СКВОЗНЫЕ ЦИФРОВЫЕ ТЕХНОЛОГИИ И БИЗНЕС-ИНФОРМАТИКА

УДК 51-78, 519.234.3, 519.257, 81-139, 519.248.6

Zenkov Andrei,
Candidate of Phys. and Math. Sciences, Assoc. Prof., Assoc. Prof.,
Dept. of Modelling of Controllable Systems,
Graduate School of Economics and Management,
Ural Federal University,
Ekaterinburg, Russian Federation

Zenkov Eugene,
Cand. of Phys. and Math. Sciences, Assoc. Prof.,
Dept. of Theoretical Physics and Applied Mathematics,
Institute of Physics and Technology,
Ural Federal University,
Ekaterinburg, Russian Federation

Zenkov Miroslav,
master student,
Engineering School of Information Technologies,
Telecommunications and Control Systems,
Ural Federal University,
Ekaterinburg, Russian Federation

DATA ANALYSIS ON THE BASIS OF NUMERALS STATISTICS

Abstract:
Two approaches to content analysis of text data are suggested, both based on the statistical study of numerals occurrence in texts. The first approach is related to counting the frequency distribution of various leading digits of numerals occurring in the text. These frequencies are unequal: the digit 1 is strongly dominating; usually, the incidence of subsequent digits is monotonically decreasing. The frequencies 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, manifested in all (sufficiently long) literary texts of any author. This approach is convenient for testing whether a group of texts has common authorship: 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 numerals themselves (not their leading digits). The approach yields non-trivial information about the author, stylistic and genre peculiarities of the texts and is suited for the advanced stylometric analysis.
The proposed approaches are illustrated by examples of computer analysis of the literary texts in Lithuanian – by S. Daukantas, A. Baranauskas, Maironis, and J. Tumas-Vaižgantas.

Keywords:
Stylometry, quantitative linguistics, numerals, numerals in text, numerals statistics, first significant digit, leading digit, Benford's Law, text processing, text attribution, text authorship, textual criticism, Lithuanian literature, Daukantas, Baranauskas, Maironis, Tumas-Vaižgantas

1. Introduction

The paper is devoted to the development of an original method of content analysis of text data, rather than the software implementation of known algorithms. Our content analysis aims at searching for the characteristic features of the author's style of literary texts as well as the attribution of texts. More generally speaking, it is the search for the author's invariant – characteristic (quantitative) features inherent in all or most of the texts of a given author (or at least texts long enough for statistically significant analysis) and distinguishing his texts from the texts of other authors.

We have proposed a new method of statistical study of connected author texts. This method is based on the analysis of the occurrence of numerals contained in the text, as well as counting the frequencies of their first significant digits. The proposed approach complements the existing methods of stylometry, favorably differing from most of them by the possibility of meaningful linguistic interpretation of the results.

Consider some well-known stylometric techniques:

1. Historically, the earliest, widespread and fairly easy to implement (including on a computer) are various versions of the statistical method. Some varieties of it:

a) Taking into account the length of sentences, length of words, frequencies of use of certain significant parts of speech, etc. [1];

b) Counting the frequency of service words (such as not, and, or, etc.) used by the author (see, for example, [2]);

c) It would seem that a suitable tool for attribution of texts, searching for proximity/differences of style and visualization of textual data obtained at steps *a* and/or *b* may be the hierarchical cluster analysis that distributes a set of n-dimensional vectors into groups so that vectors of one group are "closer" to each other than vectors of different groups [3]. Unfortunately, cluster analysis is not free from subjectivity associated with the choice of clustering method and metrics; this choice cannot be rigorously substantiated (meanwhile, it significantly affects the results obtained);

d) A characteristic feature of the author's style can be the so-called “rare pairs” – individual author's collocations: the appearance of certain words in the text may, for a particular author, entail the use of other predictable words [2];

e) Attempts have been made to attribute texts by counting the frequencies of letter combinations (digrams and trigrams) [4]. If for a poetic text with alliterations deliberately introduced by the author, this seems to be acceptable (although, in any case, alliterations are occasional and therefore cannot be a subject of statistical study), then with regard to prose it is doubtful.

2. Neural networks, a form of statistical machine learning methods, can be used to examine the authorship of texts. The texts of undisputed authorship are used to train the neural network. The network is able to generalize its recognition ability to new texts that have not yet been presented to it, classifying them with a certain degree of confidence. For examples of successful application of the method to stylometry problems, see [5, 6]. The use of neural networks can yield good results, but the technique itself is a *black box*: understanding the results is usually difficult. Meanwhile, in science, in contrast to business, the methodology for obtaining a result is no less important than the result itself.

The approach we are developing is statistical one. Of course, it does not cancel, but complements the approaches listed above.

2. Methods

We will proceed to the new ideas presented in this research after recalling the statistical Benford’s Law. It describes the probability of the appearance of a certain first significant (leftmost non-zero) digit in various real life numerical data sets. Common sense suggests that the probability distribution is uniform (the appearance of any first significant digit should be equally possible), but for many data sets this is not the case: as the first significant digit, 1 is noticeably more common than other digits. According to Benford's Law, in the decimal system, the first significant digit *d* appears with probability

$$P(d)=\lg\left(1+\frac{1}{d}\right).$$

Thus, $d = 1$ should occur with probability $\lg 2 = 0.30$, $d = 2$ – with probability 0.18, etc.

A complete explanation of Benford's Law, applicable to all cases of implementation, is missing, although some sufficient conditions for its appearance have been formulated. The incompleteness of understanding does not prevent the application of Benford’s Law to identify all kinds of "falsifications": fraud in accounting, in elections, etc. [7, 8].

In our papers [9–11], the prospects of counting the frequencies of various first significant digits of numerals in quantitative linguistics are explored – for textual problems. Both for a random combination of texts and for connected texts, the frequency distribution turned out to resemble Benford's one, but the proportion of the first significant digit 1 noticeably exceeds 30 per cent – if only because, formally being a numeral, the word *one* can act as the indefinite article. The psychological tendency to round numbers also plays a role.

In contrast to the traditional methodology of applying Benford's Law, which considers deviations from the law as an indication of possible falsifications (in a broad sense), we focus on comparing these deviations for texts by different authors. It is shown that the frequencies of occurrence of the first significant digits 1, 2 (and, in part, 3) are statistically stable authorial features distinguishing between the texts by different authors (under certain conditions, the most important of which is a sufficiently long text).

To date, we applied our methodology to literary texts in Russian, English, Lithuanian, Czech, and from non-Indo-European languages – in Turkish.

From analyzing the statistics of the first significant digits of numerals, we have taken the next step in the latest works – to analyzing the use of numerals themselves in authorial texts [12–15]. Each of the approaches has its advantages and disadvantages.

Counting the first significant digits makes sense only in relation to the significant digits 1, 2 and, possibly, 3, since the occurrence of subsequent digits is subject to strong fluctuations even in the texts by one author. Thus, only a small part of the statistical information about the numerals contained in the text is available for analysis. In addition, there is a problem with texts in languages (German, French, ...) in which the numeral *one* is formally indistinguishable from the indefinite article (although this can be overcome by switching to an intermediary language without such a problem; the distortions introduced by translation are small). On the other hand, the information is presented here in a generalized form, which makes it possible to average out specific particular features of individual works of the author.

The analysis of the usage of the numerals themselves (not the first significant digits) provides richer information about the author's features of the text and, to a large extent, is not hindered by the indistinguishability of the numeral *one* and the indefinite article. However, the analysis of numerals statistics is technically more difficult.

3. Program realization of new research methods

We have created a computer program that calculates the frequencies of occurrence of various cardinal and ordinal numerals in texts, as well as their first significant digits. The specificity of the use of numerals in a literary text is the noticeable predominance of the verbal expression of numerals over digital. In the first case, numerals (in different word forms) were first converted into digital notation, so that, for example, for the numeral *one thousand four hundred ninety-second* (1492nd), only the first significant digit **1** will be taken into account. Declension of numerals made the accounting of word forms in Russian, Czech and Lithuanian especially laborious. To identify the author's use of numerals, idiomatic expressions and stable phrases that accidentally contain numerals (*as clear as two and two makes four* or *to drink like seven lords*), as well as list markers (like 1., 2., 3.,…), pagination, etc. were previously removed from the text.

4. Objects of research

We have analyzed the numerals usage in a large amount of literary texts in Lithuanian. Lithuanian literature is young, in the XIX–beginning XX century it went from actually the first samples of secular literary texts to mature realism, which allows us to explore the relationship between the historical epoch in which the literary text was created and the frequency distribution of numerals in it. At the same time, the authors who wrote in the Lithuanian language in the XIX century are still few in number, which makes it possible to study this relationship almost exhaustively using the example of one national literature. The major works of Simonas Daukantas (1793-1864) [16, 17], Antanas Baranauskas (1835-1902) [18], Maironis (born Jonas Mačiulis, 1862 – 1932) [19], and Juozas Tumas-Vaižgantas (1869-1933) [20] are studied.

The texts by Daukantas we analyzed are

1. *Darbai senųjų lietuvių ir žemaičių* [Deeds of the Ancient Lithuanians and Samogitians],
2. *Būdas senovės lietuvių, kalnėnų ir žemaičių* [The Character of the Ancient Lithuanians, Highlanders, and Samogitians],
3. *Pasakojimas apie veikalus lietuvių tautos senovėje* [Story of the Deeds of the Lithuanian People in the Ancient Times].

The text by Baranauskas analyzed is the poem *Anykščių šilelis* [The Forest of Anykščiai].

The texts by Maironis we analyzed are the poems

1. *Jaunoji Lietuva* [Young Lithuania],
2. *Raseinių Magdė* [Magde from Raseiniai],
3. *Tarp skausmų į garbę* [From the sorrows to honor].

The texts by Tumas-Vaižgantas analyzed are

1. *Aukštaičių vaizdeliai* [Scenes of Aukštaičiai],
2. *Pragiedruliai* [Cloud Clearing],
3. *Dėdės ir dėdienės* [Uncles and Aunts].

The texts are very different in size. For the comparability of the analysis results, we introduced the size correction factors.

5. Results and Discussion

A more detailed presentation of the results with an attachment of the graphics will be published elsewhere.

Here, merely a short resume:

- The works by Daukantas are scholarly texts, albeit having an archaic style and pertaining to the humanities. This is reflected in the distribution of the numerals in his text. There are relatively few of them, mainly dates and round numbers.
- The works by Baranauskas and Maironis are poetic texts. In them, in comparison with the prose texts of other authors, there are very few numerals.
- The texts of Tumas-Vaižgantas are critical realism replete with numbers.
- The distribution of numbers in the texts is specific to each author and allows one to separate the texts from each other by authorship.
- So, the use of numerals in texts depends on the nature of the text (prose/poetry, scientific/literary) and allows one to check the authorship of the text.

We made similar conclusions earlier when analyzing literary texts in other languages (Russian, Czech, English, Turkish).

6. Conclusion

The frequencies of occurrence of the first significant digits of numerals in coherent literary texts are not the same: the digit one sharply prevails; the occurrence of subsequent digits usually decreases gradually. This is found directly for texts in Russian, Czech, Lithuanian, English, and Turkish.

The frequencies of occurrence of the first significant digit 1, as well as digits 2 and 3 (to a lesser extent), are usually a characteristic peculiarity of the author's style, consistently manifested in all (sufficiently long) literary texts of this author and proven by statistical tests.

Significant differences in these frequencies for given texts are an indication that the texts may have different authorship. Thus, the analysis of the frequencies of the first significant digits of the numerals can be used to solve the stylometric problems. Of course, frequencies similarity still does not prove the authorship identity with absolute certainty.

Taking into account the occurrence of numerals themselves (not the first significant digits) in literary texts can provide information about the authorial, stylistic and genre features of texts. Sometimes, an analysis of the occurrence of numerals allows to reject the hypothesis of the common authorship of texts.

We believe that our methodology can be a useful addition to the traditional stylometric practices such as analysis of words and sentences length, analysis of certain words use, etc.

7. Acknowledgements

The work was supported by the grant No. 19-012-00199A from the Russian Foundation for Basic Research.

A partial financial support was received from Slovenská akademická informačná agentúra, thanks to which the author's research stay at the Filozofická fakulta, Univerzita Pavla Jozefa Šafárika v Košiciach became possible. A special thank is due to Dr. Renáta Panocová, Vice-Dean for International Relations of the Faculty of Arts, for her hospitality.

REFERENCES

1. B. Ryabko, J. Astola, M. Malyutov, Compression-Based Methods of Statistical Analysis and Prediction of Time Series, Springer International Publishing AG Switzerland, Basel, 2016.
2. E. Stamatatos, A Survey of Modern Authorship Attribution Methods, Journal of the American Society for Information Science and Technology 60 (2008) 538–556, doi: 10.1002/asi.21001.
3. Guojun Gan, Chaoqun Ma, and Jianhong Wu, Data Clustering: Theory, Algorithms, and Applications, SIAM, Philadelphia, 2007.
4. C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT Press, Cambridge, Massachusetts, 2000.
5. M. L. Brocardo, I. Traore, I. Woungang, M. S. Obaidat, Authorship verification using deep belief network systems, Int. J. Commun. Syst. (2017). doi:10.1002/dac.3259.
6. R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics, Oxford University Press, Oxford, 2003.
7. M. J. Nigrini, Benford's Law: applications for forensic accounting, auditing, and fraud detection, John Wiley & Sons, Hoboken, 2012.
8. Benford Online Bibliography. URL: http://benfordonline.net/
9. A. V. Zenkov, A novel method of stylometry based on the statistic of numerals, Computer Research and Modeling 9 (2017) 837–850 (In Russ.).
10. A. V. Zenkov, A Method of Text Attribution Based on the Statistics of Numerals, Journal of Quantitative Linguistics, 25 (2018) 256–270. doi: 10.1080/09296174.2017.1371915.
11. A. V. Zenkov, M. Místecký, The Romantic Clash: Influence of Karel Sabina over Macha's Cikani from the Perspective of the Numerals Usage Statistics, Glottometrics, 46 (2019) 12–28.
12. A. V. Zenkov, Numerals in authorial texts and the discourse analysis, Bulletin of NCSPU 3 (2020) 45–51.
13. A. Zenkov, E. Zenkov, A. Belke, A Novel Text Analysis Method: Numerals Reveal the Author, in: Proceedings of the 3rd International Scientific Conference on New Industrialization and Digitalization (NID 2020), SHS Web of Conferences, Vol. 93, 2021, Article No. 03026, 6 pages, doi https://doi.org/10.1051/shsconf/20219303026, published online 12 January 2021.
14. A. Zenkov, E. Zenkov, M. Zenkov, L. Sazanova, Numerals in authorial Turkish-language texts and the stylometric analysis, in: Proceedings of the International scientific forum on computer and energy sciences (WFCES 2021), E3S Web of Conferences, Vol. 270, 2021, Article No. 01038, 5 pages, doi https://doi.org/10.1051/e3sconf/202127001038, published online 9 June 2021.
15. A. Belke, A. Zenkov, L. Sazanova, Education and sustainable development: interplay and implications, E3S Web of Conferences, 208 (2020) 09010, doi: https://doi.org/10.1051/e3sconf/202020809010.
16. S. Daukantas, Raštai: 2 t., Vaga, Vilnius, 1976. – (Lituanistinė biblioteka).
T.1 [Darbai senųjų lietuvių ir žemaičių; Būdas senovės lietuvių, kalnėnų ir žemaičių; Smulkieji raštai]; T.2 [Pasakojimas apie veikalus lietuvių tautos senovėje; Laiškai].
17. Antologija: klasikinė lietuvių literatūra. URL: http://antologija.lt/text/simonas-daukantas-budas-senoves-lietuviu-kalnenu-zemaiciu
18. Vikišaltiniai. URL: https://lt.wikisource.org/wiki/Anykščių_šilelis
19. Vikišaltiniai. URL: https://lt.wikisource.org/wiki/Maironis
20. J. Tumas-Vaižgantas, Rinktiniai raštai: 2 t., Valstybinė Grožines Literatūros Leidykla, Vilnius, 1957.