Litera
Правильная ссылка на статью:
Zenkov A.V. — Under a False Flag: Literary Hoaxes and the Use of Numerals // Litera. – 2023. – № 10. DOI: 10.25136/2409-8698.2023.10.68743 EDN: TYDRFD URL: https://nbpublish.com/library_read_article.php?id=68743

Under a False Flag: Literary Hoaxes and the Use of Numerals / Под чужим флагом: литературные мистификации и использование числительных

Зенков Андрей Вячеславович

ORCID: 0000-0002-1233-9082

кандидат физико-математических наук

доцент кафедры Моделирования управляемых систем, Уральский федеральный университет

620002, Россия, Свердловская область, г. Екатеринбург, ул. Мира, 19, оф. 434

zenkow@mail.ru

Статья из рубрики "Авторская маска"

DOI:

10.25136/2409-8698.2023.10.68743

EDN:

TYDRFD

Дата направления статьи в редакцию:

16-10-2023

Дата публикации:

23-10-2023

Аннотация: Настоящее исследование относится к стилометрии. Имеются случаи, когда писатель, добившийся известности, по разным причинам начинает творить под другим именем, пытается писать в другой манере и, порой, снова добивается успеха в новом воплощении. Способен ли автор существенно изменить присущий ему литературный стиль или невозможно уйти от самого себя – изучению этого вопроса посвящена наша работа. Предметом исследования является авторский литературный стиль произведений, подписанных подлинным именем автора и подписанных псевдонимом. Методом исследования является анализ встречаемости числительных в текстах того или иного автора. В настоящей работе продемонстрировано на ряде примеров из англо-, франко- и русскоязычной литературы, что использование числительных является авторской

особенностью, которая проявляется во всех или большинстве достаточно длинных текстов данного автора. Полученные результаты показали, что вопреки попыткам автора писать «по-новому» манера использования числительных консервативна и позволяет распознавать фиктивное авторство. Этот вывод сделан на основании анализа произведений Р. Гари и Б. Акунина (Г. Чхартишвили), известных своими литературными мистификациями. Анализ употребления числительных применён также к проблеме авторства романа «Убить пересмешника» Харпер Ли. Выводы о схожести/различии литературных стилей сделаны на основе иерархического кластерного анализа и подкреплены критерием Пирсона. Научная новизна работы состоит в новом подходе к поиску авторского инварианта и атрибуции текстов.

Ключевые слова:

стилометрия, квантитативная лингвистика, атрибуция текстов, авторство текстов, числительные в тексте, Ромен Гари, Борис Акунин, Харпер Ли, Трумен Капоте, кластерный анализ

The research was supported by the grant No. 23-28-00750 from the Russian Science Foundation; see https://rscf.ru/en/project/23-28-00750/.
Исследование выполнено за счет средств гранта Российского научного фонда № 23-28-00750, https://rscf.ru/project/23-28-00750/, проект «Разработка нового метода стилометрии на основе статистики использования числительных в авторских текстах».

1. Introduction

In stylometry, there is an actual, not yet fully resolved problem of finding an author's invariant (fingerprint) – a quantitative feature (or set of features), the value of which is approximately constant and individual for all (or most) of the texts of a given author. The author invariant would be useful, in particular, in problems of determining the authorship of texts: were these texts written by the same author? Are they written by this particular author? Who among the supposed authors is most likely the author of this text? Etc. Of course, the answers to the questions posed are always probabilistic; Type I and II errors are inevitable.

Unfortunately, in stylometry, with all the abundance of quantitative methods proposed, there is still no one that would not give an obviously absurd result on some test example. Traditional practices in stylometry include finding the average length of words and sentences, the frequencies of certain content and/or function words, the frequencies of *n*-grams, etc. [1, 2]. Then the obtained numerical data are processed within the framework of some computational apparatus from probability theory and mathematical statistics to cybernetics and information theory. It would seem that the joint use of several methods should increase the reliability of the results obtained, but, alas, these results often contradict each other.

The use of artificial intelligence inspires great hopes [3, 4], but the problem is the opacity of neural networks and the difficulty of interpreting the results.

We have developed a new approach to stylometry problems, based on the analysis of the use of numerals in the (literary) author's text [5–11]. This approach has many advantages over traditional ones. Firstly, due to the very nature of numerals, they are easily

quantifiable. Secondly, the results obtained allow a transparent philological interpretation. Thirdly, the occurrence of numerals in the text is practically invariant with respect to the translation of the text into another language (see this below, in Section 2).

Like any statistical method of stylometry, the method of taking into account numerals requires a fairly large text length (files ranging in size from tens of kB in UTF-8 encoding).

It turned out that the manner of using numerals is largely individual for each writer, i.e. it is the *author's invariant*. This can be explained by the author's psychology, which influences the creative result regardless of his conscious intention.

In the history of literature, there are examples when the author wrote under different pen names, and sometimes these literary hoaxes turned out to be successful. In the examples that we will consider below, it was not just about changing the name, but about trying to write "differently". The question arises: can the conscious intention of the author to change his literary style affect the use of numerals in the text?

Our work is devoted to considering this issue. It is constructed as follows.

After the description of the research methodology (Section 2), there follows a comparative analysis of literary texts by different authors, demonstrating the constancy of the author's features of the occurrence of numerals in texts. This is shown by the examples of English-, French- and Russian-language texts (Section 3).

In Section 4, our stylometric technique is applied to the analysis of literary texts by Romain Gary and Boris Akunin, known for their literary hoaxes, experiments with style, and publications under several *noms de plume*. We then examine the issue of authorship of Harper Lee's literary texts, which Truman Capote is suspected of having a significant influence on.

The work ends with a discussion of the results and conclusions.

The Appendix contains some computational issues.

2. Subject and Method of Research

We have developed a computer program that searches for cardinal as well as ordinal numerals expressed both in numbers and (considerably more often) verbally (in different word forms) in the English-, French- and Russian-language texts.

Numerals not related to the author's creative idea were deleted from the text beforehand – such as idiomatic expressions and set phrases accidentally containing numerals (for example, *seventh heaven* and *fifty-fifty* in English), page and chapter numbering, itemizations 1), 2), 3), ..., etc. As for itemizations, they are analogous to page numbering. Not to mention the fact that they are not always set just by the author (it may depend on editorial corrections), they are merely the usual system of markings rather than the author's intention. Anyway, the deleted items (owing to their rare occurrence) have a negligible influence on results obtained.

We have taken into account the numerals written in whatever form (e.g., *five men*, *5 men*, *The Fifth Element*, but not *The Fifth republic* – the latter is a set phrase).

Multiplicative (adverbial) numbers (*once*, *twice*, …), multipliers (*single*, *double*, …), distributive numbers (*singly*, *doubly*, …), collective numbers which describe sets, such as *pair* or *dozen* in English have been excluded. As for fractional numerals (*two fifths*, *seven*

tenths, ...), we separately took into account numerators and denominators, as if they were independent numerals.

It would seem that the method of taking into account numerals encounters an insurmountable obstacle in languages in which the numeral *one* is formally indistinguishable from the indefinite article (*ein* in German, *un* in French, etc.). But the set of numerals found in the text is perhaps the only feature that is almost completely preserved when translated into another language (note that the idioms can be translated into other languages without numerals). This allows, if necessary (the text in a language in which such a coincidence takes place, or the unavailability of the text in the original language), to analyze the author's style, resorting to translation into an intermediary language.

In previous works [8-11], we have already demonstrated by numerous examples that the use of numerals in texts is specific to each author, depends on the artistic direction, genre and style. Of course, no empirical verification, no matter how extensive, that a new method really works, can be considered conclusive, and there will always be skeptical voices claiming that the evidence collected is insufficient. Therefore, in this paper we will present new confirmations of the conclusions about the author's use of numerals and apply this methodology to answer the question posed in the Introduction.

The following works in the original language were subjected to comparative analysis from the point of view of the occurrence of numerals (see Section 3; the works are listed in the order in which they lined up on the dendrograms (Figs. 1–4); among the selection criteria were the large size of the works and free access to them on the Internet):

· English-language works

1. Charles Dickens: *Our Mutual Friend*; *Little Dorrit*; *David Copperfield*; *Dombey and Son*;

2. W. M. Thackeray: *The History of Henry Esmond*; *The History of Pendennis*; *The Memoires of Barry Lyndon*; *Vanity Fair*;

3. H. G. Wells: *The War of the Worlds*; *The Island of Doctor Moreau*; *The Invisible Man*; *The Time machine*;

4. V. V. Nabokov (works written in English): *Pale fire*; *Ada, or ardor*; *Look at the Harlequins!*; *The real life of Sebastian Knight*; *Transparent things*; *Bend sinister*;

· French-language works

1. M. Proust: *A la recherche du temps perdu* – the whole heptalogy: *Du cфté de chez Swann*; *A l'ombre des jeunes filles en fleurs*; *Le cфté de Guermantes*; *Sodome et Gomorrhe*; *La prisonnière*; *Albertine disparue*; *Le Temps retrouvé*;

2. Émile Zola: *Germinal*; *La débâcle*; *Le Ventre de Paris*; *L'assommoir*; *La faute de l'abbé Mouret*; *La fortune des Rougon*; *L'argent*; *Au bonheur des Dames*; *La terre*; *Le rêve*; *Le docteur Pascal*; *La Joie de vivre*; *Une page d'amour*; *La curée*; *Son Excellence Eugène Rougon*; *Nana*;

3. Guy de Maupassant: *Bel ami*; *Pierre et Jean*; *Une vie*; *Fort comme la mort*; *Notre cœur*;

4. F. Mauriac: *Thérèse Desqueyroux*; *Le Nœud de vipères*;

5. A. Daudet: *L'Immortel*; *Le petit chose*;

6. A. Gide: *La porte étroite*; *Les cahiers d'André Walter*; *L'école des femmes*; *Geneviève*; *Les faux-monnayeurs*;

7. J. Verne: *Un capitaine de quinze ans*; *Le tour du monde en quatre-vingts jours*;

· Russian-language works

1. F. M. Dostoevsky: *The Idiot*; *Crime and Punishment*; *The Brothers Karamazov*; *The Adolescent*; *Humiliated and Insulted*; *Demons*; *The House of the Dead*;

2. I. A. Goncharov: *Oblomov*; *The Same Old Story*;

3. A. I. Herzen: *Who is to Blame?*; *My Past and Thoughts*;

4. N. S. Leskov: *A Decayed Family*; *Lady Macbeth of Mtsensk*; *The Enchanted Wanderer*;

5. I. S. Turgenev: *On the Eve*; *Virgin Soil*; *Home of the Gentry*; *Fathers and Sons*.

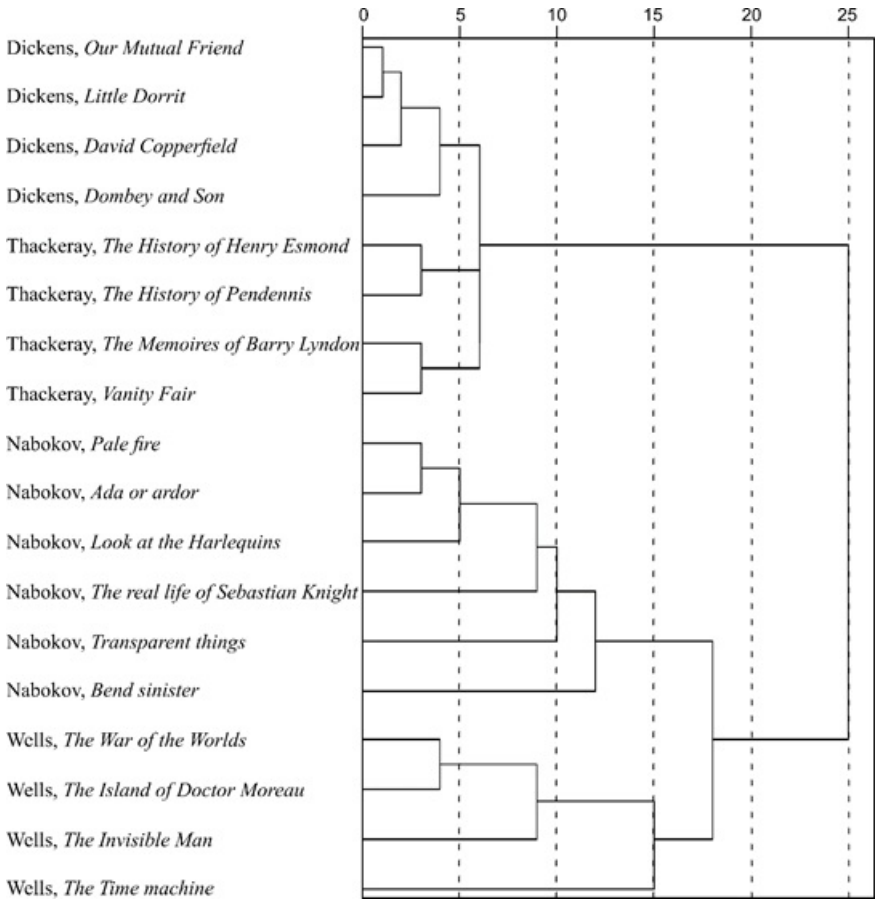

Figure 1. The result of applying hierarchical cluster analysis to literary texts by C. Dickens, W.M. Thackeray, V.V. Nabokov and H.G. Wells. The horizontal axis shows the "distance" between texts in arbitrary units

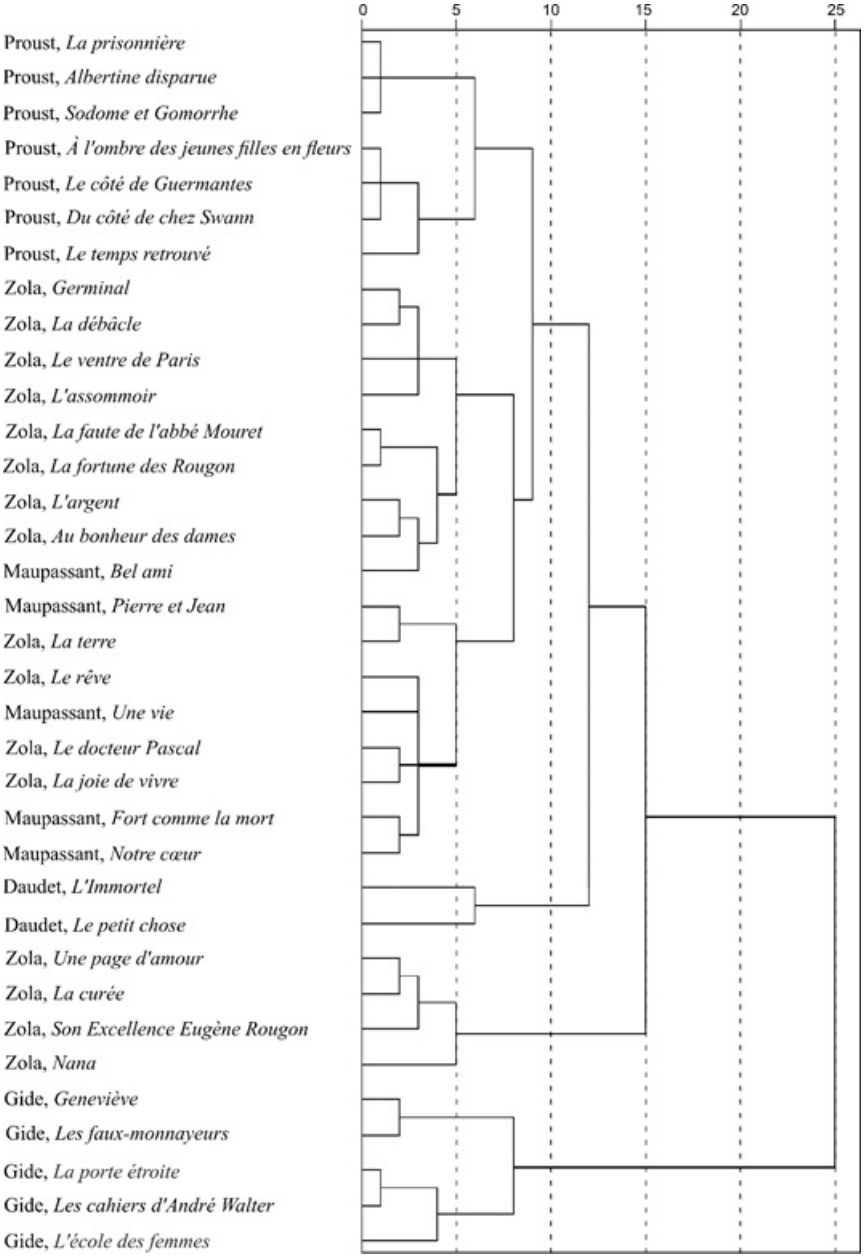

Figure 2. The result of applying hierarchical cluster analysis to French-language literary texts. The horizontal axis shows the "distance" between texts in arbitrary units

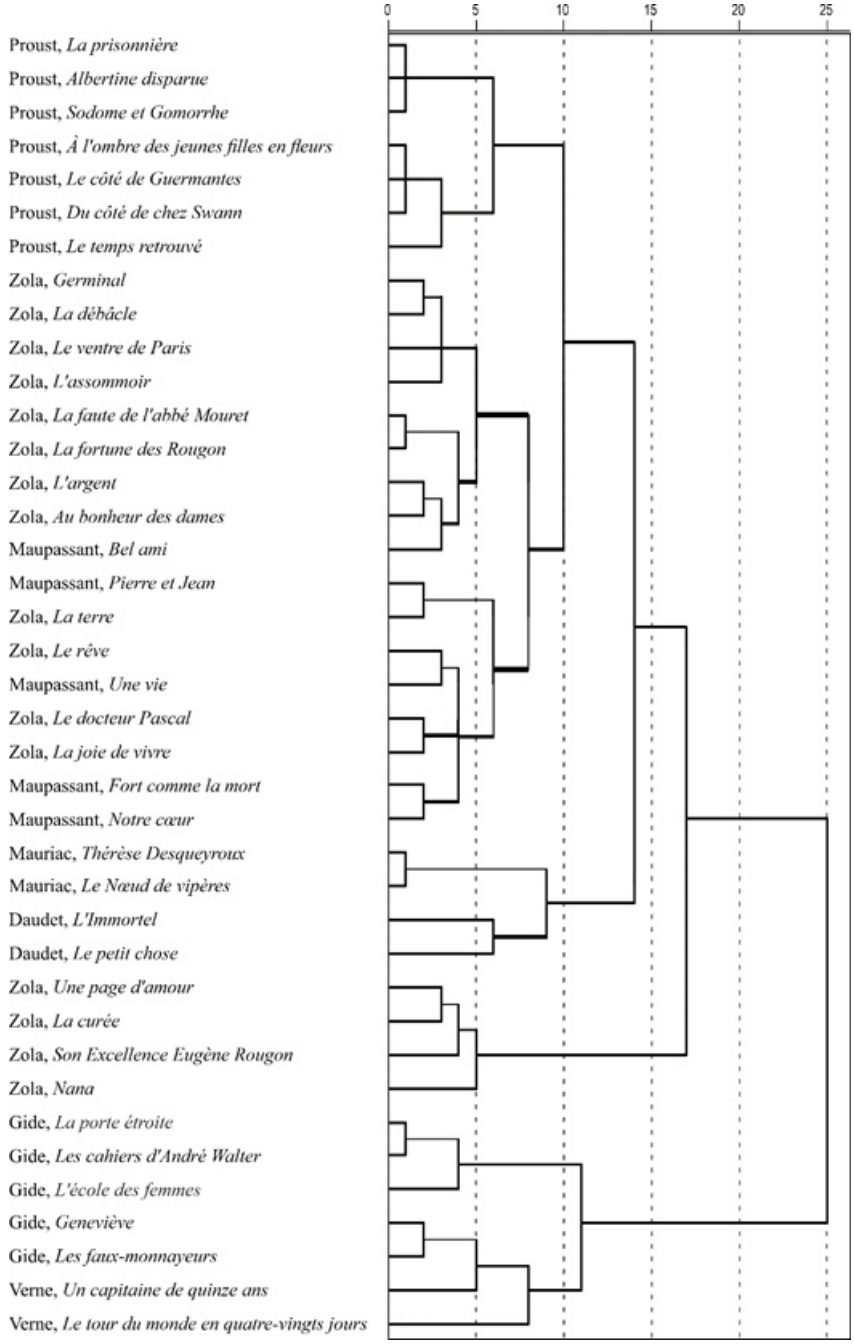

Figure 3. The result of applying hierarchical cluster analysis to French–language texts, with the addition of texts by other authors (*impostors*) – F. Mauriac and J. Verne

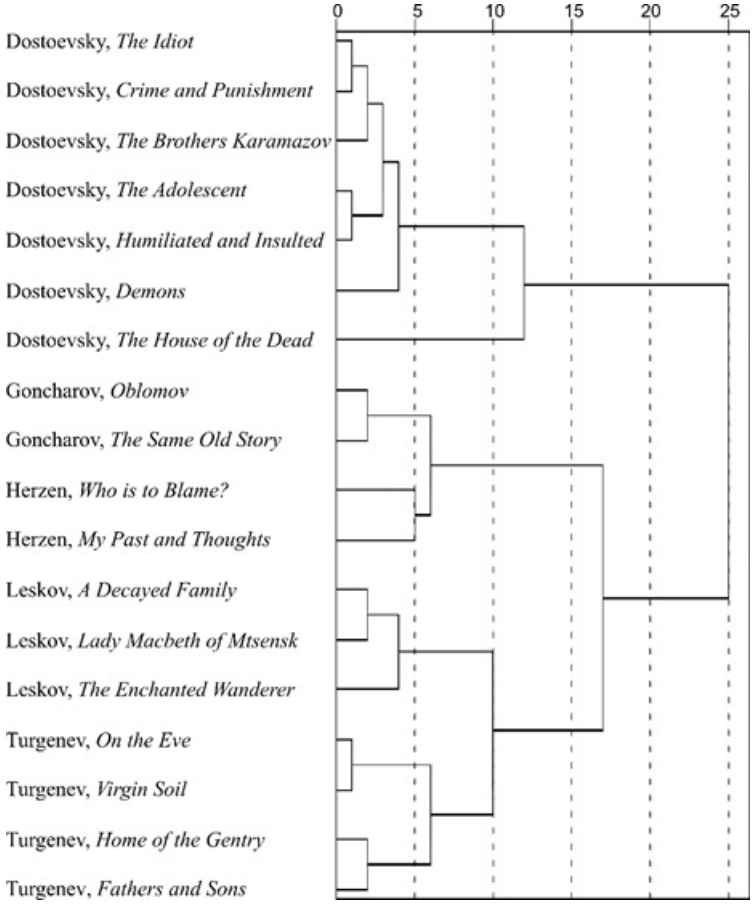

Figure 4. Results of hierarchical clustering of the works by Dostoevsky, Goncharov, Herzen, Leskov, and Turgenev

The analysis of the texts was carried out as follows. Using a computer program, numerals were extracted from the texts, and for each text a summary of the detected numerals and their absolute frequencies was generated. Since the texts differ in volume, for the comparability of absolute frequencies in different texts, the volume of one of them was chosen as a reference, and the corrected absolute frequencies were obtained by multiplying the absolute frequencies by a correction factor. To identify the internal structure in the array of corrected absolute frequencies, hierarchical cluster analysis was used, combining objects into clusters based on their similarity. Its measure is the metric ρ ("distance"): the smaller the "distance" between objects, the greater the similarity between them.

Depending on the nature of the data, different metrics are used in cluster analysis, such as Euclidean

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_i^n (x_i - y_i)^2}, \quad (1)$$

and the Manhattan metric (*aka* City Block distance)

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \sum_i^n |x_i - y_i|, \quad (2)$$

where in our case $\boldsymbol{x}$ and $\boldsymbol{y}$ are n-dimensional vectors, the components of which are the corrected absolute frequencies of the first n natural numbers in the two analyzed texts.

Each subsequent numeral occurs in texts, generally speaking, with ever decreasing frequency (see Section 4 and Appendix), therefore the presence of a square in formula (1) means that the "distance" between texts is, in fact, determined by differences in

frequencies of only the numeral *one* – they make an overwhelming contribution to the sum. We applied the Manhattan metric (2), which more evenly takes into account differences between texts in the frequencies of not only the numeral *one*, but also 2, 3, …, *n*.

In the clustering process, the Average Linkage method was used. It is the golden mean between the Single Linkage and Complete Linkage methods, which, respectively, exaggerate and underestimate the similarity between objects [12].

In quantitative linguistics, it is generally accepted that even when comparing the texts of two authors, only an analysis in which extraneous texts of other authors (so-called *impostors*) are added to the texts being studied will have evidentiary force about their similarity [13]. We took this requirement into account in our analysis.

3. The Manner of Using Numerals is the Author's Style Feature

We will confirm this thesis on the examples of English-, French- and Russian-language texts.

A. English-language texts

Figure 1 shows the results of clustering data on the occurrence of numerals in the texts of the English-language authors listed above. The approaches formulated at the end of Section 1 were applied; $n = 7$ was taken in formula (2), since in all the studied texts there were numerals from *one* to *seven*.

The works were distributed on the dendrogram in full accordance with the authorship. Dickens's literary style is characterized by the most uniform use of numerals (the height of the amalgamation is low). The greatest differences are between Wells's works. It is interesting to note that the two superclusters (Dickens–Thackeray and Nabokov–Wells) generally correspond to the chronological division into literature of the 19th and 20th centuries.

B. French-language texts

Figure 2 shows a dendrogram for French-language texts within the framework of the above approaches; $n = 8$ was taken in formula (2), since in all the studied texts there were numerals from *one* to *eight*.

Again, we can state the distribution of works on the dendrogram in accordance with authorship. The only two authors whose works are not completely localized on the dendrogram are Zola and Maupassant. In the critical literature, comments have been made repeatedly about the similarity of their styles [14–16].

But isn't a successful dendrogram just an accident? Let's try to add more authors (*impostors*) – F. Mauriac and J. Verne (Fig. 3). The structure of the dendrogram has remained almost unchanged, only new branches have been added, which indicates the reasonableness of our idea of numerals in texts as a stable feature of the author's style. Note that adding the works by Verne increases the height of the merger of two of Gide's texts (*Geneviève*, *Les faux-monnayeurs*) with his other three texts – this is a feature of the calculations in the average linkage method.

C. Russian-language texts

Figure 4 shows the dendrogram for works by Russian-language authors.

We again state that the works were distributed on the dendrogram in accordance with authorship. Dostoevsky's literary style is characterized by a very uniform use of numerals (the height of the amalgamation is low). The natural exception is *The House of the Dead*, which is semi-documentary in nature ("fictionalized memoir").

The examples given show that the manner of using numerals in texts is an author's invariant and can be used in stylometry problems. Of course, cluster analysis in itself does not have evidentiary value, but rather is a means of data visualization. But if necessary, it is possible to demonstrate the similarity/difference of data on numerals for different authors using mathematical statistics (see Appendix), which confirms the results obtained.

4. Does the Manner of Using Numerals Change When the Author Writes under a Pen Name?

We now proceed to the main task of this work.

A. Literary heritage of R. Gary

French novelist Romain Gary (1914–1980) was prone to literary hoaxes. In addition to works published under the name "Romain Gary" (which is itself a nom de plume), he also published under the names "Émile Ajar", "Fosco Sinibaldi" and "Shatan Bogat". Merely his first novel, *Le vin des morts* (1937), was published under his real name, "Roman Kacew". The only writer to twice receive the Prix Goncourt (first as Gary and again as Ajar), R. Gary, in his own words, left many hints in the texts of Ajar's works that made it possible to identify the true author, but critics, for the most part, turned out to be blind and did not recognize hints [17–19].

How much do the literary styles of Romain Gary and fictional authors differ from the point of view of our methodology?

To answer this question, we have analyzed

· Works released under the name "Romain Gary":

Education européenne, 1945,

Tulipe, 1946,

Le grand vestiaire, 1949,

Les racines du ciel, 1956, Prix Goncourt,

La promesse de l'aube, 1960,

Gloire à nos illustres pionniers (*Les oiseaux vont mourir au Pérou*), 1962,

Lady L., 1963,

Les mangeurs d'étoiles, 1966,

La danse de Gengis Cohn, 1967,

La tête coupable, 1968,

Adieu Gary Cooper, 1969,

Chien blanc, 1970,

Europa, 1972,

Les enchanteurs, 1973,

Au-delà de cette limite votre ticket n'est plus valable, 1975,

Clair de femme, 1977,

Charge d'âme, 1977,

Les clowns lyriques, 1979,

Les cerfs-volants, 1980,

· Works released under the name "Émile Ajar":

Gros calin, 1974,

La vie devant soi, 1975, second Prix Goncourt,

Pseudo, 1976,

L'Angoisse du roi Salomon, 1979,

· A work released under the name "Fosco Sinibaldi":

L'homme à la colombe, 1958,

· A work released under the name "Shatan Bogat":

Les têtes de Stéphanie, 1974,

· A work released under the real name "Roman Kacew":

Le vin des morts, 1937.

Figure 5 (left panel) shows the dendrogram of data concerning the occurrence of numerals (clustering principles are described in the Introduction; $n = 8$ is taken in formula (2)).

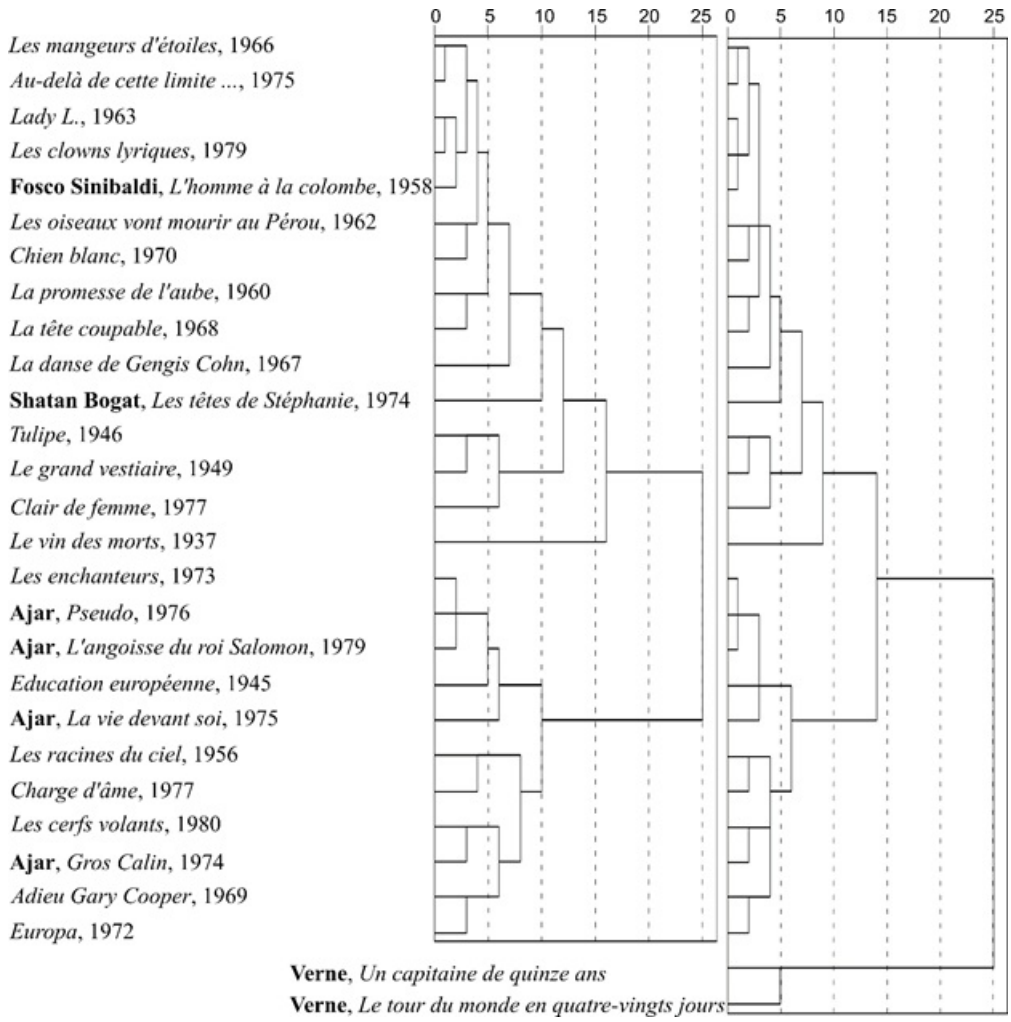

Figure 5. Left panel: results of hierarchical clustering of works by R. Gary, published under his own name and under pen names (in the latter case, the name is explicitly indicated). Right panel: the same, but with the addition of *impostors*: the works of J. Verne *Un capitaine de quinze ans* and *Le tour du monde en quatre-vingts jours*

It would seem that the graph does not confirm our idea about the specificity of the use of numerals by authors. But when adding two works by J. Verne to the analysis (right panel of Fig. 5), it becomes clear that it's all a matter of scale (on the horizontal axis): the height of the amalgamation of impostors and Gary's texts is almost twice the height of the internal amalgamation of Gary's texts. Note that the maximum height is always normalized to 25.

Any chronological sequence in the distribution of works in the dendrogram is not visible, however, note that in one of the two superclusters the early novel *Le vin des morts* (1937) stands out clearly – it is obvious that R. Gary was just developing his style in literature.

Works signed with R. Gary's own name are interspersed without any system with works published under pen names. So, there were no substantial changes in the manner of using numerals when R. Gary tried to change his literary style.

B. Literary hoaxes of Boris Akunin

Russian-language writer, literary critic, translator, liberal public figure Grigory Chkhartishvili (born 1956) publishes non-fiction texts under his real name, but as an author of novels since 1998 he is incomparably better known under the pen name "B. Akunin". Since 2007, works have been published under the pseudonyms "Anatoly Brusnikin" (*The Ninth Savior*, *A Hero of a Different Time*, *Bellona*) and "Anna Borisova" (*There...*, *The Idea-Man*, *Vremena*

goda). Subsequently, G. Chkhartishvili recognized the authorship of these works.

Did his literary style (as far as numerals are concerned) change when writing under a *nom de plume*?

Figure 6 shows the results of clustering data on the use of numerals in the works by "B. Akunin", "Anatoly Brusnikin" and "Anna Borisova" (clustering principles are described in the Introduction; n = 10 is taken in formula (2)).

The patterns in the distribution of names along the dendrogram are not clear enough to state with certainty that G. Chkhartishvili uses numerals substantially differently in works written under different names.

So, the examples of R. Gary and G. Chkhartishvili lead us to the (preliminary) conclusion that the manner of using numerals is invariant for each writer, and it is almost impossible to change it. Of course, this conclusion, to be fully justified, needs to be supported by other examples.

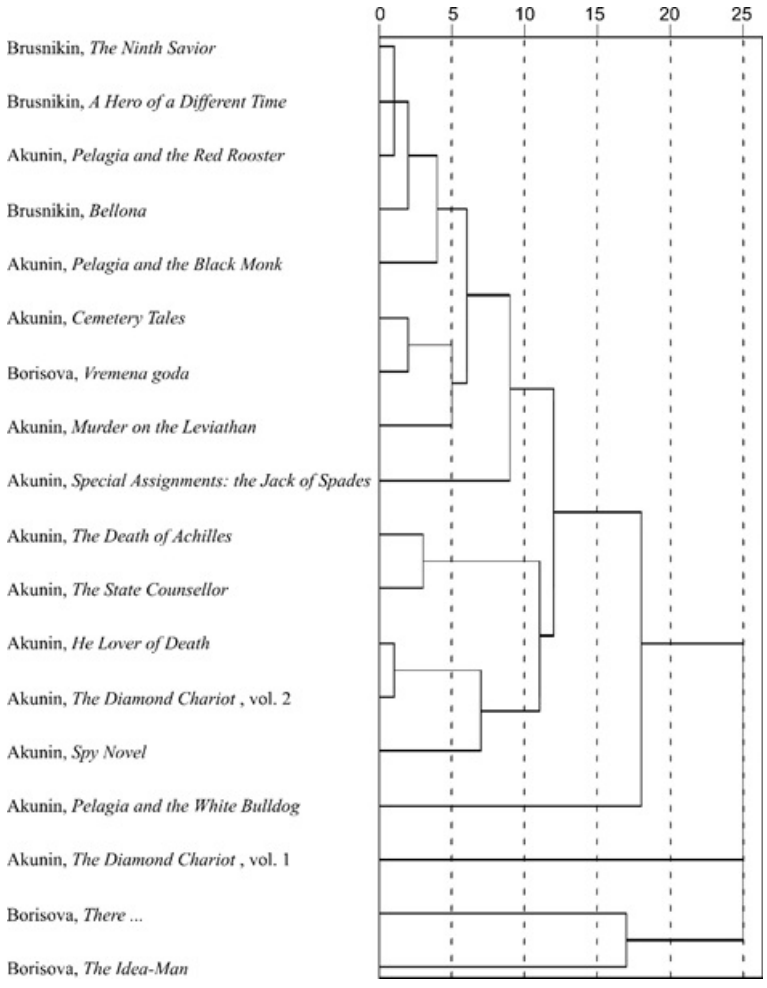

Figure 6. The results of hierarchical clustering of G. Chkhartishvili's works published under the most famous pseudonym "B. Akunin" and under the lesser-known pen names of "Anatoly Brusnikin" and "Anna Borisova"

C. The problem of authorship of Harper Lee's work

In previous examples (R. Gary and G. Chkhartishvili), our method of taking into account numerals served the purpose of ascertaining, and now we are starting on a problem that has not yet been definitively solved.

Harper Lee (1926–2016) – American writer, author of the famous novel *To Kill a Mockingbird* (1960), which is her only major literary work. In 2015, H. Lee's book *Go Set a Watchman* was published, which was written earlier than *To Kill a Mockingbird*, but was not published at the time. According to critics, this is not a separate novel, but merely the original version of *To Kill a Mockingbird*, which has now been tried, based on commercial interest, to be presented as an independent work [20].

Lee was a lifelong friend of Truman Capote (1924–1984) until his death. One of the characters in *To Kill a Mockingbird* was based on him. His numerous literary and documentary works are considered literary classics. In light of the above, it is understandable that suspicions have been repeatedly expressed that the novel *To Kill a Mockingbird* was also written by Capote.

Figure 7 presents the results of clustering data on the use of numerals in two novels by Lee, as well as in Capote's main works *The Grass Harp*, *A Christmas Memory*, *Breakfast at Tiffany's*, *Answered Prayers*, *Summer Crossing*, *Other Voices, Other Rooms* (clustering principles are described in the Introduction; in formula (2) n = 10 is taken).

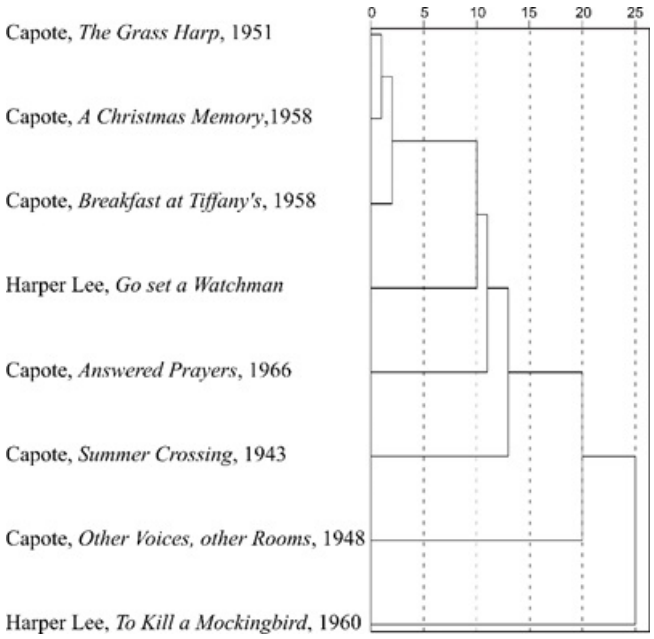

Figure 7. Results of hierarchical clustering of works by Harper Lee and Truman Capote

The dendrogram shows that the primary version of H. Lee's novel is close in terms of the use of numerals to Capote's novels, and, therefore, he could influence H. Lee's text. In the final version, *To Kill a Mockingbird*, Capote's influence, if any, is less pronounced.

Note that an early version of our stylometric method, based on taking into account the first significant digits of numerals in the text [7], led to a similar conclusion about the authorship of the novel by H. Lee.

In this analysis, the introduction of additional authors (*impostors*) is inappropriate, since the range of possible authors is initially limited to Lee and Capote. In [21], the authors come to a similar conclusion regarding Capote's influence on H. Lee's novel. However, they also consider the texts of Therese von Hohoff Torrey, "Tay Hohoff", 1898–1974, a literary editor who devoted much effort to improving H. Lee's original manuscript. But among Hohoff's four own works, two are not fiction, but documentary (*A Ministry to Man: The life of John Lovejoy Elliott, a biography*; *The Author and his Audience: With a Chronology of Major*

Events in the Publishing History of J. B. Lippincott Company), one work is intended for a children's audience (*The Cat Who Wanted Out*) and one more work is a memoir (*Cats and Other People*). These are very special texts, hardly suitable for analysis for the use of numerals; unfortunately, these books were not available to us.

Figure 8 presents the frequency dependence of numerals found in the above-mentioned works by Harper Lee and T. Capote. Absolute frequencies are recalculated taking into account different text sizes. For ease of perception, numerals are limited to the range (1; 40).

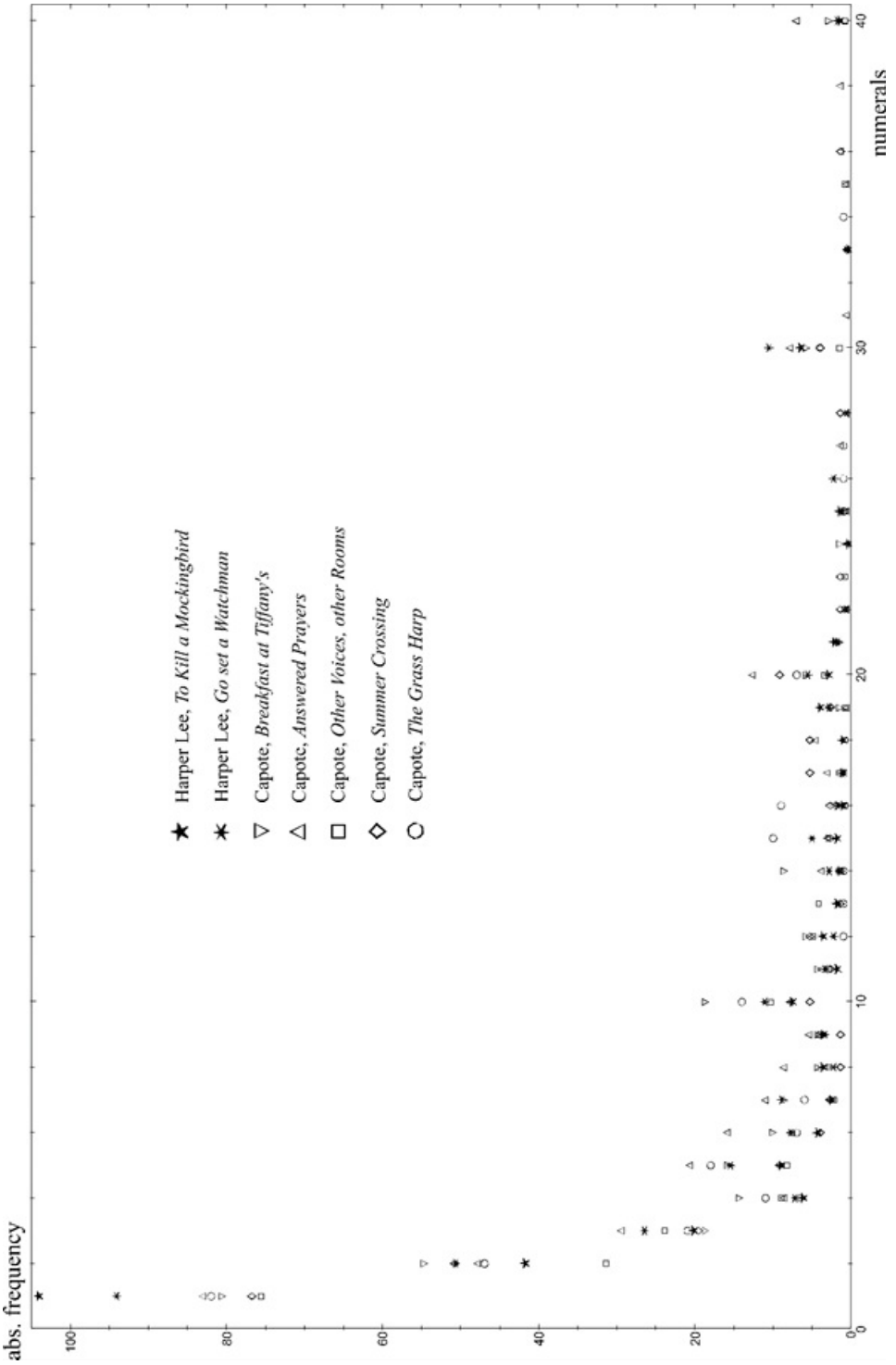

Figure 8. Frequency dependence of numerals occurring in the works by H. Lee and T. Capote

The following can be seen directly from the graph:

1) Common properties for all texts are a decrease in frequency with an increase in the numeral (i.e., the number denoted by it); the presence of local maxima on round numbers (10, 20, 30, ...), the height of which also gradually decreases; gradual rarefaction of the numbers series (the appearance of gaps on the axis of numerals). These conclusions are universal and valid for all texts we have analyzed.

2) In the texts by H. Lee, in contrast with the texts of T. Capote, the frequency of the numeral *one* is especially high, but the numeral *two* (and subsequent ones) has a relatively lower frequency, i.e. Lee resorts to numerals less often.

3) There is a greater variety of numerals in Capote's texts.

4) The frequency dependence of numerals in the text *Go set a Watchman* is closer to that for Capote's texts than in *To Kill a Mockingbird*. This is consistent with the conclusion obtained above from the dendrogram that the early version of H. Lee's novel is more close to the works of T. Capote.

5. Discussion of the Results and Conclusions

We have studied the use of numerals in literary texts of a large number of authors in English, French, and Russian. It was found that the manner of using numerals among different authors can vary substantially, and for each author it is consistently reproduced in different texts. Author's differences in the use of numerals are not only observed visually by means of cluster analysis on dendrograms, but are also confirmed by the Pearson's chi-squared test (see Appendix).

It turned out that in the series of numerals *one*, *two*, *three*, ... each subsequent element is found in texts, generally speaking, less and less often (with understandable maxima on the round numbers *ten*, *twenty*, ..., *one hundred*, ...). The reasons for the decrease in frequency are not entirely clear; some explanation is given by the experimentally discovered phenomenon of "Benford bias" [22] in human psychology: when people generate numbers, the frequency distribution of the first significant digits of numerals is distorted towards Benford's Law [9]: smaller digits are more common.

The study of the peculiarities of the use of numerals in author's literary texts published under different pen names has shown that the authors' attempts to change their creative style and write "differently" practically do not affect the occurrence of numerals in texts.

Thus, the manner of using numerals is an author's invariant (fingerprint), and this can be used when solving problems about the authorship of texts. The conclusion we have made regarding the attribution of two novels by the American writer Harper Lee is consistent with the conclusions obtained by other researchers using other methods.

So, the method we suggested, based on the analysis of the occurrence of numerals in texts, is another effective method of stylometry, which, of course, does not cancel the existing ones, but complements them.

6. Appendix

We will present here in more detail some computational aspects of our work.

According to Fig. 7, the final version of Lee's *To Kill a Mockingbird*, in terms of the use of numerals, is far from Capote's *Breakfast at Tiffany's*, and the original version of Lee's *Go Set a Watchman*, on the contrary, is close to this novel. Visual similarity/difference can be

supported by the Pearson's chi-squared test.

The comparison of empirical distributions (in our case, the distributions of absolute frequencies of numerals in the texts of various authors) is related to testing the statistical hypotheses about the significance/insignificance of differences between distributions [23].

We now formulate the hypotheses. The null hypothesis *H*0 asserts that the tested populations are distributed identically. The alternative hypothesis *H*1: The distributions differ from each other.

The parametric Pearson's chi-squared test, among other things, is also used as a test of homogeneity – it compares the distribution of counts for two or more groups using the same categorical variable. In the form we need, the corresponding procedure is not available in standard statistical packages, so we will describe it in detail.

Our initial statistical data concerning the occurrence of numerals *one*, *two*, ..., *ten* in three texts are given in the following table 1. Of course, larger numerals also appear in texts, but with ever less frequency.

Table 1. Empirical absolute frequencies of numerals in the analyzed texts.

Numeral	Absolute frequencies of numerals		
	Harper Lee, *To Kill a Mockingbird*	Harper Lee, *Go set a Watchman*	Capote, *Breakfast at Tiffany's*
1	289	171	56
2	116	92	38
3	56	48	13
4	17	13	10
5	25	28	11
6	12	14	7
7	7	16	6
8	10	4	3
9	10	6	3
10	21	20	13

For applicability of Pearson's test, the frequency in each cell should be not less than 5, so rows 8 and 9 will have to be merged:

Table 2. Empirical absolute frequencies of numerals after cell merging.

Numeral	Absolute frequencies of numerals		
	Harper Lee, *To Kill a Mockingbird*	Harper Lee, *Go set a Watchman*	Capote, *Breakfast at Tiffany's*
1	289	171	56
2	116	92	38
3	56	48	13
4	17	13	10
5	25	28	11
6	12	14	7
7	7	16	6

8 and **9**	20	10	6
10	21	20	13

We will compare each of H. Lee's texts separately with the text by T. Capote.

Table 3. Empirical absolute frequencies of numerals in *To Kill a Mockingbird* and *Breakfast at Tiffany's* after cell merging.

Numeral	Harper Lee, *To Kill a Mockingbird*		Capote, *Breakfast at Tiffany's*		Sum of frequencies over the row
	Empirical absolute frequency	Cell label	Empirical absolute frequency	Cell label	
1	289	I	56	II	289 + 56 = 345
2	116	III	38	IV	116 + 38 = 154
3	56	V	13	VI	56 + 13 = 69
4	17	VII	10	VIII	17 + 10 = 27
5	25	IX	11	X	25 + 11 = 36
6	12	XI	7	XII	12 + 7 = 19
7	7	XIII	6	XIV	7 + 6 = 13
8 and **9**	20	XV	6	XVI	20 + 6 = 26
10	21	XVII	13	XVIII	21 + 13 = 34
	Σ = **563**		Σ = **160**		ΣΣ = **723**

We will juxtapose empirical and theoretical frequencies, the latter obtained by taking into account that the numbers of the numerals (not exceeding ten) in the texts are different: 563 in Lee's text and 160 – in that by Capote. Thus, out of the total quantity 563 + 160 = 723 numerals in two texts, the first one accounts for the share 563/723 = 0.78, and the second for 160/723 = 0.22. In all the rows, the theoretical frequencies related to the first and second texts should thus be, respectively, 0.78 and 0.22 out of the total frequency of the corresponding row. If the empirical distributions to be compared do not differ from one another, the empirical frequencies should not significantly deviate from the theoretical ones, obtained from the proportion.

Now, we recompose the data of Table 3, placing the relative frequencies for both texts in the order indicated by the labels in one column (these will be the empirical frequencies *f*emp); in the other column, we will place the theoretical frequencies *f*theor, calculated according to the previous as

$$f_{\text{theor}} = \frac{(\Sigma \text{ frequencies over the row}) \cdot (\Sigma \text{ frequencies over the column})}{\Sigma\Sigma}.$$

Here, ΣΣ = 723.

Table 4. Calculations for Pearson's chi-squared test.

Cell	empirical frequency *f*emp	theoretical frequency *f*theor	$(f_{emp} - f_{theor})^2 / f_{theor}$
I	289	345·563/723 = 268.65	1.54
II	56	345·160/723 = 76.35	5.42
III	116	154·563/723 = 119.92	0.13
IV	38	154·160/723 = 34.08	0.45
V	56	69·563/723 = 53.73	0.10
VI	13	69·160/723 = 15.27	0.34
VII	17	27·563/723 = 21.02	0.77
VIII	10	27·160/723 = 5.98	2.70
IX	25	36·563/723 = 28.03	0.33
X	11	36·160/723 = 7.97	1.15
XI	12	19·563/723 = 14.80	0.53
XII	7	19·160/723 = 4.20	1.87
XIII	7	13·563/723 = 10.12	0.96
XIV	6	13·160/723 = 2.88	3.38
XV	20	26·563/723 = 20.25	0.00
XVI	6	26·160/723 = 5.75	0.01
XVII	21	34·563/723 = 26.48	1.13
XVIII	13	34·160/723 = 7.52	3.99
	Σ = 723	**Σ = 723**	**Σ = 24.81** = χ^2_{emp}

Now, we determine the number of degrees of freedom, *df*. For the test of homogeneity, $df = (r-1)(c-1)$, where *r* corresponds to the number of categories (i.e. rows in the table of empirical frequencies; *r* = 9 – see Table 2), and *c* corresponds the number of independent groups (here, *c* = 2). Therefore, *df* = 8.

With such *df*, the tabulated critical values of the χ^2 distribution for two significance levels α are:

$$\chi^2_{cr} = \begin{cases} 15.5 \ (\alpha = 0.05), \\ 20.1 \ (\alpha = 0.01). \end{cases} \quad (3)$$

Since the empirical χ^2_{emp} = 24.81 exceeds each of these critical values, hypothesis *H*0 (asserting that both the tested populations are distributed identically) is rejected; in other words, the distribution of numerals in *To Kill a Mockingbird* by Harper Lee and *Breakfast at Tiffany's* by Truman Capote differ significantly.

Now we compare the primary version of the novel, *Go Set a Watchman*, by H. Lee with the same novel by T. Capote.

Table 5. Empirical absolute frequencies of numerals in *Go Set a Watchman* and *Breakfast at Tiffany's* after cell merging.

Numeral	Harper Lee, *Go set a Watchman*		Capote, *Breakfast at Tiffany's*		Sum of frequencies over the row
	Empirical absolute frequency	Cell label	Empirical absolute frequency	Cell label	
1	171	I	56	II	171 + 56 = 227
2	92	III	38	IV	92 + 38 = 130
3	48	V	13	VI	48 + 13 = 61
4	13	VII	10	VIII	13 + 10 = 23
5	28	IX	11	X	28 + 11 = 39
6	14	XI	7	XII	14 + 7 = 21
7	16	XIII	6	XIV	16 + 6 = 22
8 and **9**	10	XV	6	XVI	10 + 6 = 16
10	20	XVII	13	XVIII	20 + 13 = 33
	Σ = **412**		Σ = **160**		ΣΣ = **572**

Out of the total quantity 412 + 160 = 572 numerals in two texts, the first one accounts for the share 412/572 = 0.72, and the second for 160/572 = 0.28. In all the rows, the theoretical frequencies related to the first and second texts should thus be, respectively 0.72 and 0.28 out of the total frequency of the row.

Performing calculations similar to those done above, we get

Table 6. Calculations for Pearson's chi-squared test.

Cell	empirical frequency *f*emp	theoretical frequency *f*theor	$(f_{emp} - f_{theor})^2 / f_{theor}$
I	171	227·412/572 = 163.50	0.34

II	56	227·160/572 = 63.50	0.89
III	92	130·412/572 = 93.64	0.03
IV	38	130·160/572 = 36.36	0.07
V	48	61·412/572 = 43.94	0.38
VI	13	61·160/572 = 17.06	0.97
VII	13	23·412/572 = 16.57	0.77
VIII	10	23·160/572 = 6.43	1.98
IX	28	39·412/572 = 28.09	0.00
X	11	39·160/572 = 10.91	0.00
XI	14	21·412/572 = 15.13	0.08
XII	7	21·160/572 = 5.87	0.22
XIII	16	22·412/572 = 15.85	0.00
XIV	6	22·160/572 = 6.15	0.00
XV	10	16·412/572 = 11.52	0.20
XVI	6	16·160/572 = 4.48	0.52
XVII	20	33·412/572 = 23.77	0.60
XVIII	13	33·160/572 = 9.23	1.54
	Σ = 572	**Σ = 572**	**Σ = 8.59** = χ^2_{emp}

The critical values of the χ^2 distribution remain the same (3) as above. Since the empirical χ^2_{emp} = 8.59 is *less* than these critical values at both significance levels, we fail to reject the hypothesis *H*0: the patterns of numerals usage in *Go Set a Watchman* by Harper Lee and *Breakfast at Tiffany's* by Truman Capote are indistinguishable (at given significance levels).

We performed similar (very cumbersome!) calculations for all the above dendrograms, and it turned out that the visual similarities/differences in the use of numerals by the authors can be trusted.

Библиография

1. Stamatatos E. A survey of modern authorship attribution methods // J. of the American Society for information Science and Technology. 2009. No. 60(3), Pp. 538–556.
2. Tempestt N., Kalaivani S., Aneez F., Yiming Y., Yingfei X., Damon W. Surveying Stylometry Techniques and Applications // ACM Comput. Surv. 2017, 50(6), Article 86, 36 pages.
3. Brocardo M. L., Traore I., Woungang I., Obaidat M. S. Authorship verification using deep belief network systems // Int. J. Commun. Syst., 2017. DOI:10.1002/dac.3259.
4. La Inteligencia Artificial ayuda a descubrir una obra desconocida de Lope de Vega en

los fondos de la BNE, Biblioteca Nacional de España, https://www.bne.es/es/noticias/inteligencia-artificial-ayuda-descubrir-obra-desconocida-lope-vega-fondos-bne (Accessed: October 18, 2023).

5. Зенков А. В. Отклонения от закона Бенфорда и распознавание авторских особенностей в текстах // Компьютерные исследования и моделирование. 2015. № 7(1). С. 197–201.
6. Зенков А. В. Новый метод стилеметрии на основе статистики числительных // Компьютерные исследования и моделирование. 2017. № 9(5). С. 837–850.
7. Zenkov A. V. A Method of Text Attribution Based on the Statistics of Numerals // J. of Quantitative Linguistics. 2018. No. 25(3). Pp. 256–270.
8. Zenkov A. V., Místecký M. The Romantic Clash: Influence of Karel Sabina over Macha's Cikani from the Perspective of the Numerals Usage Statistics // Glottometrics. 2019 No. 46, Pp. 12–28.
9. Zenkov A. V. Stylometry and Numerals Usage: Benford's Law and Beyond // Stats 2021. No. 4, Pp. 1051–1068.
10. Zenkov A. and Místecký, M. Young Vladimír Vašek? – A Numerals Analysis Contribution to the Bezruč–Hrzánský Identity Issue // Naše řeč. 2022 No. 105(3). Pp. 151–161.
11. Зенков А. В., Ермаков Н. Е. Числительные в текстах как характерная особенность авторского стиля // Russian Linguistic Bulletin. 2023. № 9(45). 6 с.
12. Moisl H. Cluster Analysis for Corpus Linguistics. Berlin, München, Boston: De Gruyter Mouton, 2015.
13. Koppel M., Winter Y. Determining if Two Documents are Written by the Same Author // J. of the Association for Information Science and Technology. 2014. No. 65(1). Pp. 178–187.
14. Artinian A. Maupassant criticism in France, 1880 – 1940, with an inquiry into his present fame and a bibliography. N. Y.: Kings Crown Press, 1941.
15. Dugan J. R. Illusion and Reality, A Study of Descriptive Techniques in the Works of Guy de Maupassant. The Hague, Paris: Mouton, 1973.
16. Lloyd C. Guy de Maupassant. Reaktion Books, 2020.
17. Boisen J. Un Picaro métaphysique: Romain Gary et l'art du roman. Odense University Press, 1996.
18. Hocus Bogus. Romain Gary writing as Émile Ajar, Transl. by D. Bellos, New Haven, London: Yale University Press, 2010.
19. Poier-Bernhard A. Romain Gary – das brennende Ich: literaturtheoretische Implikationen eines Pseudonymenspiels. Tübingen: Niemeyer, 1996.
20. Shields C. J. Mockingbird: A Portrait of Harper Lee: From Scout to Go Set a Watchman. Henry Holt and Company, 2016.
21. Michał Choiński, Maciej Eder, Jan Rybicki, Harper Lee and Other People: A Stylometric Diagnosis // Mississippi Quarterly. 2017/2018. No. 70/71(3). Pp. 355–374.
22. Burns B. D. Do People Fit to Benford's Law, or Do They Have a Benford Bias? Available online: https://cognitivesciencesociety.org/cogsci20/papers/0379/index.html (Accessed: October 18, 2023).
23. Clarke G. M., Cooke D. A basic Course in Statistics. London: Hodder Arnold, 2004.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не

раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Представленная на рассмотрение статья «Под чужим флагом: литературные мистификации и использование числительных», предлагаемая к публикации в журнале «Litera», представленная на английском языке, несомненно, является актуальной, ввиду возрастающего интереса к изучению языка художественной прозы.
Автор обращается к методам стилометрии. Автор предлагает новый подход к проблемам стилометрии, основанный на анализе использования числительных в (литературном) авторском тексте. В работе демонстрируются преимущества предлагаемой методики.
В работе рассматриваются литературные мистификации, под которыми понимается публикация произведений одним автором под разными псевдонимами, а также анализируются числительные.
Автором предложена компьютерная программа, которая ищет как кардинальные, так и порядковые числа, выраженные как числами, так и (что значительно чаще) устно (в разных словоформах) в англоязычных, франкоязычных и русскоязычных текстах.
Отметим наличие сравнительно небольшого количества исследований по данной тематике в отечественном языкознании. Статья является новаторской, одной из первых в российской лингвистике, посвященной исследованию подобной проблематики. В статье представлена методология исследования, выбор которой вполне адекватен целям и задачам работы. Автор обращается, в том числе, к различным методам для подтверждения выдвинутой гипотезы.
К сожалению, автор не указывает на объем корпуса, отобранного для практической части исследования, принципы и методы отбора.
Данная работа выполнена профессионально, с соблюдением основных канонов научного исследования. Исследование выполнено в русле современных научных подходов, работа состоит из введения, содержащего постановку проблемы, основной части, традиционно начинающуюся с обзора теоретических источников и научных направлений, исследовательскую и заключительную, в которой представлены выводы, полученные автором. Автор иллюстрирует теоретические положения языковым материалом, а также графиками и диаграммами, часть материала представлена в табличных формах, что облегчает восприятие читателем.
Библиография статьи насчитывает 23 источника, среди которых представлены научные труды на русском и английском языках. К сожалению, в статье отсутствуют ссылки на фундаментальные работы отечественных исследователей, такие как монографии, кандидатские и докторские диссертации. Технически при оформлении библиографического списка нарушены общепринятые требования ГОСТа, а именно несоблюдение алфавитного принципа оформления источников, смешение работ на иностранном и русском языках. В общем и целом, следует отметить, что статья написана простым, понятным для читателя языком. Опечатки, орфографические и синтаксические ошибки, неточности в тексте работы не обнаружены. Высказанные замечания не являются существенными и не умаляют общее положительное впечатление от рецензируемой работы. Работа является новаторской, представляющей авторское видение решения рассматриваемого вопроса и может иметь логическое продолжение в дальнейших исследованиях. Практическая значимость исследования заключается в возможности использования его результатов в процессе преподавания вузовских курсов по стилистике, литературоведению, а также курсов по междисциплинарным исследованиям, посвящённым связи языка и общества. Статья, несомненно, будет полезна широкому кругу лиц, филологам, магистрантам и аспирантам профильных вузов.
Статья «Под чужим флагом: литературные мистификации и использование числительных» может быть рекомендована к публикации в научном журнале.