International Journal on Data Science and Technology
2017; 3(2): 16-23
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20170302.11
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

A New Stylometry Method Basing on the Numerals Statistic

Andrei Viacheslavovich Zenkov[1, *], Larisa Anatolievna Sazanova[2]

[1]Department "Modelling of Controllable Systems", Ural Federal University, Ekaterinburg, Russia

[2]Department of Statistics, Econometrics and Computer Science, Ural State University of Economics, Ekaterinburg, Russia

Email address:
zenkow@mail.ru (A. V. Zenkov)

[*]Corresponding author

To cite this article:
Andrei V. Zenkov, Larisa A. Sazanova. A New Stylometry Method Basing on the Numerals Statistic. *International Journal on Data Science and Technology*. Vol. 3, No. 2, 2017, pp. 16-23. doi: 10.11648/j.ijdst.20170302.11

Received: March 22, 2017; **Accepted**: April 25, 2017; **Published**: May 22, 2017

Abstract: A new method of statistical analysis of texts is suggested. The frequency distribution of the first significant digits in numerals of connected authorial English-language texts is considered. Benford's law is found to hold approximately for these frequencies with a marked predominance of the digit 1. Deviations from Benford's law are statistically significant author peculiarities that allow, under certain conditions, to consider the problem of authorship and distinguish between texts by different authors. At the end of {1, 2,…, 8, 9} row, the digits distribution is subject to strong fluctuations and thus unrepresentative for our purpose. The approach suggested and the conclusions are backed by the examples of the computer analysis of works by W. M. Thackeray, M. Twain, R. L. Stevenson et al. The results are confirmed on the basis of non-parametric range Mann-Whitney and Kruskal-Wallis tests as well as the parametric Pearson's chi-squared test.

Keywords: Benford's Law, Statistic of Numerals, Text Attribution, Text Processing, English-Language Fiction, Mann-Whitney U Test, Pearson's Chi-Squared Test

1. Introduction

Recently, the scope of the practical use of Benford's law [1] has significantly expanded. Known for over a hundred years, Benford's law refers to the probability of occurrence of a certain first significant digit in the distribution of various real life data. Contrary to the common assumption that the frequency of occurrence of *any* first significant digit should be equal, the digit 1 occurs more likely for many data sets! According to Benford's law, in the decimal system, probability of occurrence of the digit d as the first significant

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

accordingly, the probability of $d = 1$ should be $\lg 2 \approx 0.30$, the probability of $d = 2$ – 0.18, etc.

An exhaustive explanation of Benford's law, covering all cases of its manifestation, has not yet been proposed, although some conditions favouring its emergence are stated. A classic experiment by Benford, showing a good agreement with (1) – analysis of the occurrence of numerals contained in articles of a randomly selected issue of a magazine – is naturally explained by the theorem by Hill [2], according to which, if one repeatedly randomly chooses a probability distribution and then randomly chooses a number according to that distribution, the resulting data set will obey Benford's law. Note that Benford himself analyzed the occurrence of numerals expressed in *figures* only.

Incomplete understanding [3] does not preclude the successful use of Benford's law to detecting fraud in accounting and auditing data [4] and election fraud [5]; the applications suggested extend from physics and astronomy [6, 7] through seismology [8] to steganography [9] and scientometrics [10].

Zenkov [11] has shown the efficacy of counting frequencies of different first significant digits of numerals for text attribution. It was found that not only for the *random* combination of heterogeneous texts, but also for the *coherent* (Russian-language) texts to which the conditions of the afore-named theorem are not applicable, frequency distribution resembles that of Benford's law (1), but the quota of digit 1 considerably exceeds 30 per cent – at least since the word "one" formally being a numeral can actually play the role of an indefinite article.

In contrast to the traditional methodology of application of Benford's law, which treats deviations from the law as an indication of the possible existence of "falsification" (broadly defined), he placed emphasis on the comparison of these deviations for texts by different authors, showing that these deviations are statistically robust author features that allow to distinguish between texts by different authors (under certain conditions, the most important of which is a sufficiently large text).

Basing on these ideas, we present here new research results concerning the distribution of the first significant digits of numerals contained in coherent *English*-language texts.

The study is of an empirical and experimental nature. The aim of the theoretical explanation of the results (if at all possible) is not intended which, however, does not diminish the possibility of the practical use of the proposed methodology for practical problems of stylometry.

For all (English-language fiction) texts subjected to computer-aided statistical analysis, we have studied the frequency of occurrence of various first significant digits of numerals, taking into account cardinal as well as ordinal numerals expressed both in figures, and (considerably more often) verbally. In the last case, the first step was to rewrite every form of a numeral with figures (e.g., 'one thousand, seven hundred and eighty-ninth' replaced by '1789') and then to take into account the first significant digit (1) only. To identify the author's use of numerals, we previously deleted from the text all idiomatic expressions and set phrases *accidentally* containing numerals ('one hand washes the other', 'five-o'clock'), as well as itemizations like 1), 2), 3), etc.

Texts analyzed are mainly taken from the Project Gutenberg website http://www.gutenberg.org

2. Distribution of First Significant Digits of Numerals in Compound Texts

The conditions of Hill theorem are best satisfied for the compound texts containing the pieces by different authors. In this case, the author peculiarities are averaged, and we obtain a Benford-like frequency-digit dependence but with a steeper drop and the occurrence of 1 much more predominant than prescribed by Benford's law (1).

The Figure 1 shows the results of the analysis of eight English-language compound fiction texts [12–19]. For each collection of stories, we see a monotone decrease of frequency; results for different collections are upon the whole similar, variations may be owing to peculiarities (for example, genre and time of creation) of texts in each collection.

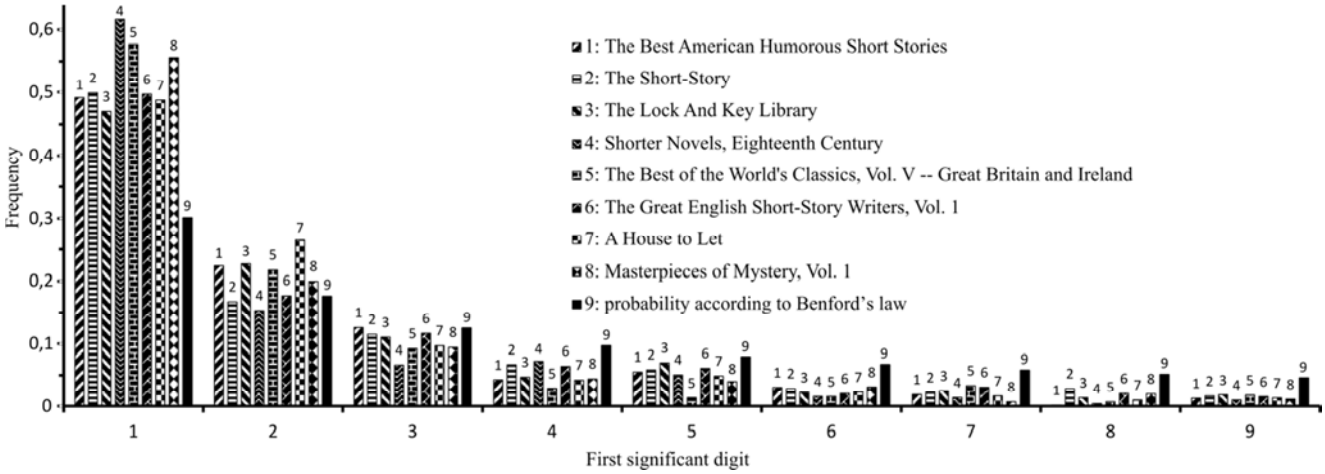

Figure 1. *The distribution of the first significant digits of numerals in English-language compound fiction texts; the results are compared with those expected according to Benford's law (1).*

3. Distribution of First Significant Digits of Numerals in Coherent Texts

Usually, texts belonging to the pen of a distinct author have persistent peculiarities in the statistics of first significant digits of numerals, and their distribution is a stable characteristic of the author.

As an example, we show here the distributions of the first significant digits of numerals in texts by W. M. Thackeray, M. Twain, and R. L. Stevenson (Figures 2–4).

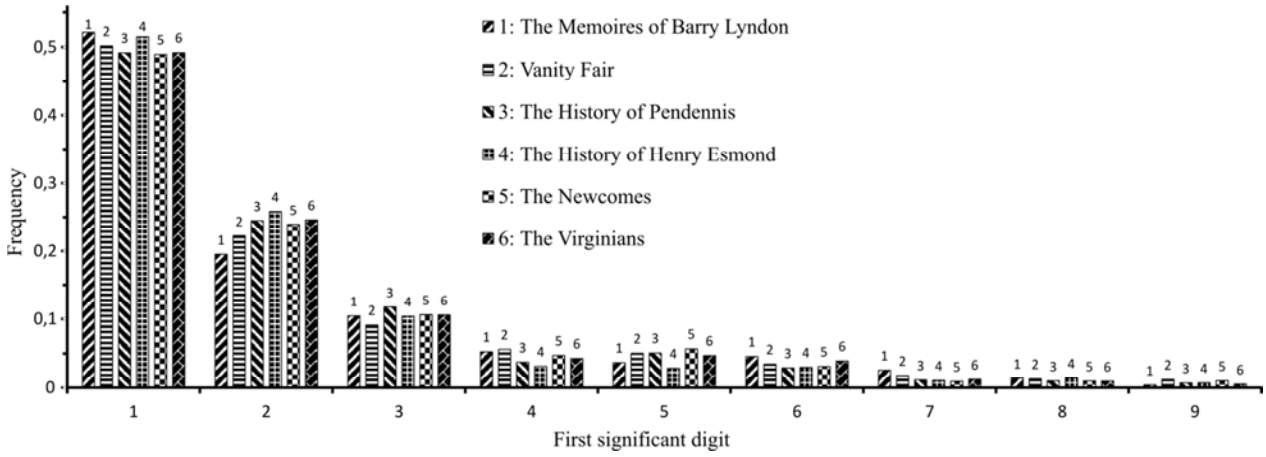

Figure 2. *The distribution of the first significant digits of numerals in texts by W. M. Thackeray.*

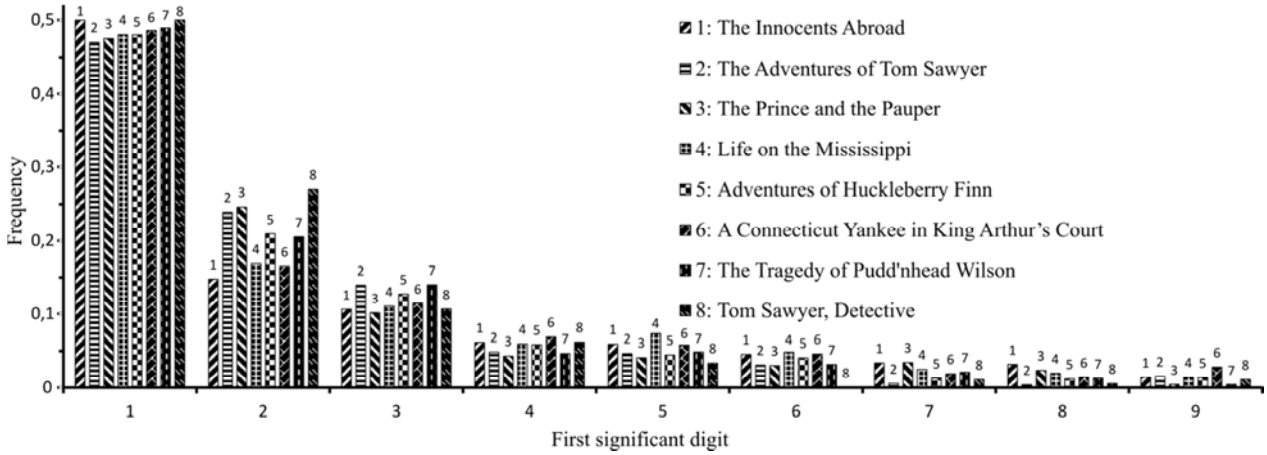

Figure 3. *The distribution of the first significant digits of numerals in texts by M. Twain.*

Note two digit 1 outliers corresponding to texts *not wholly* written by Stevenson.

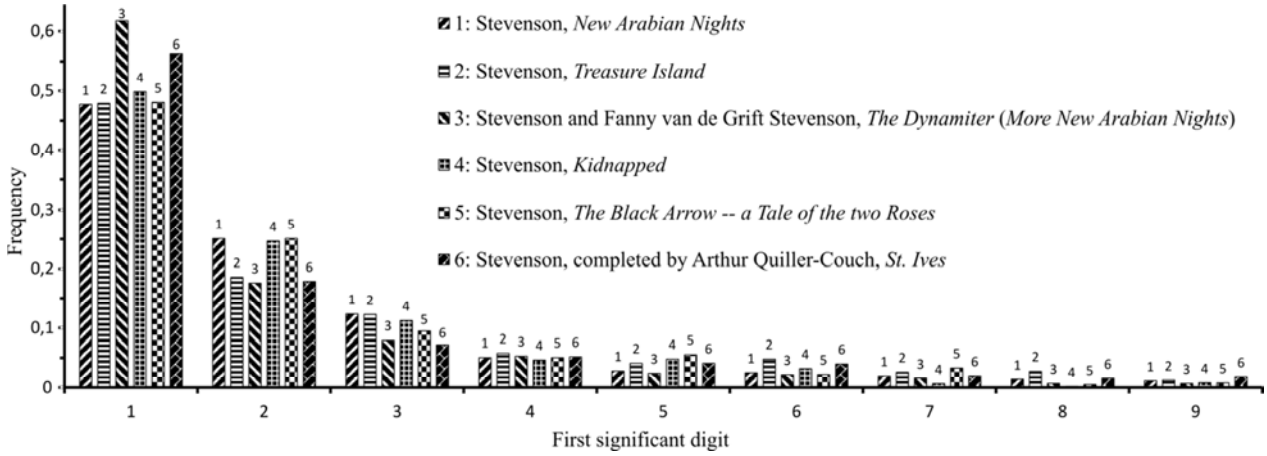

Figure 4. *The distribution of the first significant digits of numerals in texts by R. L. Stevenson.*

The differences in the statistics of first significant digits of numerals in texts by different authors may be not striking, as in case of novels by sisters Brontë. This is in fact not surprising in view of their common family and education background.

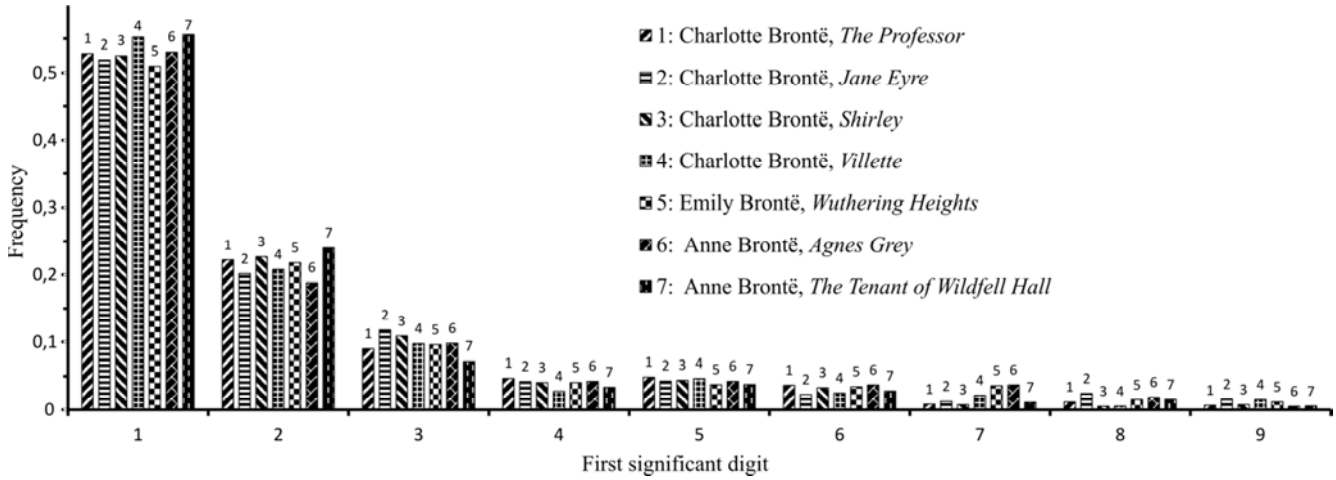

Figure 5. *The distribution of the first significant digits of numerals in texts by the Brontës.*

The frequency of digit 1 can reach the two times higher value than according to Benford's law (Figures 1–5). It is this digit as well as digits 2 and 3 (to a lesser degree) which determine the author peculiarity of texts in our approach. The occurrence of subsequent digits is subject to strong fluctuations which precludes obtaining useful information from their distribution. In Figures 2–5, the frequency of the digit 1 usually was about 0.5; as it will be shown later, this frequency can strongly differ from that value.

The frequency of digit 1 is, so to speak, a 'fingerprint' which permits to distinguish between different authors if this frequency strongly differs for their texts. How strong should be the difference, to be regarded as significant? We will answer this question at the end of the article.

4. Text Attribution

4.1. Jane Austen and Her Imitators

Domestic novels of manners by Jane Austen (1775–1817) caused numerous sequels and prequels. Related topics and even the intention to write in the same way did not prevent the imitators from stark difference in the numerals usage (Figure 6).

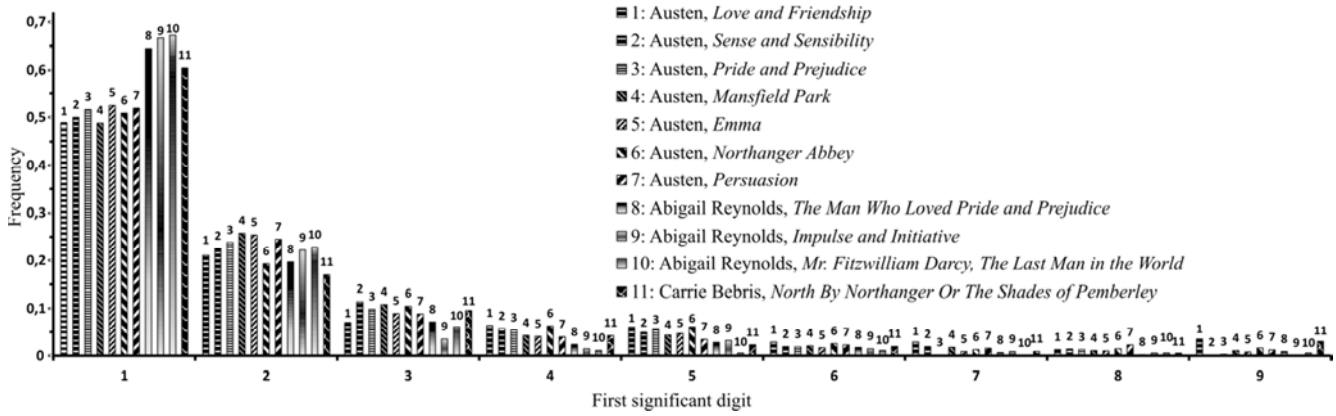

Figure 6. *The distribution of first significant digits of numerals in J. Austen's novels and in those of her epigones.*

Thus, Benfordian analysis can be useful in the study of text's authorship.

4.2. Authorship of the 15th Book of Oz

Lyman Frank Baum, a prolific writer whose "Wonderful Wizard of Oz" was a great success, wrote until his death 13 sequels of this book. The series was so popular that the publishers decided to continue it. The 15th book, 'The Royal Book of Oz', published after Baum's death, was written "by L. Frank Baum,…, Enlarged and Edited by Ruth Plumly Thompson" as noted on the title page of the first edition (1921). Subsequently, the point of view has spread (argued by linguistic and statistical means) that Thompson did not base the story on any notes Baum left behind, thus "The Royal Book of Oz" was entirely her own work [20]. This opinion is now generally accepted.

Although this particular philological question has already been solved, we will show the results of applying our methodology.

Below are the results of the statistical study of Baum's books as well as sequels by Thompson and by other authors (Figures 7–9).

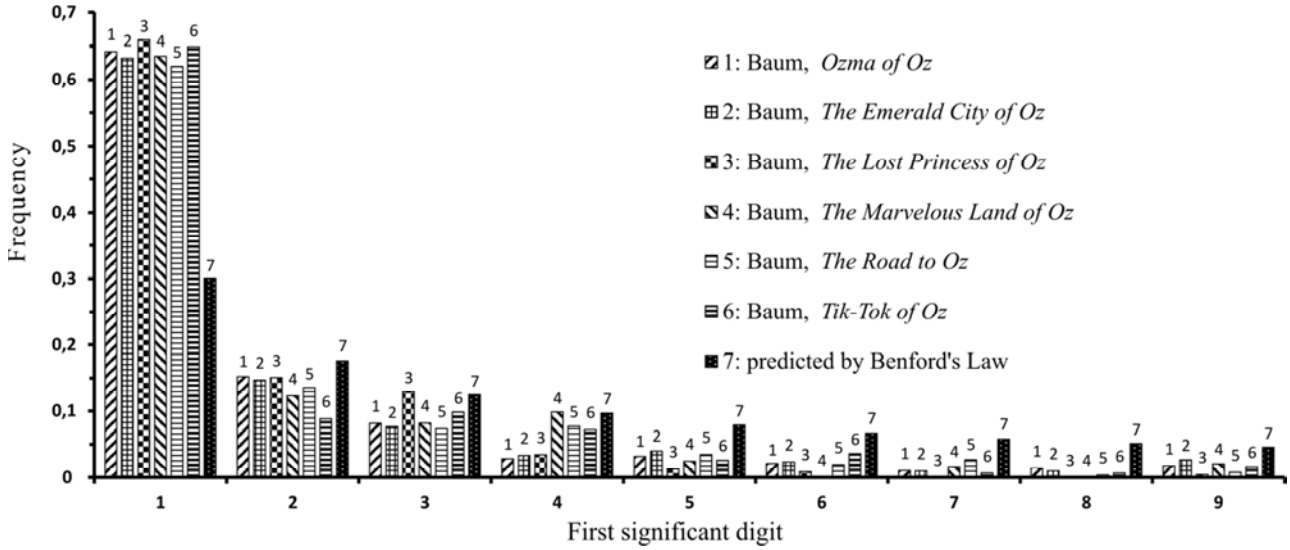

Figure 7. *The distribution of the first significant digits of numerals in texts by L. F. Baum.*

Note a dramatic difference in the occurrence of significant digit 1 in Baum's texts, on the one hand, and in texts by Thompson (in particular, in "The Royal Book of Oz"), on the other hand. In view of the length of the texts analyzed, this striking difference can hardly be explained by random fluctuations (unlike subsequent significant digits, which even in the books by the same author behave differently); it demonstrates the authorship of Thompson.

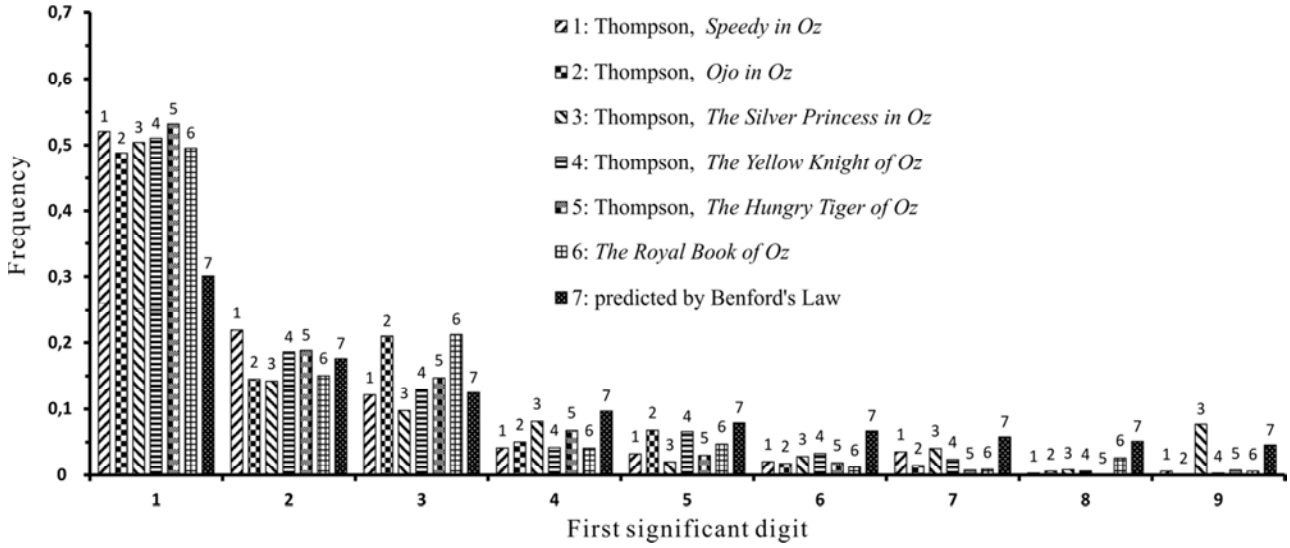

Figure 8. *The distribution of the first significant digits of numerals in texts by R. P. Thompson.*

Besides Thompson, many other writers created sequels for "Wonderful Wizard of Oz". Again, the common theme did not cause the similar distributions (Figure 9). We are prone to regard this difference as a characteristic of the author's style. We tend to associate it with the psychological peculiarities that, regardless of the will and intention of the author, influence his texts.

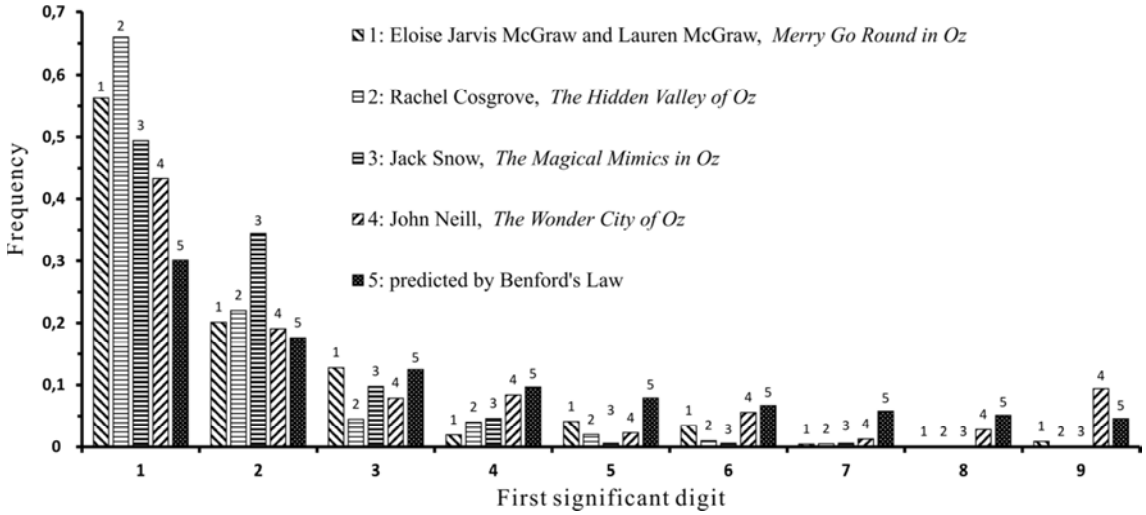

Figure 9. *The distribution of the first significant digits of numerals in sequels of Oz by other authors.*

Thus, the statistical method based on counting the first significant digits of numerals, is able to answer the question about the text authorship.

4.3. Testing of Methodology: Harper Lee and Truman Capote

Harper Lee's "To Kill a Mockingbird", published in 1960, is considered one of the greatest novels of American literature. In 2015, short before her death, another novel, "Go Set a Watchman", was published. Initially promoted by its publisher as a sequel, it is now widely accepted as a first draft of her famous novel.

Truman Capote was a lifelong friend of Harper Lee. One of the characters in "To Kill a Mockingbird" was based on him. In contrast to Lee who in fact is the author of a sole book, he was much more prolific, and many of his works are recognized literary classics. The speculation eventually grew that Capote ghosted Lee's book.

Testing this hypothesis is an interesting application of the idea about the relation of text authorship to its statistical characteristics.

We have counted the frequencies of various first significant digits of numerals in novels by Harper Lee and Truman Capote (Figure 10). Results of the analysis are unexpected: properties of the novel "To Kill a Mockingbird" are *far* from those of Capote's texts, but the primary draft, "Go Set a Watchman", is close to them. It seems that Capote *could* help Harper Lee in writing the primary text. After having gained experience, she seems to have written her famous novel by *herself*.

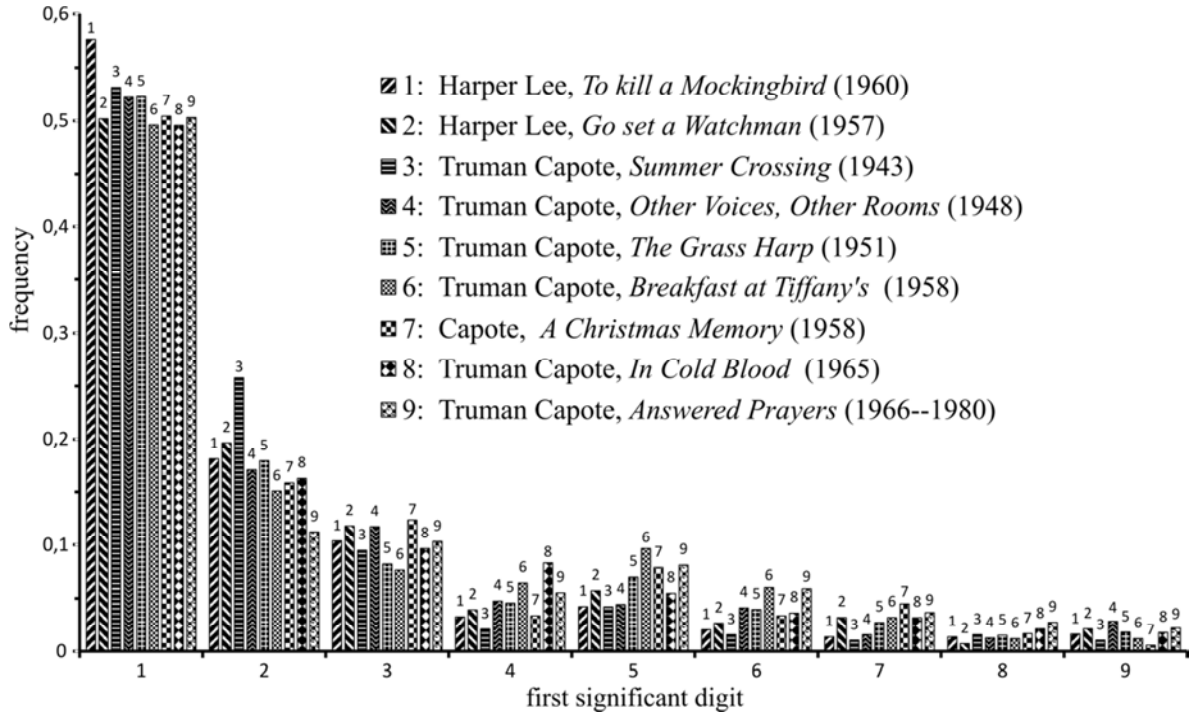

Figure 10. *The distribution of first significant digits of numerals in texts by Harper Lee and Truman Capote.*

We believe that our methodology can be a useful addition to traditional textual practices, taking into account sentence length, word length, occurrence of certain words and parts of speech, etc. [21, 22, 23].

5. Conclusion

Benford's law holds approximately for coherent texts. Deviations from Benford's law are statistically significant author features that allow, under certain conditions (the most important of which is a sufficient length), to distinguish between the texts with a different authorship.

The actual frequency of occurrence usually is higher than the probability according to Benford's law for significant digits 1, 2, 3; for the subsequent digits the situation is reversed. At the end of {1, 2,…, 8, 9} row, the digits distribution is characterized by strong fluctuations and thus is unrepresentative for our purpose.

Of course, the comparison of the distributions cannot be based merely on the detection of their subjective visual similarities/differences. To quantify, we have applied the non-parametric range Mann-Whitney U test and Kruskal-Wallis test as well as the parametric Pearson's chi-squared test. The null hypothesis, which asserts the absence of significant differences in the distributions considered, was rejected and accepted exactly in the cases, as described above, i.e. the visual assessment was *correct*.

Acknowledgements

A. V. Z. is grateful to Dr. William M. Goodman for his valuable comments.

References

[1] F. Benford, “The law of anomalous numbers”. Proceedings of American Philosophical Society. 1938. vol. 78. No. 4. pp. 551–572.

[2] T. P. Hill, “A Statistical Derivation of the Significant-Digit Law”. Statistical Science. 1995. vol. 10. pp. 354–363.

[3] W. M. Goodman, “Reality Checks for a Distributional Assumption: The Case of ‘Benford’s Law’”. JSM 2013 – Business and Economic Statistics Section, pp. 2789–2803.

[4] M. J. Nigrini, Benford’s Law: applications for forensic accounting, auditing, and fraud detection. Hoboken: John Wiley & Sons, 2012.

[5] B. F. Roukemaa, “A first-digit anomaly in the 2009 Iranian presidential election”. Journal of Applied Statistics. 2014. vol. 41. No. 1. pp. 164–199.

[6] D. Biau, “The first-digit frequencies in data of turbulent flows”. Physica A. 2015. vol. 440, pp. 147–154.

[7] T. P. Hill and R. F. Fox, “Hubble’s Law Implies Benford’s Law for Distances to Galaxies”. Journal of Astrophysics and Astronomy. 2016. vol. 37. No. 4. 8 pages.

[8] M. Sambridge, H. Tkalčić, and P. Arroucau, “Benford’s Law of First Digits: from Mathematical Curiosity to Change Detector”. Asia Pacific Mathematics Newsletter. 2011. vol. 1. No. 4. pp. 1–6.

[9] P. Andriotis, G. Oikonomou, and T. Tryfonas, “JPEG steganography detection with Benford’s Law”. Digital Investigation. 2013. vol. 9. No. 3–4. pp. 246–257.

[10] A. D. Alves, H. H. Yanasse, and N. Y. Soma, “Benford’s Law and articles of scientific journals: comparison of JCR and Scopus data”. Scientometrics. 2014. vol. 98. pp. 173–184.

[11] A. V. Zenkov, “Deviation from Benford’s law and identification of author peculiarities in texts”. Computer Research and Modeling, 2015, vol. 7, No. 1, pp. 197–201 (in Russian).

[12] The Best American Humorous Short Stories, by G. P. Morris, E. A. Poe, C. M. S. Kirkland, E. Leslie, G. W. Curtis, E. E. Hale, O. W. Holmes, M. Twain, H. S. Edwards, R. M. Johnston, H. C. Bunner, F. R. Stockton, F. Bret Harte, O. Henry, G. R. Chester, G. MacGowan Cooke, W. J. Lampton, and W. Hastings. The Project Gutenberg eBook, eBook #10947;

[13] The Short-story, by W. Irving, E. A. Poe, N. Hawthorne, F. Bret Harte, R. L. Stevenson, and R. Kipling. The Project Gutenberg eBook, transcribed from the 1916 Allyn and Bacon edition, eBook # 21964.

[14] The Lock And Key Library, Classic Mystery And Detective Stories, by R. Kipling, A. Conan Doyle, E. Castle, S. J. Weyman, W. Collins, and R. L. Stevenson. The Project Gutenberg eBook, transcribed from the 1909 Review of Reviews Co. edition, eBook # 2038.

[15] Shorter Novels, Eighteenth Century. The History of Rasselas, The Castle of Otranto, Vathek, by S. Johnson, H. Walpole, and W. Beckford. The Project Gutenberg eBook, transcribed from the 1903 Aldine House edition, eBook # 34766.

[16] The Best of the World's Classics, Vol. V – Great Britain and Ireland, by J. Boswell, W. Wordsworth, W. Scott, S. T. Coleridge, R. Southey, W. S. Landor, C. Lamb, W. Hazlitt, T. De Quincey, Lord Byron, P. Bysshe Shelley, G. Grote, T. Carlyle, Lord Macaulay. The Project Gutenberg eBook, transcribed from the 1909 Funk & Wagnalls Co. edition, eBook # 22182.

[17] The Great English Short-Story Writers, Vol. 1, by D. Defoe, J. Hogg, W. Irving, N. Hawthorne, E. A. Poe, J. Brown, C. Dickens, F. R. Stockton, M. Twain, F. Bret Harte, T. Hardy, H. James, and R. L. Stevenson. The Project Gutenberg eBook, transcribed from the 1910 Readers's Library edition, eBook # 10135.

[18] A House to Let, by C. Dickens, W. Collins, E. Gaskell, and A. A. Procter. The Project Gutenberg eBook, transcribed from the 1903 Chapman and Hall edition, eBook #2324.

[19] Masterpieces of Mystery, Vol. 1, Ghost Stories, by A. Blackwood, M. R. James, K. Rickford, W. F. Harvey, R. A. Cram, R. L. Stevenson, and W. D. Steele. The Project Gutenberg eBook, transcribed from the 1920 Doubleday, Page & Co. edition, eBook # 27722.

[20] J. N. Binongo, “Who wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution”. Chance. 2003. vol. 16. No. 2, pp. 9–17.

[21] The Oxford Handbook of Computational Linguistics (Ed. R. Mitkov). Oxford (a.o.): Oxford University Press, 2003.

[22] The Handbook of Linguistics (Eds. M. Aronoff and J. Rees-Miller). Oxford (a.o.): Blackwell Publishing, 2004.

[23] B. Ryabko, J. Astola, and M. Malyutov, Compression-Based Methods of Statistical Analysis and Prediction of Time Series. Springer International Publishing Switzerland, 2016.