applied sciences

MDPI

Article

The Question of Studying Information Entropy in Poetic Texts

Olga Kozhemyakina *, Vladimir Barakhnin, Natalia Shashok and Elina Kozhemyakina

Federal Research Center for Information and Computational Technologies, Novosibirsk 630090, Russia; bar@ict.nsc.ru (V.B.); n.shashok@alumni.nsu.ru (N.S.); kojemyakina.elina2017@yandex.ru (E.K.)
* Correspondence: olgakozhemyakina@mail.ru; Tel.: +7-(913)-946-22-80

Abstract: One of the approaches to quantitative text analysis is to represent a given text in the form of a time series, which can be followed by an information entropy study for different text representations, such as "symbolic entropy", "phonetic entropy" and "emotional entropy" of various orders. Studying authors' styles based on such entropic characteristics of their works seems to be a promising area in the field of information analysis. In this work, the calculations of entropy values of the first, second and third order for the corpus of poems by A.S. Pushkin and other poets from the Golden Age of Russian Poetry were carried out. The values of "symbolic entropy", "phonetic entropy" and "emotional entropy" and their mathematical expectations and variances were calculated for given corpora using the software application that automatically extracts statistical information, which is potentially applicable to tasks that identify features of the author's style. The statistical data extracted could become the basis of the stylometric classification of authors by entropy characteristics.

Keywords: quantitative text analysis; information entropy; author's style features

Citation: Kozhemyakina, O.; Barakhnin, V.; Shashok, N.; Kozhemyakina, E. The Question of Studying Information Entropy in Poetic Texts. *Appl. Sci.* **2023**, *13*, 11247. https://doi.org/10.3390/app132011247

Academic Editor: Valentino Santucci

Received: 31 August 2023
Revised: 21 September 2023
Accepted: 29 September 2023
Published: 13 October 2023

1. Introduction

One of the approaches to text quantitative analysis is to represent a given text in the form of a time series, which can be followed by the study of information entropy—a measure of the randomness and the appearance of the next symbol. The printed text in natural language can be translated into a time series representation since the person reading the text mostly perceives the printed symbols in a specific order. Oral speech, perceived by a person through the ear, can also be described as a sequence of irreversible states, changed in time, as a sound wave or a time series. Therefore, text representation in the form of phonetic transcription allows the study of it using the mathematical apparatus of time series theory and information theory.

Entropy methods have been successfully applied in the study of the differences between poetic texts in different languages through the definition of rhythmic structure when studying the degree of proximity of automatically generated texts to texts written by a human, as well as in determining the differences between the texts belonging to different genres.

The purpose of this article is to consider the questions that relate to the study of information entropy in poetic texts, to describe the software developed to apply the methods of information theory and mathematical statistics to poetic texts in Russian, to the task of determining the author's style of texts ("symbolic" entropy, "phonetic" entropy, "emotional" entropy), and to present the entropic characteristics of the works of Pushkin-era poets.

The entropy of the first, second and third orders can be calculated for poetic works written by the classical Russian poets, namely those of A.S. Pushkin (323 works) and poets of his era: K.N. Batyushkov, E.A. Boratynsky, V.A. Zhukovsky, A.A. Delvig (correspondingly 102, 197, 210, 199 works). This study used the corpora provided by the Fundamental Digital Library [1]. The works of these authors were chosen due to the linear graphic structure of their texts, and it was assumed that the approach to calculating the entropy values for the works of different authors and in different languages might be different.

Entropy was calculated for both the symbolic recordings of texts and their phonetic transcriptions. In addition, for all these authors, the "emotional" entropy of the first, second and third order was calculated as well. To calculate "emotional entropy", all the final punctuation marks (dot, ellipsis, question mark, exclamation mark, question mark with two dots, exclamation mark with two dots), which to some extent reflect the emotional coloring of sentences in the text, were selected from the text after the usual rules for calculating entropy were applied to the resulting sequence of symbols.

After calculating all the entropy characteristics listed above, through the use of modern machine learning methods, the classification of texts was carried out based on these characteristics, both individually and in groups. Moreover, through the usage of the Shapley method, explanations of the obtained classifications could be carried out and potentially provide insights into which entropic features can characterize the work of authors.

2. Related Works

A significant contribution to the formation of quantitative methods of analysis in Russian poetry was made by A.A. Markov (senior), whose article "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains" (1913) demonstrates the frequency and repetition of sound units (vowels and consonants) in the text of the novel by A.S. Pushkin [2]. As a system, the researcher considered the random process of the appearance of a letter in the text, while it was clear that the probability of the next letter was largely determined by the previous one: for example, the letter "з" rarely comes after "ч", while the letter "а" often does; also, for example, the combination of "space"-"ы" is impossible in the literary language. The recalculation of all the letters of "Eugene Onegin" gave a picture of the frequency of letters following each other. The mathematical model of such a process was described as follows: let there be a system with discrete states $S = \{s_1, s_2, \ldots, s_N\}$; the time is set in such a way that at each discrete time step, there is a transition to one of the existing states; only those iterations in which the state jumps occur are excluded and considered; each state is described by a set of conditional probabilities of transition to any state of the system: $s_1 = \{p(s_1 \mid s_1), p(s_2 \mid s_1), \ldots\}$. This model, which was represented using a geometric scheme (directed graph) and subsequently received the name of Markov chains, represents the historical basis of the method of authorship identification. The methodology associated with Markov chains has not become less popular in modern research; for example, in the article by I.I. Drozdova and A.D. Obukhova, "The determination of the authorship of a text by frequency characteristics" [3], the topic of determining the authorship of an anonymous text due to the frequency of characteristics is considered.

Among the studies carried out by A.N. Kolmogorov in the 60s of the 20th century, a special place was occupied by the works devoted to the analysis of speech statistics and the study of poetry. The research in this direction is closely related, on the one hand, to probabilistic and algorithmic approaches to information theory and, on the other hand, reflects the interest in analyzing the patterns that are inherent in the form and language of literary works. The global idea expressed by A.N. Kolmogorov is that the "entropy of speech" (i.e., the measure of the amount of information transmitted by speech) can be decomposed into two components: extra-verbal (semantic) and speech (linguistic) information [4]. The first of these components characterizes diversity, which makes it possible to transmit different semantic information. The second component, called "residual entropy" by Kolmogorov, characterizes the various possible ways of expressing the same or equivalent semantic information; in other words, this component describes the "flexibility" of speech or the "flexibility" of expression. The presence of "residual entropy" provides speech a special artistic including, in particular, the expressiveness of sound when transmitting intended semantic information. Thanks to the research of A.N. Kolmogorov, the estimate of the "residual entropy" can be obtained, and the "entropy costs" for individual techniques of sound expressiveness in the verse can be calculated. However, the corresponding part of

the work of A.N. Kolmogorov and A.V. Prokhorov, "Statistical methods for studying the rhythm of poetic speech", was not fully published [5].

An example of using the entropy approach in the classification problem is the system "Linguoanalyzer" [6].

The work of V.A. Gogoleva and A.P. Shkaraputa, "The mathematical approach to the determination of authorship and time of creation of a text based on the study of its entropy", is interesting [7], which is the analysis of entropy and frequency characteristics of the text, depending on the time of its creation and authorship, was carried out. The Russian prose texts were studied, where the frequencies of 32 letters of the Russian alphabet, punctuation marks and space were considered. The authors concluded that this approach, based on the calculation of entropy in the text, can be used with a high degree of accuracy to determine the authorship of the text.

Among the latest research in the considered field is the work of A.V. Poltavsky and E.Yu. Rusyaeva [8], in which the methods of performing entropy analysis on texts are used for machine learning based on precedents for the recognition of text information. In addition, the authors claim that their proposed method is also suitable for identifying the authorship of the text, but the example of calculations they give shows only the possible applicability of this method for identifying the specific works of one author.

In foreign studies, there are also a number of works related to the definition of entropy characteristics. Thus, in the work of R. Mansilla and E. Bush [9], through the definition of the rhythmic structure, the entropy characteristics of texts were analyzed. This considered the development of a method to analyze the increase in complexity from classical Greek poetry to classical Latin poetry by comparing "large samples" of this poetry with symbolic time series. Using a method from information theory, more precisely, the Renyi entropy—a generalization of Shannon entropy—and a measure of quantitative diversity, including uncertainty or randomness of some systems [10], the authors of the article showed how the rhythmic patterns in Greek poetry evolve to more complex forms in Latin poetry, and concluded that the complexity of this rhythmic structure is observed in hexameter from Greek poetry to Latin. This method, according to the authors, made it possible to distinguish Greek and Latin verses, relying on the differences in the usage of dactyl and spondei, as well as in the position and usage of caesura.

In the article by O. Calin (2020) [11], the quantitative approach to poetry based on the use of several statistical indicators (entropy, information energy, *n*-grams, etc.), applied to several iconic works of English literature, is presented. The author of the article determines the fact that the entropy of the English language changes over time, and this entropy depends on the language and the author. To estimate the information entropy between two texts, the statistical method is used; this method was developed to calculate the average information transmitted by a group of letters about the next letter in the text. In addition, the author claims that this formula has been found to calculate Shannon's language entropy [12] and introduces the concept of the *n*-gram information energy of poetry. The results also include the construction of a neural network that is capable of generating poetry close to authentic Byronic poetry and analyzing it.

In the work of this group of authors [13], the proximity of automatically generated texts to texts written by humans is studied using mathematical methods. We can also note the article by J. Ackerman, in which the "entropy of Sounds" was intended to help in determining genre differences between texts. [14]. In the last two articles, special attention is paid to the information entropy and cross-entropy of texts and the applicability of these statistical parameters in the classification task.

Thus, there is an obvious interest in research related to the calculation of entropy characteristics and the subsequent application of obtained data in the task of analyzing textual information in natural language.

However, none of the works discussed above (and especially works related to the analysis of texts in Russian) study the phonetic component of the text, despite the fact that letters and sounds are a priori different objects. Translating a text into its phonetic notation is

often a non-trivial task that requires an appeal to an expert to distinguish the pronunciation of letters in different contexts, and such a translation may not be unambiguous. From this follows the reasonable assumption that depending on the object and application of the methods—letters or sounds—the results of used statistical methods vary.

Also, none of these works analyzes the statistical characteristics or the use of punctuation marks since, in machine learning and document classification tasks, they are excluded from the studied texts. Nevertheless, from the point of view of the study of poetic texts, punctuation marks are naturally considered an "emotional characteristic", and it can be assumed that these punctuation marks do not change the meaning of sentences that carry a semantic coloring, which is important for the author when writing the work. A similar idea is expressed in [15].

The application of statistical methods both to the text itself and to its transcription and punctuation can improve the result of the algorithm for determining the author's style of the text. It should be noted that the task of determining the author's style is not equivalent to the task of determining authorship: this task only examines the features characteristic of individual authors without the assumption that the authorship can be determined on the basis of these features. As mentioned earlier, we are not aware of works based on the use of entropy methods, which consider the phonetic transcription of the text, although, undoubtedly, for Russian poetry, the phonetic transcription of this poetic text reflects the author's intention much more adequately compared to its literal notation. There is also no work on the application of entropy methods in the study of stylometry.

3. The Usage of the Information Entropy in the Tasks of Determining the Author's Style

Upon transmitting information over a certain channel, it is possible to quantify the randomness of the appearance of a particular symbol or a set of symbols. The measure of this randomness is called information entropy and is one of the basic concepts of information theory.

The concept of entropy as a measure of randomness was introduced by Claude Shannon in his article "Mathematical Theory of Communication" [12]. Entropy, in this context, is understood as a measure of the average amount of information needed to record some event taken from the probability distribution of a random variable.

There are the information entropies of different orders, which are calculated through the probabilities of using a certain number of symbols in line with the text. The entropy of the first order denotes the uncertainty of occurrence in the text of a separate unigram; the second is a separate bigram, and so on.

The information entropy of any text is calculated using the formula:

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{1}$$

in which S denotes the information entropy of power s, n equals the number of unique substrings of length s contained within the text, i denotes the ordinal number assigned to the i-th unique substring of length s, and p_i describes the probability of the i-th substring to appear within the text; therefore, it is a dimensionless quantity.

Information entropy can be both the subject of research and some "basic" concept for other statistical methods, in particular, for measuring information gain (Kullback–Leibler divergence or relative entropy) and mutual information, which is used in machine learning, for example, when building decision trees.

Another entropy characteristic is cross-entropy, which is a measure of the difference between two probability distributions, which can be interpreted as a measure of the difference between two texts. It is calculated using the following formula:

$$H = -\sum_{i=1}^{n} p_i \log_2 q_i, \tag{2}$$

in which H denotes the cross-entropy value, n equals the number of unique symbols contained within both texts, i denotes the ordinal number assigned to the i-th symbol, and p_i and q_i denote the probability of the i-th symbol within the first and the second text, respectively.

According to Yu.M. Lotman, metric calculations give the deviation of the text from the "average" case, which makes it possible to evaluate the information capabilities of various texts, genres and authors [16].

This postulate is used by Yu.N. Orlov and K.P. Osminin in the monograph "Methods of statistical analysis of literary texts" [17], in which, in particular, the deviation of texts by the parameter of information entropy from the "average" entropy of the language is considered. According to the authors' conclusions, it seems that the functional S given by form (1) is more effective in the tasks of formal classification of texts than in determining authorship. However, despite the fact that the result of the functional S can be compared to the works of the analyzed authors, one-letter distributions do not allow the genre of a work to be determined.

Later, in the same work, the researchers stated: "the authors have significant differences in the frequency of usage of letter pairs" (although these differences can rather be used in assessing the general trends). This serves as a confirmation of a postulate from the earlier work by the same authors: "the selective distribution functions of texts by letters and letter pairs can serve as a tool for grouping works by authors and genres" [18]. Based on this, the next assumption could be made when using a higher-order functional S, where some general tendencies could be estimated both for authors and for the genres of their works.

At the same time, these authors additionally suggested that the second-order functional S only allowed to clearly separate time epochs from each other. This could be stated as a contradiction: if the high-order functional S only allows works to be divided by time epochs, then it may not be able to determine the features of the author's style or genres. Moreover, these researchers themselves suggest that such a conclusion about time epochs could be dictated by the specifics of the selected works since, from later epochs and according to the authors of the study, works that are not part of the literary heritage of the classical period were selected for analysis, and, perhaps, therefore, their authors do did have such mastery of the word as authors of earlier epochs selected for analysis.

It makes sense to consider the authors' statements [17] in relation to information entropy, which was calculated with the help of transcription of the text and not of its literal notation, especially when applied to poetic texts, as well as to punctuation marks; that is, the "emotional" component of the text.

4. Methods

A software application for the automatic extraction of statistical data from Russian poetic texts was implemented in the Python programming language. The software application received a number N, denoting the entropy power to be studied, and a CSV document contained the complete works of the author or authors, with each work annotated with the year of writing. The references, epigraphs, line numbers, page numbers, as well as other elements not considered to be a part of the texts were removed from the document beforehand.

The texts of the works were then transcribed using the modified *epitran* library, which provided the phonetic representation of the texts. The initial document was then divided into letters (the symbolic part of the corpus) and punctuation marks (the emotional part of the corpus). The three received representations of the text—symbolic, phonetic and composed of punctuation marks—were then divided into n-grams, or the sequences of language characters (letters, sounds, punctuation marks) of the same length N. In the case of splitting the phonetic transcription into "symbols" of oral speech, it was taken into account that in the Russian language, there are soft consonants and stressed vowels denoted by special symbols.

After receiving the transcription of a poem, the values of the probabilities of occurrence in the text of each n-gram were calculated for the three sets of n-grams, and the entropy value was calculated according to the form (1). These values were sorted in descending order.

The obtained data—date of writing, the title, probabilistic characteristics in the json format, and the degree of entropy and entropy value—were stored in the CSV format, suitable for further use in other research tasks, such as the task of evaluating the author's style.

The output data of the application included nine CSV format files: this included the values of the probability of occurrence of n-grams in the text, phonetic transcription and punctuation marks, where $n \in \{1, 2, 3\}$, given that in this work only the entropy values of the 1st, 2nd and 3rd orders were analyzed.

The block algorithm of the software application is presented in Figure 1.

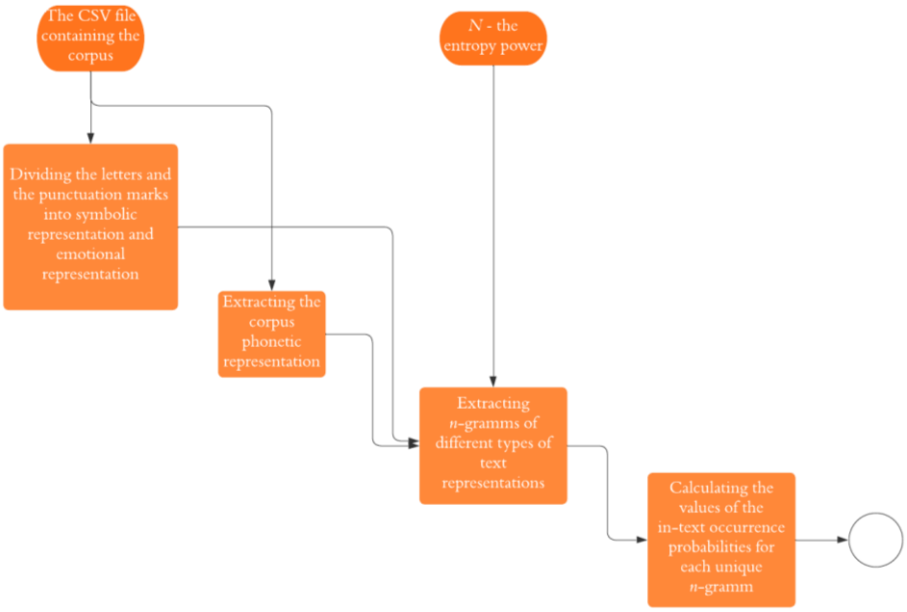

Figure 1. Software application algorithm.

5. Results

The obtained entropy values across the corpora are presented in Tables 1–5; given that the form (1) operates with dimensionless quantities, the entropy values, as well as the mathematical expectations, and the standard deviations are dimensionless.

Table 1. Informational entropy of A.S. Pushkin's works.

Order	Entropy Type	Mathematical Expectation	Standard Deviation
1st	Symbolic	4.409	0.012
	Phonetic	4.61	0.018
	Emotional	0.957	0.493
2nd	Symbolic	7.144	0.44
	Phonetic	7.191	0.482
	Emotional	1.311	1.265
3rd	Symbolic	8.12	1.398
	Phonetic	8.134	1.412
	Emotional	1.375	1.972

Table 2. Informational entropy of K.N. Batyushkov's works.

Order	Entropy Type	Mathematical Expectation	Standard Deviation
1st	Symbolic	4.470	0.003
	Phonetic	4.709	0.006
	Emotional	1.328	0.327
2nd	Symbolic	7.654	0.216
	Phonetic	7.744	0.26
	Emotional	2.195	1.014
3rd	Symbolic	9.061	1.059
	Phonetic	9.084	1.091
	Emotional	2.628	1.886

Table 3. Informational entropy of Ye.A. Baratynsky's works.

Order	Entropy Type	Mathematical Expectation	Standard Deviation
1st	Symbolic	4.429	0.008
	Phonetic	4.639	0.012
	Emotional	0.997	0.28
2nd	Symbolic	7.281	0.234
	Phonetic	7.331	0.266
	Emotional	1.477	0.879
3rd	Symbolic	8.282	0.818
	Phonetic	8.298	0.828
	Emotional	1.545	1.492

Table 4. Informational entropy of V.A. Zhukovsky's works.

Order	Entropy Type	Mathematical Expectation	Standard Deviation
1st	Symbolic	4.424	0.0098
	Phonetic	4.653	0.02
	Emotional	1.048	0.416
2nd	Symbolic	7.329	0.503
	Phonetic	7.406	0.585
	Emotional	1.656	1.325
3rd	Symbolic	8.497	1.741
	Phonetic	8.516	1.792
	Emotional	1.979	2.198

Table 5. Informational entropy of A.A Delvig's works.

Order	Entropy Type	Mathematical Expectation	Standard Deviation
1st	Symbolic	4.419	0.013
	Phonetic	4.634	0.024
	Emotional	0.909	0.301
2nd	Symbolic	7.205	0.412
	Phonetic	7.261	0.451
	Emotional	1.352	0.943
3rd	Symbolic	8.176	1.249
	Phonetic	8.187	1.256
	Emotional	1.458	1.672

The obtained values of the mathematical expectation and the standard deviation of entropy are shown in Figures 2–10 ("lower bound" denotes mathematical expectation minus standard deviation, "upper bound" denotes the sum of mathematical expectation and standard deviation).

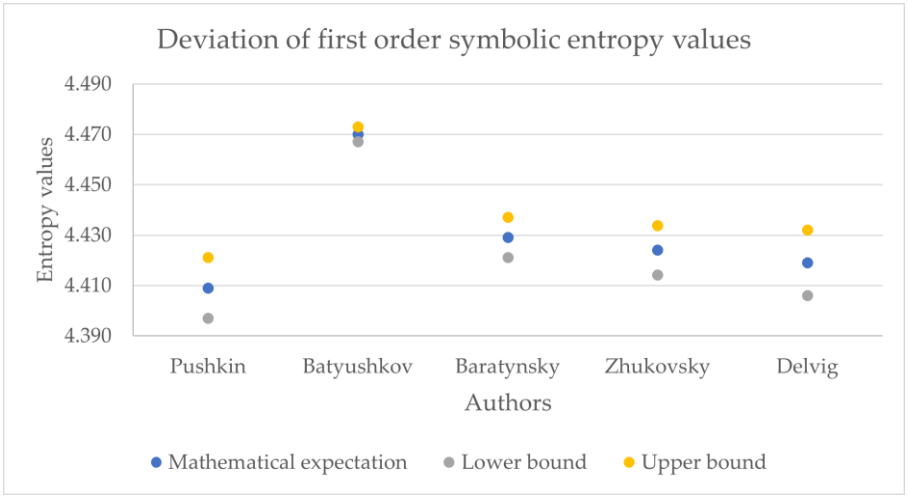

Figure 2. First-order symbolic entropy values.

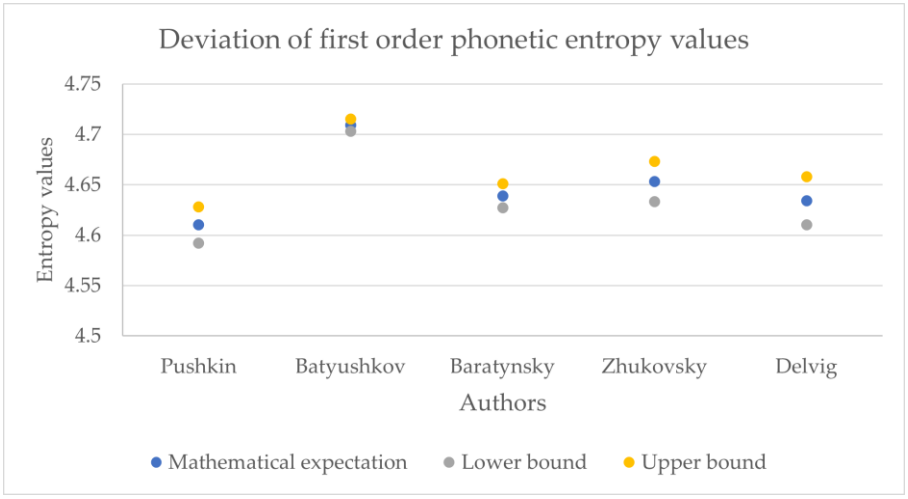

Figure 3. First-order phonetic entropy values.

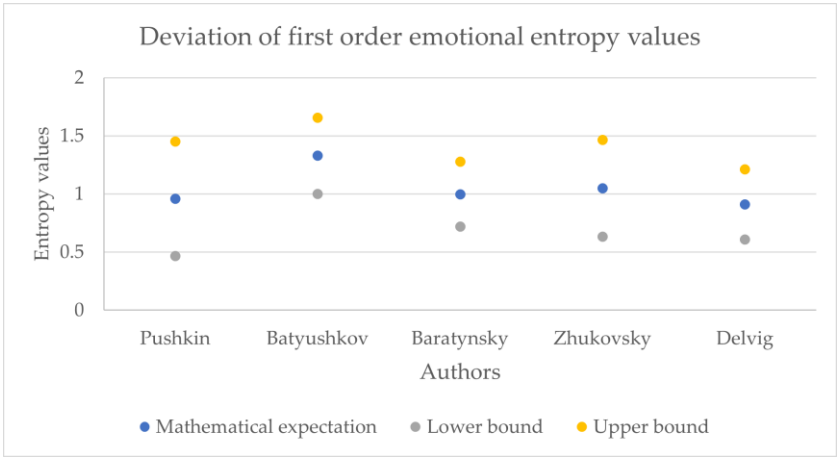

Figure 4. First-order emotional entropy values.

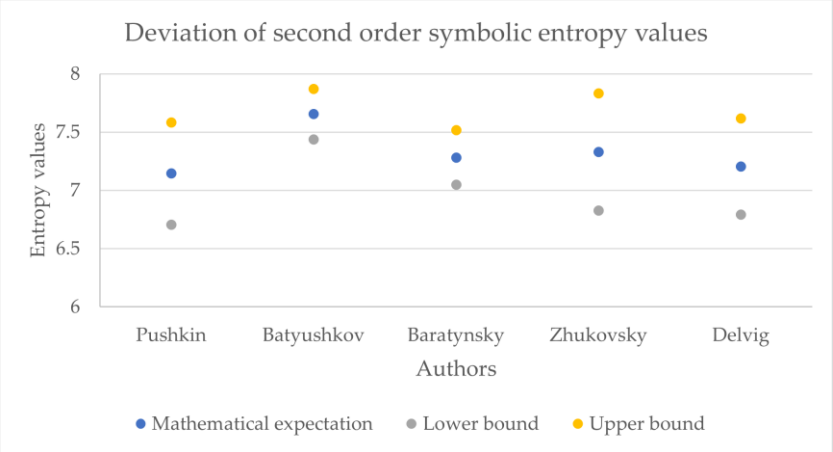

Figure 5. Second-order symbolic entropy values.

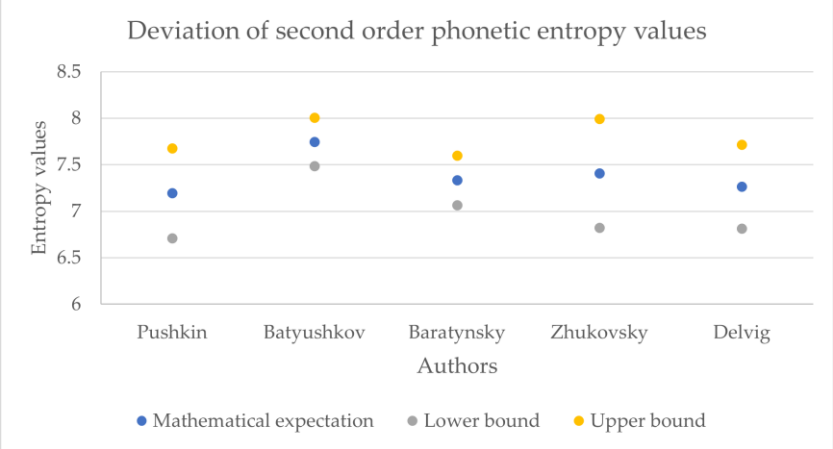

Figure 6. Second-order phonetic entropy values.

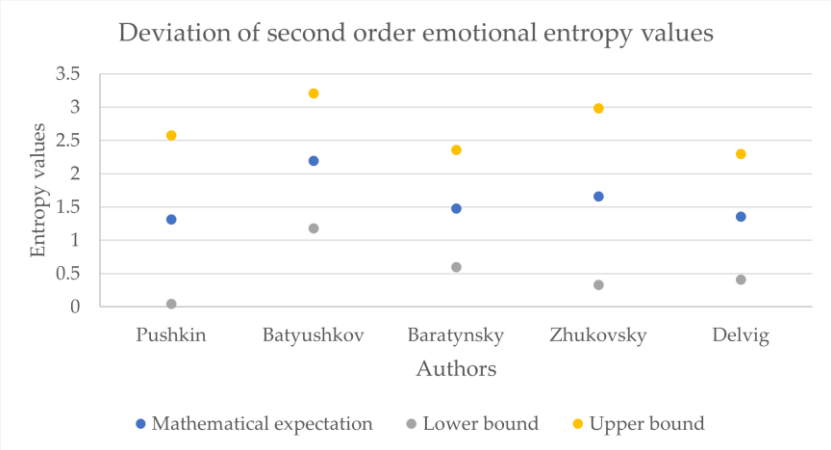

Figure 7. Second-order emotional entropy values.

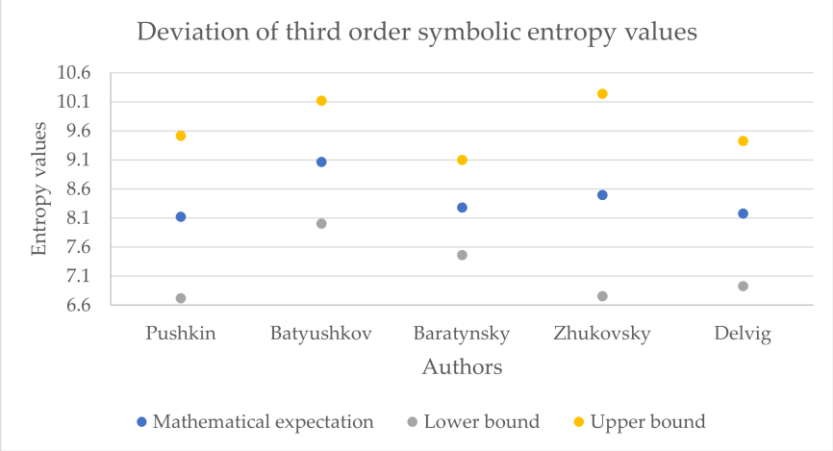

Figure 8. Third-order symbolic entropy values.

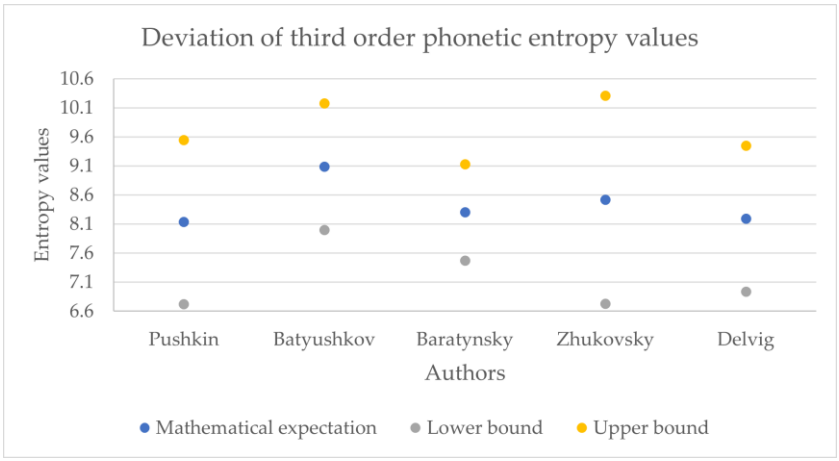

Figure 9. Third-order phonetic entropy values.

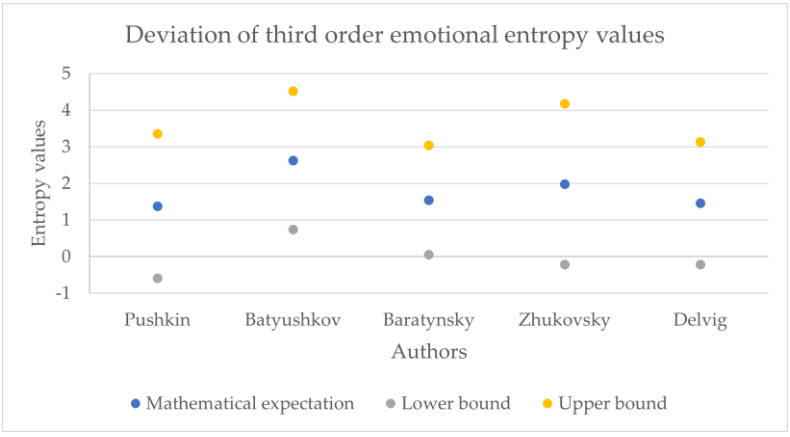

Figure 10. Third-order emotional entropy values.

Figures 11–19 show the boxplot diagrams of the entropy span, on which the median, upper and lower quartile boundaries and outliers' limiters can be found. The values of the latter, in this case, were calculated using the formula 1.5 * IQR, where IQR = $Q_{25} - Q_5$ for the lower bound, and IQR = $Q_{95} - Q_{75}$ for the upper bound (here Q_n is the value of the *n*-th percentile).

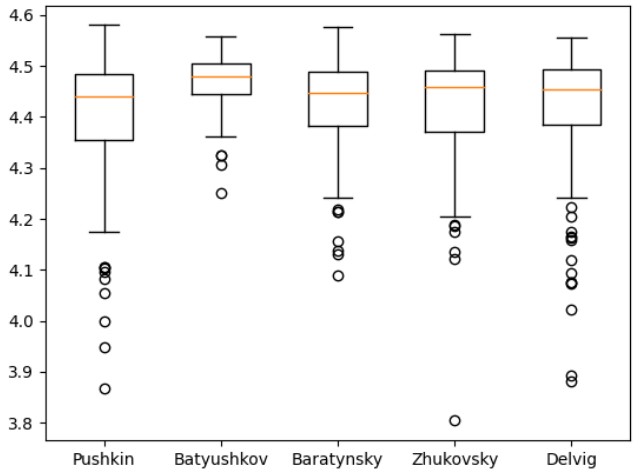

Figure 11. Boxplot of the first-order symbolic entropy span.

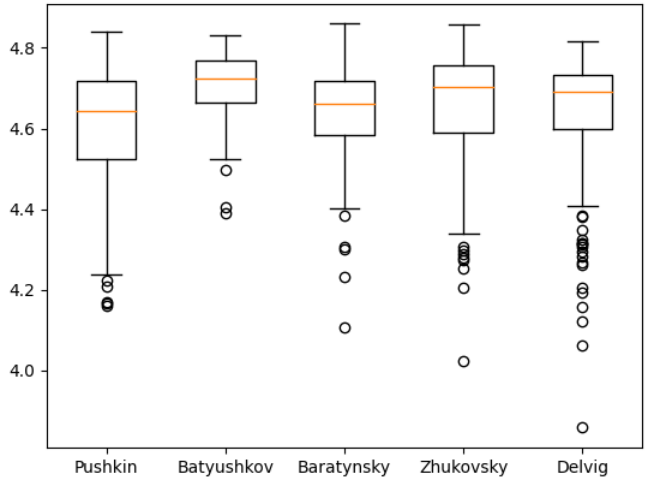

Figure 12. Boxplot of the first-order phonetic entropy span.

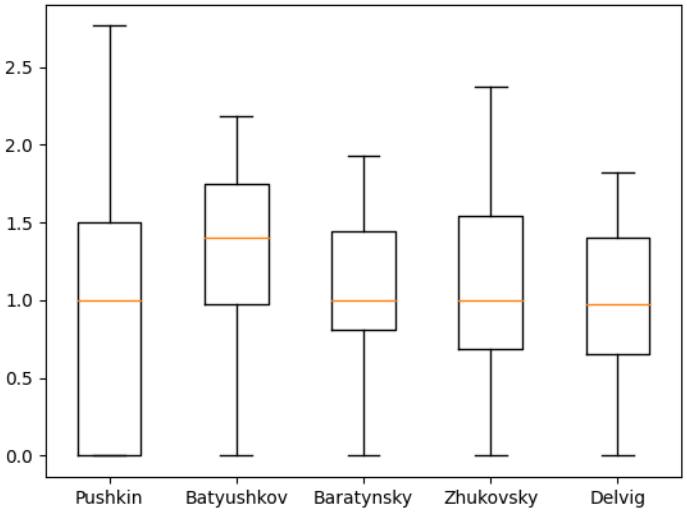

Figure 13. First-order emotional entropy values.

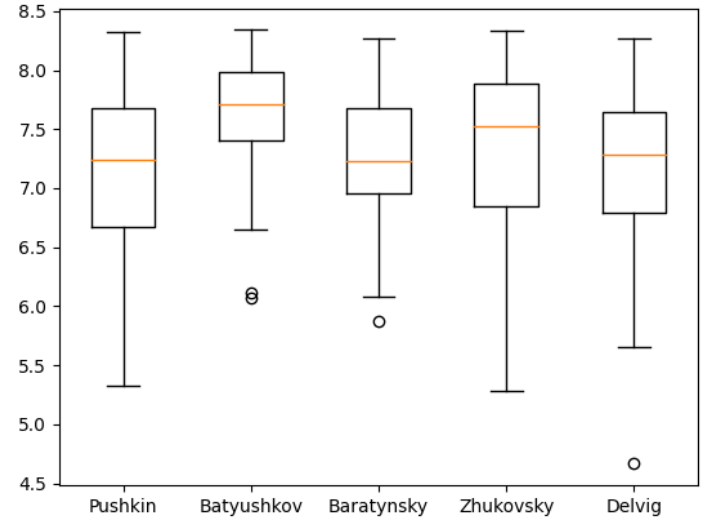

Figure 14. Boxplot of the second-order symbolic entropy span.

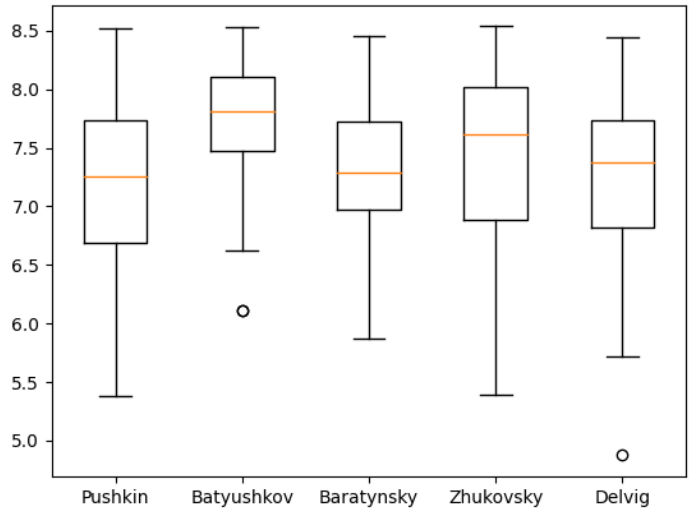

Figure 15. Boxplot of the second-order phonetic entropy span.

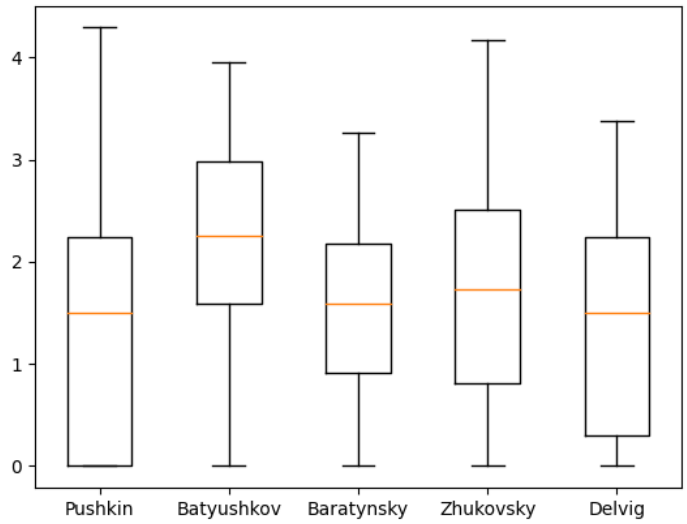

Figure 16. Boxplot of the second-order emotional entropy span.

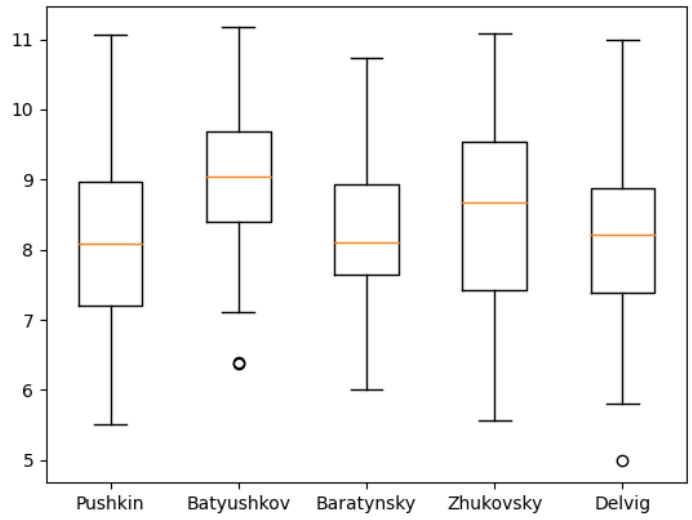

Figure 17. Boxplot of the third-order symbolic entropy span.

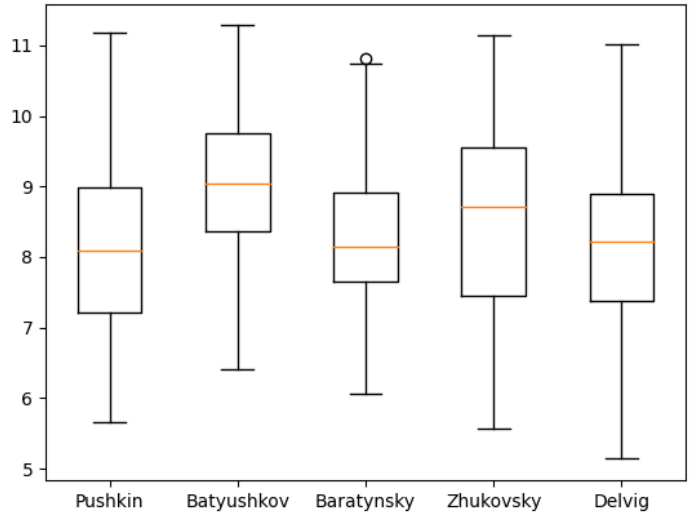

Figure 18. Boxplot of the third-order phonetic entropy span.

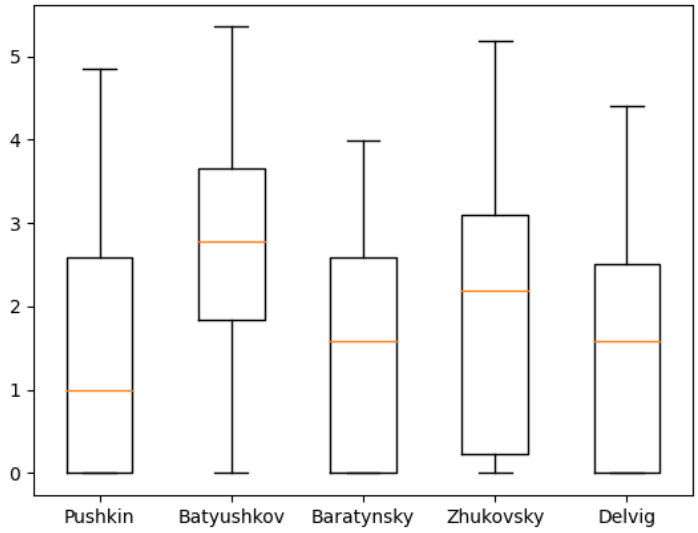

Figure 19. Boxplot of the third-order emotional entropy span.

6. Discussion

We estimated the difference between the values of letter and numeric entropy (in Tables 1–5, the phonetic entropy always exceeds the symbolic entropy). As noted above, for Russian poetry, the phonetic transcription of a poetic text reflects the author's intention more adequately compared to its symbolic notation; therefore, when calculating the relative error, the phonetic entropy can be considered as an exact value, while the symbolic entropy can be considered as an approximation. Table 6 presents the relative error values of the symbolic entropy mathematical expectation.

Table 6. Relative error of symbolic entropy mathematical expectation, %.

Order	1st Order	2nd Order	3rd Order
Pushkin	4.359017	0.652964	0.178770
Batyushkov	5.068891	1.161830	0.251289
Baratynsky	4.527102	0.682906	0.191871
Zhukovsky	4.932839	1.031219	0.221830
Delvig	4.629639	0.764526	0.136160

Thus, the relative error for the 1st order entropy was about 5%, the 2nd order was about 1%, and the 3rd was about 0.2%; therefore, counting the phonetic entropy instead of the symbolic entropy has a certain effect on the accuracy of the analysis. This influence can hardly be called fundamental and could signify the general difference between the written language and spoken language; additional studies on prose and non-classical poetry could offer deeper insights regarding this statement.

However, these values may also offer a deeper understanding of authors who are more exact in the textual representations of their work, like Pushkin, and who are less exact, like Batyushkov. In particular, it can be noted that the symbolic and phonetic entropy of the 1st and 2nd order of Batyushkov's works had a much smaller standard deviation value than the same entropy of other authors. This was an unexpected result compared with other authors despite being in good agreement with the characteristics of his work described in the Big Russian Encyclopedia [19]: "The musical harmony of verse, created by numerous alliterations and gaping, combined with the refinement of the poetry form, is the characteristic sign of the Batyushkov's style". From these values, it can be seen that the strictness of the poetic form provides a relatively small spread of entropy values.

Another unexpected result is the relatively small value of the lower quartile of the 1st and 2nd order of emotional entropy in Pushkin's works. Despite being unexpected due to the overall perception of Pushkin's works as emotional, similar results were obtained in the study of his works using machine learning methods: "Pushkin, unlike other authors, preferred the usage of dots and ellipses rather than exclamation marks, which differs him from the other authors. Additionally, according to the study of Pushkin's word usage, Pushkin was also not inclined to use the interjection «ax»; therefore, it can be noted that the other poets of Pushkin's era were much more emotional and wrote with a more pronounced enthusiastic syllable" [20].

7. Conclusions

Within the framework of this work, the calculations of entropy values of the first, second and third order for the corpus of poems by A.S. Pushkin and other poets of Pushkin's era were carried out. The calculations were made for the symbolic notation of the texts of poems with punctuation marks eliminated for the phonetic transcription of the texts of poems and for sequences of punctuation marks in poems (dot, ellipsis, question mark, etc. as indivisible units of punctuation), which reflected the emotional coloring of sentences in the text to varying degrees. The mathematical expectation and variance of this type of entropy calculated in each series were obtained. The calculations were carried out to find out the significance of entropy characteristics for describing the author's style. The software application was implemented to automatically extract statistical information,

which could be potentially applicable in the task of identifying features of the author's style (regardless of language or text style). The statistical data were extracted from the poems of A.S. Pushkin and other authors. These data could become the basis for the stylometric classification of authors using entropy characteristics. The confirmation of some features of stylometric characteristics of Russian poets in Pushkin's era, established earlier by empirical observations or using machine learning methods, was obtained. Thus, there are serious reasons to suppose that the use of entropy methods for the stylometric analysis of poetic texts, due to the consideration of the text as a whole, allows important results to be obtained in relation to the phonetic and emotional characteristics not only of Russian classic poetry but potentially for other texts if integrated with other methods of textual analysis.

Author Contributions: Conceptualization, O.K. and V.B.; methodology, O.K. and V.B.; software, N.S.; validation, V.B.; formal analysis, N.S.; investigation, N.S. and E.K.; resources, O.K. and V.B.; data curation, N.S. and E.K.; writing—original draft preparation, O.K., V.B., N.S. and E.K.; writing—review and editing, O.K.; visualization, N.S.; supervision, O.K.; project administration, O.K.; funding acquisition, V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed can be found in [1].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. FEB-Web: The Fundamental Digital Library of Russian Literature & Folklore. Available online: http://feb-web.ru/indexen.htm (accessed on 20 September 2023).
2. Markov, A.A. An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains (Trans.). *Sci. Context* **2006**, *19*, 591–600. [CrossRef]
3. Drozdova, I.I.; Obukhova, A.D. Determining the authorship of a text by frequency characteristics. In Proceedings of the VII International Scientific Conference "Technical Sciences in Russia and Abroad", Moscow, Russia, 20–23 November 2017.
4. Kolmogorov, A.N.; Prokhorov, A.V. On the basics of Russian classical metrics. In *The Commonwealth of Sciences and the Secrets of Creativity*; Iskusstvo: Moscow, Russia, 1968; pp. 397–432.
5. Kolmogorov, A.N. *Works on Poetry*; Prokhorov, A.V., Ed.; MCCME: Moscow, Russia, 2016.
6. Khmelev, D.V. Classification and Markup of Texts Using Data Compression Methods. Available online: http://compression.ru/download/articles/classif/intro.html (accessed on 30 August 2023).
7. Gogoleva, V.A.; Shkaraputa, A.P. The mathematical approach to the determination of authorship and time of creation of a text based on the study of its entropy. *Bull. Perm Univ. Math. Mech. Comput. Sci.* **2014**, *4*, 22–28.
8. Poltavsky, A.V.; Rusyaeva, E.Y. Entropic foundations of machine translations and text analysis in a computer network. In Proceedings of the International Symposium "Reliability and Quality", Penza, Russia, 24–31 May 2021.
9. Mansilla, R.; Bush, E. Increase of complexity from classical Greek to Latin poetry. *Complex Syst.* **2003**, *14*, 201–213.
10. Rényi, A. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley and Los Angeles, CA, USA, 20 June–30 July 1960.
11. Calin, O. Statistics and machine learning experiments on English and Romanian Poetry. *Sci* **2020**, *2*, 92. [CrossRef]
12. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
13. Lippi, M.; Montemurro, M.A.; Degli Esposti, M.; Cristadoro, G. Natural Language Statistical Features of LSTM-Generated Texts. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3326–3337. [CrossRef] [PubMed]
14. Ackerman, J. Entropy of Sounds: Sonnets to Battle Rap. In Proceedings of the 42th Annual Meeting of the Cognitive Science Society—Developing a Mind: Learning in Humans, Animals, and Machines, CogSci, Virtual, 29 July–1 August 2020.
15. Parlar, T.; Özel, A.S.; Song, F. Analysis of data pre-processing methods for the sentiment analysis of reviews. *Comput. Sci.* **2019**, *20*, 123. [CrossRef]
16. Lotman, Y.M. *The Structure of a Literary Text*; Iskusstvo: Moscow, Russia, 1970.
17. Orlov, Y.N.; Osminin, K.P. *Methods of Statistical Analysis of Literary Texts*; LIBROCOM Book House: Moscow, Russia, 2017.
18. Orlov, Y.N.; Osminin, K.P. Definition of the genre and author of a literary work by statistical methods. *Appl. Inform.* **2010**, *2*, 95–108.

19. Yurchenko, T.G. Batyushkov. In *Big Russian Encyclopedia*; Osipov, Y.S., Ed.; Big Russian Encyclopedia: Moscow, Russia, 2005; Volume 3, p. 109.
20. Barakhnin, V.; Kozhemyakina, O.; Grigorieva, I. Determination of the features of the author's style of Pushkin's poems by machine learning methods. *Appl. Sci.* **2022**, *12*, 1674. [CrossRef]