International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 5, 2020, PP 14-20
ISSN No. (Online) 2347-3142
DOI: https://doi.org/10.20431/2347-3142.0805003
www.arcjournals.org

Statistics of Numerals in Authorial Texts and Stylometry

Zenkov A. V.*

Ural Federal University, Ekaterinburg, Russia

Corresponding Author:** ***Zenkov A. V., *Ural Federal University, Ekaterinburg, Russia*

Abstract: *Two approaches to the statistical analysis of texts are suggested, both based on the study of numerals occurring in coherent literary texts. The first approach is related to the study of 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 discourse analysis.

The proposed approaches are illustrated by examples of computer analysis of the literary texts by L. Dobychin, A. Platonov, I. Ilf and E. Petrov, V. Kataev, and M. Bulgakov. Frequency distributions of the leading digits of numerals in Dobychin's and Platonov's texts are shown, which differ markedly in appearance and confirm the significant stylistic originality of the texts of the two authors. For all of the above authors, we obtained the frequency distributions of the numerals found in the texts. The hypothesis that Ilf and Petrov are fake authors of the novels "The Twelve Chairs" and "The Little Golden Calf", and they were ghosted by Bulgakov, is investigated. The frequency distribution of numerals, as well as its cluster analysis, do not confirm this hypothesis.

Keywords: *Quantitative linguistics, stylometry, attribution of texts, text authorship, numerals in texts, leading digit, novel "The Twelve Chairs", novel "The Little Golden Calf".*

1. Introduction

The problems of this research relate to stylometry (statistical study of texts to find individual features of the author's style – in particular, for attribution of texts). The conventional methods used – taking into account the frequency of occurrence of certain words and collocations in the text, the average length of words and sentences, etc. [1] – often lead to contradictory results, and the very abundance of methods indicates a lack of reliability of each of them individually. In this case, the emergence of new stylometric techniques is not redundant, and they are all complementary rather than mutually exclusive.

We have proposed the idea of studying numerals found in the text as a means of characterizing the author's style [2, 3]. The analysis of numerals has many advantages. The results of this analysis allow direct linguistic interpretation (unlike, for example, the neural network method [1], which is able to successfully recognize the authorship of texts, but the recognition procedure is a *black box*). The use of numerals in the text is directly related to its authorship, style, and genre (see below).

The approach to stylometry problems we are developing has two varieties. First, we studied the frequency distribution of the leading digits of numerals. The idea may seem bizarre, but it is in line with research related to Benford's Law [4] – a mysterious, not fully understood manifestation of the Law of large numbers, according to which in large arrays of numerical data describing various objects and phenomena, numbers starting with digit 1 (their share according to Benford's Law is 30.1 per cent) are more common than those starting with digit 2, and the latter, in turn, are more common than numbers starting with digit 3, and so on. According to our research, the leading digits of numerals in coherent texts are distributed even more unevenly than prescribed by Benford's Law: the proportion of numerals starting with 1 can reach 50 per cent. Usually, the frequency distribution of the leading digits of numerals is characteristic of each author and appears in all (large enough) of his works. Sometimes this allows to check the authorship of texts: if the distributions of the leading digits significantly differ for two texts, then the same authorship of the texts is doubtful.

The second variation of our stylometric method consists in analyzing the numerals occurring in the text (and not their leading digits). The frequency distribution of numerals is also, to a large extent, specific for the author [3]. The first of the two approaches can be considered a convolution of the second. Each approach has its own advantages and disadvantages.

Counting the leading digits makes sense only for significant digits 1, 2, and, possibly 3, since the occurrence of subsequent digits is subject to strong fluctuations even in the texts of the same author (see Fig. 1). Thus, only a small part of the statistical information about the numerals contained in the text is available for analysis. In addition, a problem arises with texts in languages in which the numeral *one* is formally indistinguishable from the indefinite article (although this is surmountable by switching to an intermediary language without this problem). On the other hand, the information here is presented in a generalized form, which allows to average specific features of individual works of the author.

Analysis of the use of the numerals themselves (not the leading digits) provides richer information about the author's features of the text and, to a large extent, is devoid of indistinguishability of the numeral *one* and the indefinite article. However, analyzing the statistics of numerals is technically more difficult. This article is devoted to the possibilities and comparison of both types of our method. Along the way, we consider one problem related to the Russian literature of the 20th century.

2. Object of Research

Literary texts by L. Dobychin and A. Platonov are notable for distinct stylistic originality in the Russian fiction, they have common literary sources and analogues in foreign literature [5]. We will show how this affects the statistics of the use of numerals in their texts.

The literary work of I. Ilf and E. Petrov has repeatedly become the subject of discussion. The novels *The Twelve Chairs* and *The Little Golden Calf* are full of literary allusions; thematically and stylistically they are related to the texts by V. Kataev, M. Bulgakov, Yu. Olesha, and others [6]. There is nothing comparable to these two works in the literary heritage of Ilf and Petrov. According to the radical point of view [7], Ilf and Petrov are the fake authors of *The Twelve Chairs* and *The Little Golden Calf*, and they were ghosted by Bulgakov. In this paper, we will apply our methodology to a comparative analysis of the corpus of literary texts by Ilf and Petrov. Along the way, we study Kataev's *The Lord of Iron* (1924) and *The Embezzlers* (1926), contemporary to *The Twelve Chairs* (1928) and *The Little Golden Calf* (1931), as well as Bulgakov's *The Master and Margarita*.

3. Method of Research

We have prepared a computer program that searches in the text for numerals expressed both in numbers and verbally. The texts analyzed were pre-cleaned of numerals that do not reflect the author's creative intent (numbering of pages, chapters, etc.) or accidentally included in idioms (*to the four winds*). Since the analyzed texts have different sizes, correction coefficients were used to equalize the results on the occurrence of numerals.

The numerals extracted from the text were displayed on frequency graphs, which directly allowed us to draw conclusions about the author's style. Information about numerals found in texts was also systematized using the hierarchical cluster analysis. The farthest neighbour clustering was used (which exaggerates differences but provides clearly defined clusters). The smaller the difference in the occurrence of the same numbers in two texts, the greater the similarity (the smaller the "distance" ρ) between these texts, so the Manhattan metric was used

$$\rho(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}|x_i-y_i|, \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are n-dimensional vectors whose components are the absolute frequency of occurrence of the first n natural numbers found in both analyzed texts.

4. Results and Discussion

Fig1. shows the frequency distribution of the leading digits of numerals in the most voluminous works by Dobychin and Platonov.

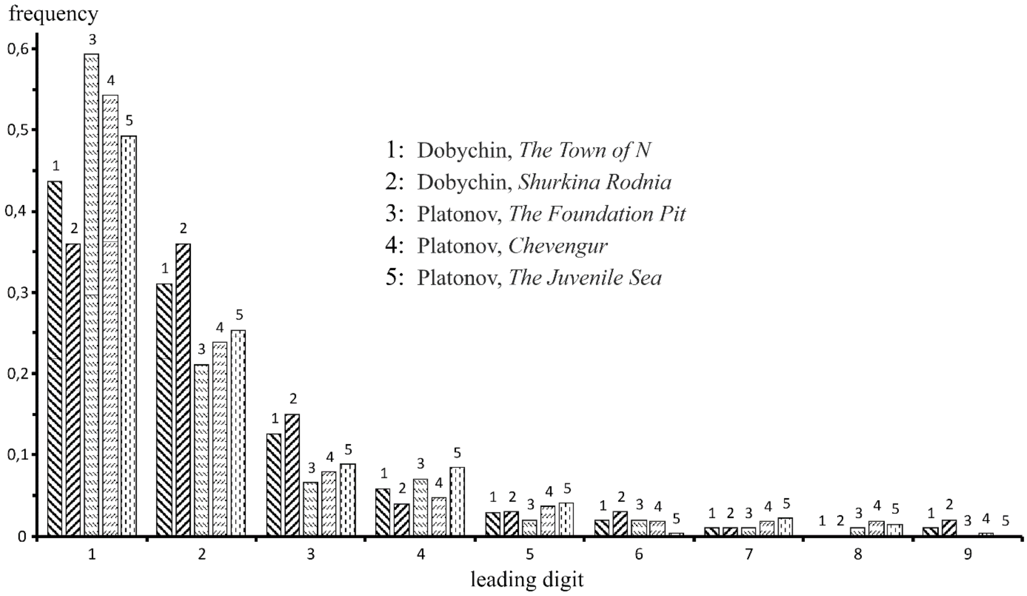

Fig1. *Frequency distribution of the leading digits {1, 2, ..., 9} of numerals in texts by Dobychin and Platonov*

The leading digits 1, 2, and 3 appear in texts by Dobychin, on the one hand, and Platonov, on the other hand, in very different manner. The visual distinction is supported by Pearson's statistical test [2, 3]. Thus, the analysis of the distribution of the leading digits indicates certain stylistic differences in texts of the two authors.

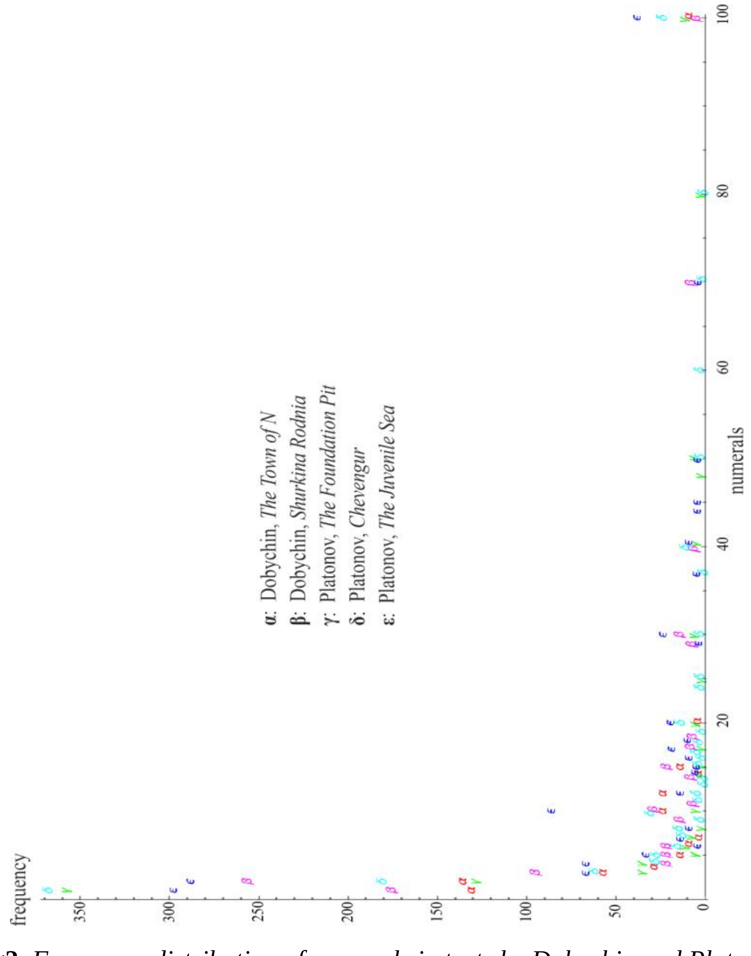

Fig2. *Frequency distribution of numerals in texts by Dobychin and Platonov*

The results of using an extended statistical method that analyzes the occurrence of numerals themselves are much richer. Fig. 2 shows the frequency of occurrence of numerals from the range [0, 100] in the same texts by Dobychin and Platonov.

Some results:

1. Platonov's texts tend to use numerals more often than Dobychin's texts.

2. Platonov less often resorts to rounding numerals (10, 20, 30, ...), which, together with item 1, may indirectly indicate a greater tendency to detail.

3. the numeral *one* (in various word forms) is the undisputed leader among the numerals found in Platonov's texts [1]. But in Dobychin's texts, the numeral *one* is inferior in frequency to the numeral *two*!

4. Note the psychologically understandable rarefaction of the numerals sequence and a decrease in their occurrence as they increase, as well as a noticeable local maximum on the numeral *hundred*, which, of course, plays here the role of an indefinitely large number.

Fig. 3 shows the frequency distribution of numerals in Ilf and Petrov's *The Twelve Chairs* and *The Little Golden Calf*, as well as in Bulgakov's *The Master and Margarita*, and Kataev's *The Embezzlers* and *The Lord of Iron*. For clarity, we restrict the graph to the range [1; 50] on the horizontal axis; the conclusions formulated below are valid for the entire set of numerals found.

Some conclusions directly following from the figure:

1. For all the analyzed texts, there are peaks in the occurrence of "round" numbers 10, 20, ... , 100, 200, …

2. In the texts by Ilf and Petrov, as well as in Bulgakov's *The Master and Margarita*, the numeral 1 has the highest frequency (which is consistent with Benford's Law), but in Kataev’s texts the number 2 leads.

3. Between *The Twelve Chairs* and *The Little Golden Calf*, there is a conspicuous similarity in the numerals frequency (we will return to this later – see the conclusions to Figs. 4 and 5).

4. These two texts are characterized by the greatest variety of numerals.

5. On the contrary, Kataev's texts are distinguished by the least variety of numerals.

6. In terms of the variety of numerals, *The Master and Margarita* occupy an average position, but the frequencies of the numerals (after the initial frequent *ones* and *twos*) are usually lower than in other texts analyzed. In fact, many numbers occur once.

Based on the frequency distributions of numerals in the works by Ilf and Petrov included in their 5-volume collection of works [8], we performed clustering and built a dendrogram (Fig. 4). It turned out that all the analyzed texts contained natural numbers in the range [1; 12]; the frequencies of these numbers were used for clustering; $n = 12$ in the formula (1). The numbers to the left of the dendrogram refer to the following texts:

1) *The Twelve Chairs*; a joint work by Ilf and Petrov, 1927-28; vol. 1 [8],

2) Joint works 1932-37 (stories, feuilletons, articles, speeches, vaudevilles, screenplays) by Ilf and Petrov, included in vol. 3,

3) *The Little Golden Calf*; a joint work by Ilf and Petrov, 1929-30; vol. 2,

4) Works (stories, essays, feuilletons) written individually by Petrov in 1924-32 and included in vol. 5,

5) Works (essays, articles, memoirs) written individually by Petrov in 1937-42 and included in vol. 5,

6) *One-storied America* (travel essays; sometimes translated as *Little Golden America*), 1936, vol. 4,

7) Works (stories, essays, feuilletons) written solely by Ilf in 1923-29, as well as his notebooks from 1925-37, included in vol. 5.

From the dendrogram, it is clear that the closest (in terms of the use of numerals) are Nos. 1 and 2 – joint, mainly literary texts; to this initial cluster, No. 3 is soon added with the same characteristic. Nos.

[1] This is a common property of most texts by different authors; for the interpretation we may refer to both psychological factors and the technical indistinguishability of the actual numeral *one* and the same word acting as an indefinite article in Russian; cf.: «*V odin prekrasny den’*» (*One Fine Day*).

5 and 6 – late non-fiction works – form the next cluster, internally not as unified as the cluster {1, 2, 3}. At the last stage, No. 7 is added to the cluster – texts by Ilf.

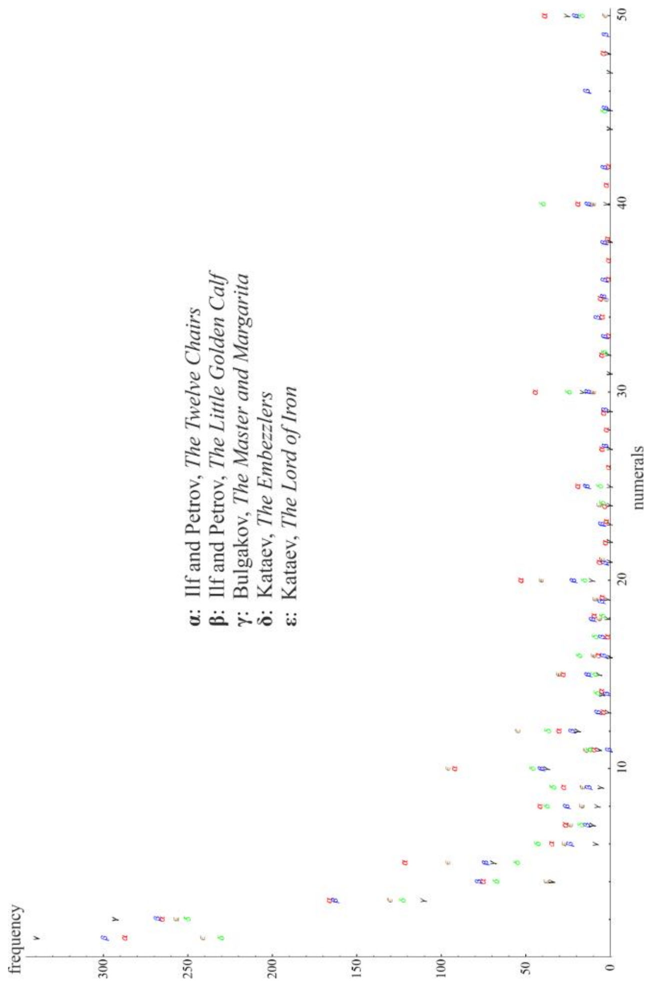

Fig3. *Frequency distribution of numerals in texts by Ilf and Petrov, Bulgakov, and Kataev*

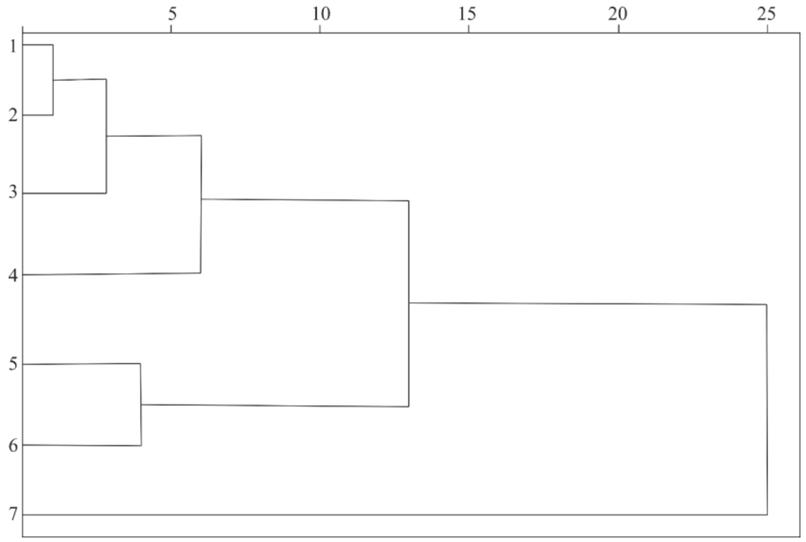

Fig 4. Results of hierarchical cluster analysis based on the occurrence of numerals in the texts by Ilf and Petrov. The horizontal scale indicates the “distance” between clusters in conventional units. Texts Nos. 1–7, combined into clusters, are indicated in the article

So, from Fig. 4 it follows that the analysis of the use of numerals in texts can distinguish between genres and “feel” the authorship of texts.

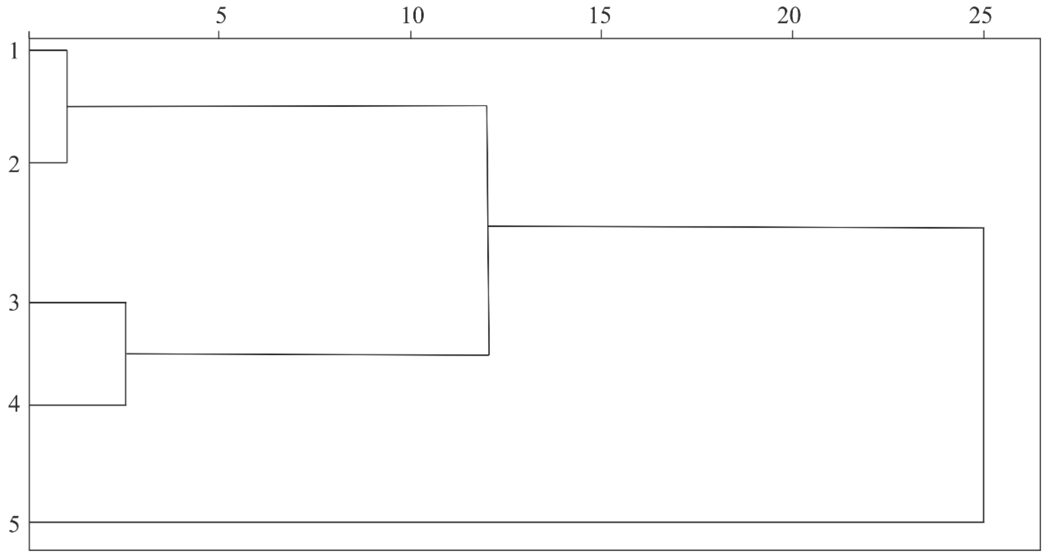

Fig. 5. The results of hierarchical cluster analysis based on the occurrence of numerals in texts by Ilf and Petrov (Nos. 1, 2), Kataev (Nos. 3, 4), and Bulgakov (No. 5). The horizontal scale indicates the "distance" between clusters in conventional units. Texts Nos. 1-5, combined into clusters, are indicated in the article

Fig. 5 shows the results of hierarchical cluster analysis of the occurrence of numerals (again from the range [1; 12], as in Fig. 4) in the novels *The Twelve Chairs* (No. 1) and *The Little Golden Calf* (No. 2) by Ilf and Petrov, as well as in the works of writers who were named as possible true authors of these two novels – *The Embezzlers* (No. 3) and *The Lord of Iron* (No. 4) by Kataev, and Bulgakov's *The Master and Margarita* (No. 5). Clustering took place in accordance with the generally accepted authorship of the texts. The distance between clusters {1, 2} and {3, 4}, not to mention the height of fusion with No. 5 – is so great that it casts doubt on the hypothesis that Bulgakov or Kataev wrote *The Twelve Chairs* and *The Little Golden Calf*. Of course, taken separately, the results of Fig. 5 do not confirm the authorship of Ilf and Petrov themselves, but the results of Fig. 4 indirectly indicate their authorship. Note that in the article [9], based of formal accounting of statistics of service words in texts of *The Twelve Chairs* and *The Master and Margarita*, the hypothesis [7] that Bulgakov is the author of *The Twelve Chairs* is also questioned.

So, the analysis of the use of numerals in texts can be used to test the authorship of texts.

5. Conclusion

The analysis shows that taking into account the occurrence of numerals in literary texts can provide information about the author's, 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 the methodology we have developed can be a useful addition to the traditional stylometric practices of taking into account the length of sentences and words, the frequency of use of service words and certain significant parts of speech, etc.

Acknowledgement

This work was supported by a grant from the Russian Foundation for Basic Research, project No. 19-012-00199A, "A New Method of Text Attribution Based on Statistics of Numerals".

References

[1] Tempestt N., Kalaivani S., Aneez F., Yiming Y., Yingfei X. and Damon W., Surveying Stylometry Techniques and Applications, ACM Comput. Surv. 50(6), Article 86, 36 pages (2017), https://doi.org/10.1145/3132039.

[2] Zenkov A.V., A Method of Text Attribution Based on the Statistics of Numerals, Journal of Quantitative Linguistics. 25(3), 256 (2018).

[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. 46, 12 (2019).

[4] Benford F., The law of anomalous numbers, Proceedings of American Philosophical Society. 78(4), 551 (1938).

[5] Eidinova V.V., A. Platonov and L. Dobychin: Stylistic Convergence and Repulsion. Andrei Platonov’s “Land of Philosophers”: Problems of Creativity. Anniversary Issue 5: Based on the materials of the International Scientific Conf. dedicated to the 50th anniversary of the death of A.P. Platonov. April 23–25, 2001. Moscow. 2003. p. 211–219.

[6] Ščeglov Yu.K., The Novels by Ilf and Petrov. Readers’s Companion. Saint Petersburg, Ivan Limbach Publishing House, 2009. 656 p.

[7] Amlinski I., 12 Chairs from Mikhail Bulgakov. Berlin, Kirschner Verlag, 2013. 328 p. ISBN: 978-3-00043-284-2

[8] Ilf I., Petrov E., Collected works in five volumes. Moscow, Khudozhestvennaia Literatura, 1961.

[9] Mironova L.B., Vereshchagina O.V., To the question of the authorship of the novel “The Twelve Chairs», Baltic Humanitarian Journal. 8(1), 108 (2019).

Citation: *Zenkov A. V, Statistics of Numerals in Authorial Texts and Stylometry, International Journal of Scientific and Innovative Mathematical Research (IJSIMR), vol. 8, no. 5, pp. 14-20, 2020. Available : DOI: https://doi.org/10.20431/2347-3142.0805003*