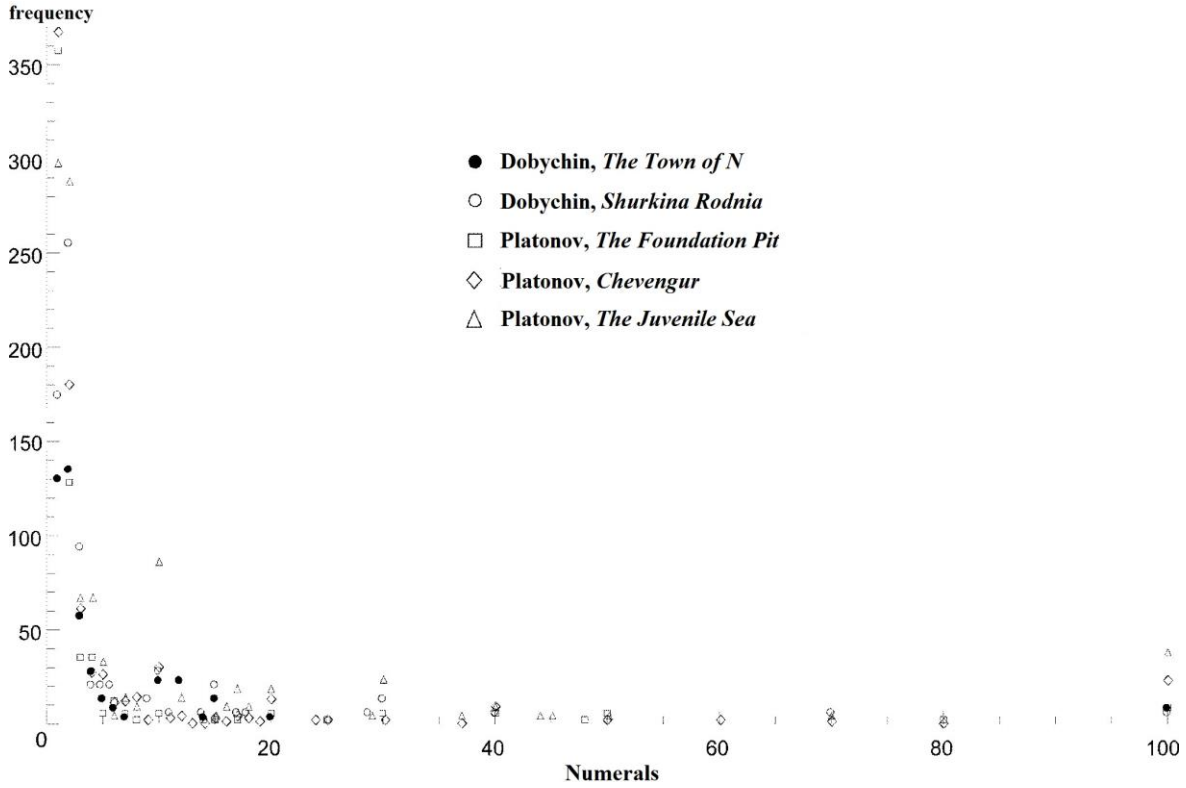

Fig. 2. Distribution of occurrence frequencies of numerals in the texts by Dobychin and Platonov

IV. Conclusion

We believe that the methodology developed by us can be a useful comple-ment to traditional stylometric practices for recording the length of sentences and words, the frequency of use of function words and certain significant parts of speech, etc.

Impressive results in stylometry can be achieved using neural networks, but the technique itself, unfortunately, is a “black box”: comprehension of the results is usually difficult. Our approach to stylometry is linguistically more meaningful.

Acknowledgment

This work was supported by a grant from the Russian Foundation for Basic Research (RFBR), project No. 19-012-00199A “A New Method for Attribution of Texts Based on Statistics of Numerals”.

This work was partially supported by a grant from SAIA (Slovenská akademická informačná agentúra) – Slovak Academic Information Agency.

References

[1] Zenkov A.V. A New Stylometry Method Based on Statistics of Numbers. Computer Research and Modeling. 2017. Vol. 9. No. 5. Pp. 837–850.

[2] Zenkov A.V. A Method of Text Attribution Based on the Statistics of Numerals. Journal of Quantitative Linguistics. 2018, Vol. 25, Issue 3, pp. 256–270.

[3] Zenkov A.V., Místecký M. The Romantic Clash: Influence of Karel Sabina over Mácha’s Cikáni from the Perspective of the Numerals Usage Statistics. Glottometrics. 2019, Vol. 46, pp. 12–28.

[4] Eidinova V.V. A. Platonov and L. Dobychin: Stylistic Convergence and Repulsion. The Land of Philosophers by Andrei Platonov: Problems of

Creativity. Issue 5. Anniversary: On the materials of the International Scientific Conference dedicated to the 50th anniversary of the death of A.P. Platonov. April 23–25, 2001. Moscow. – M., 2003. Pp. 211–219.

[5] Urban text in the XX century: Andrey Platonov. L. Dobychin. URL: http://dobychin.lit-info.ru/dobychin/articles/nazarenko-platonov-dobychin.htm