P.G. Demidov Yaroslavl State University

as a manuscript

Lagutina Ksenia

AUTOMATIC ANALYSIS OF RHYTHM FEATURES OF TEXTS IN NATURAL LANGUAGE

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Yaroslavl — 2021

The PhD dissertation was prepared at P.G. Demidov Yaroslavl State University

Academic Supervisor: Valery A. Sokolov, Doctor of Sciences, P.G. Demidov Yaroslavl State University

PhD Dissertation Relevance. Stylometry is a branch of computational linguistics that studies the quantification of linguistic features in natural language texts. Stylometry is closely related to the definition of an author's individual style and idiolect that are a system of linguistic features used by the author [1]. Distinctive features of the style in which the text is written can be formalized, algorithms for their automatic determination can be developed and used for the tasks of authorship verification, classification of natural language texts by publication date or genre, as well as for statistical analysis of the texts features. The search for the stylometric text features is quite laborious and requires a significant amount of time for manual processing, so it needs to be automated.

The choice of stylometric features of the text is the most important stage of the investigation. Researchers identify about a thousand features at different levels of analysis: lexical (including levels of characters and words), syntactic, semantic, structural, and subject-oriented [2; 3]. This indicates the complexity and versatility of the text, therefore, it is necessary to evaluate the text units selected for quantitative analysis and their ability to express the originality of the author's style.

Today there is no consensus on how to choose the optimal stylometric features for solving any of the problems of text classification or analysis. Most state-of-the-art researchers use practically the same set of standard features that model text at the level of words and characters, and to improve the quality of solving the task, they pay attention to methods for reducing the dimension of feature vectors and the selection of classifiers and their parameters. In contrast, in classical linguistics, researchers concentrate on complex linguistic features when analyzing the text style. Therefore, the search and analysis of new style features is an actual task of computational linguistics.

One of the important aspects of the specifics of the text style is the rhythm. Rhythm is defined as a regular repetition of similar and commensurable units of speech, that performs structuring, text-forming and expressive functions [4]. In classical linguistics, the main goal of rhythm analysis is a deep penetration into the author's creative method, his/her design, the originality of individual creativity and skill, therefore, identifying the specifics of the rhythm of literary texts allows to solve the problem of determining the individual author's style more successfully. This method is used in the analysis of poetic texts, while its application to prose fiction is understudied [5]. In particular, algorithms for rhythm feature searching are difficult to formalize, and there are no available software tools for their extraction. Therefore, the development of automated tools for analyzing rhythm in a prose text and their testing in the text classification and analysis is one of the important tasks of natural language processing.

PhD Dissertation Goal is the development and study of a complex of rhythm features of the text and their comparison with standard stylometric features in solving problems of text classification by authors and publication periods.

To achieve this goal, it is necessary to solve the following **tasks**.

1. Development of algorithms and the software tool for automatic search for rhythm features in prose texts.

2. Statistical analysis of the dynamics of changes in rhythm features in prose texts of the XIX–XXI centuries.

3. Classification of literary texts of the XIX–XXI centuries by centuries and half a century of their publication on the basis of rhythm and standard stylometric features.

4. Verification of authors of literary texts of the XIX–XXI centuries based on rhythm and standard stylometric features.

The object of research is prose texts in natural language.

The subject of research is a complex of rhythm features of the text.

Methodology and research methods. The methodology of the dissertation research is based on the formulation and formalization of goals and tasks, the development of text models, methods and algorithms for text analysis, experimental evaluation using statistical experiments, testing and analysis of results. To solve the set tasks, the methods of automatic preprocessing of texts, searching for statistical and lexico-grammatical features of the text were used. The analysis of the dynamics of the rhythm of the texts was carried out using statistical metrics and methods of their visualization. The classification of texts was carried out on the basis of machine learning methods and neural networks.

The following **key aspects** that have scientific **novelty**, are submitted for defense.

1. Algorithms have been developed for automatic search and visualization of lexical and grammatical rhythm features in prose texts for Russian, English, French, and Spanish.

2. A complex of numerical rhythm features for a prose text has been developed. Its suitability for carrying out volumetric experiments is demonstrated on the example of statistical analysis of the dynamics of changes in rhythm features for prose texts of the XIX–XXI centuries.

3. The effectiveness of the use of rhythm features for the classification of literary texts of the XIX–XXI centuries by centuries and half a century of their publication is shown. The comparison of rhythm and standard stylometric features for solving this problem is carried out.

4. It is shown that the rhythm features in terms of the quality of the authorship verification correspond to standard features, and in combination with them, they increase the effectiveness of verification of authors of literary texts of the XIX–XXI centuries.

The **Practical Value.** The results of research on the authorship verification and the classification of texts by centuries of publication show that a text model based on rhythm features can be successfully used to solve problems of classification of literary texts. A software tool based on the proposed algorithms for searching and visualizing rhythm features, developed under the guidance of the author of this dissertation, is useful for linguistic experts to automate their investigations and reduce the time for voluminous routine work in research.

Approbation of the work. The main results of the work were reported at international scientific conferences:

1. "AIST 2019 — The 8th International Conference on Analysis of Images, Social Networks and Texts" (Kazan, Russia, 2019);

2. "The 25th Conference of Open Innovations Association FRUCT" (Helsinki, Finland, 2019);

3. "The 26th Conference of Open Innovations Association FRUCT" (Yaroslavl, Russia, 2020);

4. "The 27th Conference of Open Innovations Association FRUCT" (Trento, Italy, 2020);

5. "The 28th Conference of Open Innovations Association FRUCT" (Moscow, Russia, 2021);

6. "The 29th Conference of Open Innovations Association FRUCT" (Tampere, Finland, 2021).

Personal contribution. The content of the dissertation and the key aspects for the defense reflect the personal contribution of the author to the published works. From the works performed in co-authorship, the dissertation includes results that correspond to the personal participation of the author.

Publications. The main results on the topic of the dissertation are presented in the following publications.

Second-tier publications

1. Lagutina N.S. Automated Search of Rhythm Figures in a Literary Text for Comparative Analysis of Originals and Translations Based on the Material of the English and Russian Languages. / Lagutina N.S., Lagutina K.V., Boychuk E.I., Vorontsova I.A., Paramonov I.V. //Modeling and Analysis of Information Systems. – 2019. – Vol. 26. – No.. 3. – pp. 420-440. In Russian. (list of journals recommended by HSE)

2. Lagutina K.V. Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th-21st Centuries / Lagutina K.V., Manakhova A.M. //Modeling and Analysis of Information Systems. – 2020. – Vol. 27. – No.. 3. – pp. 330-343. In Russian. (list of journals recommended by HSE) — the main co-author.

3. Lagutina N. S. Automated Rhythmic Device Search in Literary Texts Applied to Comparing Original and Translated Texts as Exemplified by English to Russian Translations / Lagutina, N. S., Lagutina, K. V., Boychuk, E. I., Vorontsova, I. A., Paramonov, I. V. //Automatic Control and Computer Sciences. – Springer, 2020. – Vol. 54. – No.. 7. – pp. 697-711. (Scopus, Q3)

4. Lagutina K. V. Comparison of Style Features for the Authorship Verification of Literary Texts / Lagutina K. V. //Modeling and Analysis of Information Systems. – 2021. – Vol. 28. – No.. 3. – pp. 250-259. (list of journals recommended by HSE)

5. Lagutina K.V. Text Classification by Genre Based on Rhythm Features / Lagutina K.V., Lagutina N.S., Boychuk E.I. //Modeling and Analysis of Information Systems. – 2021. – Vol. 28. – No.. 3. – pp. 280-291. In Russian. (list of journals recommended by HSE) — the main co-author.

Other publications

6. Lagutina K. A Survey on Stylometric Text Features / Lagutina K., Lagutina N., Boychuk E., Vorontsova I., Shliakhtina E., Belyaeva O., Paramonov I. // Proceedings of the 25th Conference of Open Innovations Association FRUCT, IEEE, 2019 – Vol. 25. – No. 1. – pp. 214-219. (Web of Science, Scopus) — the main co-author.

7. Boychuk E. Automated Approach to Rhythm Figures Search in English Text. /Boychuk E., Vorontsova I., Shliakhtina E., Lagutina K., Belyaeva O. // International Conference on Analysis of Images, Social Networks and Texts. CEUR Workshop Proceedings. Springer, Cham CCIS, Vol. 1086, 2020. pp. 107-119. (Web of Science, Scopus)

8. Lagutina K. Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries / Lagutina K., Poletaev A., Lagutina N., Boychuk E., Paramonov I. // Proceedings of the 26th Conference of Open Innovations Association FRUCT. IEEE, 2020. – Vol. 26. – No. 1. – pp. 247-255. (Web of Science, Scopus) — the main co-author.

9. Lagutina K. The Influence of Different Stylometric Features on the Classification of Prose by Centuries / Lagutina K., Lagutina N., Boychuk E., Paramonov I. //Proceedings of the 27th Conference of Open Innovations Association FRUCT. – IEEE, 2020. – Vol. 27. – No. 1. – C. 108-115. (Web of Science, Scopus) — the main co-author.

10. Boychuk E. Evaluating the Performance of a New Text Rhythm Analysis Tool / Boychuk E., Lagutina K., Vorontsova I., Mishenkina E., Belyayeva O. // English Studies at NBU. – New Bulgarian University, 2020. – Vol. 6. – No.. 2. – pp. 217-232. (Web of Science)

11. Lagutina K. Authorship verification of literary texts with rhythm features. / Lagutina K., Lagutina N., Boychuk E., Larionov V., Paramonov I. // Proceedings of the 28th Conference of Open Innovations Association FRUCT, IEEE, 2021 – Vol. 28. – No. 1. – pp 240–251. (Web of Science, Scopus) — the main co-author.

12. Lagutina K. A Survey of Models for Constructing Text Features to Classify Texts in Natural Language. / K. Lagutina, N. Lagutina. // Proceedings of the 29th Conference of Open Innovations Association FRUCT. – IEEE, 2021 – Vol. 29. – No. 1. – pp. 222-233. (Scopus) — the main co-author.

Certificates of official registration of computer programs:

1. The program that implements an automated algorithm for analyzing the rhythm of a text based on phonetic, lexical-grammatical and structural-compositional parameters of the rhythm for texts in Russian, English, and French / Ratnikov E.S., Tumanova A.D., Boychuk E.I., Lagutina N. S., Lagutina K. V. // The certificate of official registration of computer programs No. 2019619380, 16.07.2019.

2. The program for statistical analysis of stylometric and rhythm features of texts in Russian, English, French, and Spanish / Manakhova A. M., Lagutina K. V., Lagutina N. S. // The certificate of official registration of computer programs No. 2020618648, 30.07.2020.

3. The program for automatic search of stylometric features of various levels from texts and classification of texts by author / Lagutina K.V. // The certificate of official registration of computer programs No. 2021616718, 26.04.2021.

4. The software prototype for automatic identification of the quality parameters of the style of texts / Lagutina K.V. // The certificate of official registration of computer programs No. 2021664205, 1.09.2021

5. The software prototype for processing the rhythm of texts, comparative analysis of rhythm in their translation and authorization of texts / Lagutina K.V., Lagutina N.S., Boychuk E.I. // The certificate of official registration of computer programs No. 2021664248, 2.09.2021.

The PhD dissertation was funded by RFBR according to the research project No.20-37-90045.

Content of work

The **introduction** substantiates the relevance of the research carried out within the framework of this dissertation, formulates the goal, lists the work tasks, sets out the scientific novelty and practical significance of the work presented, provides new scientific results to be defended.

The **first chapter** is devoted to the survey and analysis of stylometric features in the text used for authorship attribution, authorship verification, authors profiling, style changes detection and classification of texts by genre and sentiment. Methods for solving these problems are based on the assumption that it is possible to identify the features of the text that confirm authorship [6].

Stylometric features can be divided into two categories: simple statistical features, to calculate which the text is considered as a set of characters or words, and complex linguistic features, whose search requires knowledge of the language.

Simple statistical (or standard) features include character- and word-level features. As the analysis of state-of-the-art literature shows, they are the easiest and fastest to calculate and are used much more often than others [7].

At the character level, the text is presented as a sequence of characters, whereas the features themselves present the simplest document structure. N-gram defined as a contiguous sequence of n items from a given sample of text is a regular characteristic at the character level.

At the word level, the text is often seen as a bag-of-words regardless of the word order, grammar or context. In such case word frequency, word character length, average word length, word n-grams and vocabulary richness are measured.

Standard stylometric features also include character and word embeddings that are based on the simple statistical features described above.

Complex linguistic features include syntactic, rhythm, topical, semantic and other features.

Syntactic features are based on sentence structure. Punctuation mark frequency, sentence length, average sentence length, and functional word frequency are among the simplest and most common. More complex characteristics include syntactic tree features.

Text rhythm include lexical and grammatical features, for example, anaphora, epiphora, or aposiopesis, based on the repetition of words or punctuation marks, and also phonetic features, for example, alliteration and assonance, based on the repetition of sounds.

Topical features are based on extracting keywords and analyzing their occurrence.

Semantic features are based on the relationship between words: synonymous, associative, etc.

Thus, the number of stylistic features used in computer linguistics is very large and heterogeneous. However, researchers pay insufficient attention to systematization of these features, study of their influence on the quality of solving tasks and justification of feature choice. Most authors experimentally compare algorithmic approaches like [8]. Much less often, researchers set the task of studying the influence of various parameters on the quality of text classification by the author's style [9]. Almost none of the researchers consider the reasons why features or feature groups are relevant and efficient.

Comparing studies with the highest quality scores (about 90 % and higher) of algorithms with different feature categories, we can conclude that these results are most often achieved under one or more of the following conditions:

— a relatively small text corpus (not more than 200–250 texts), and the texts are quite voluminous in size;

— texts belong to a small number of authors, usually 10 or less;

— a large number of texts of a given author is analyzed, then one of the best classification results is obtained for this author;

— researchers successfully selected stylistic features according to which the classifier makes decisions, and the features may differ for texts with different topics and genres.

In addition, researchers most often take into account only some features of the idiolect or the linguistic specificity of an author's style, which consist, as a rule, in reflecting quantitative indicators of rather low-level text features, such as the number of words, syllables, sentence size, etc. However, the idiostyle is expressed in features that are rather complicated for the search and related to the personality of the author. The added complexity is that “there is no taxonomy or checklist of the elements of individual style, since anything can be an element of individual style if it is consistently used in such a way as to contribute to the expression of the personality of the author” [10].

Thus, the implementation of a comprehensive analysis of an author's individual style is a rather difficult task. The analysis of author's language specificity is only one of its many stages. Automating the search for these formal features is the first step towards a comprehensive understanding of an individual author's style.

The **second chapter** is devoted to the development of algorithms for finding rhythm features and their implementation in the application called ProseRhythmDetector. This software tool is designed to automatically identify repetitive lexical and grammatical figures and visualize them.

Existing tools turn out to be focused on the analysis of the rhythm of the text at the phonetic, lexical and/or syntactic levels or on the solution of a specific problem with practically no analysis of intermediate steps and linguistic interpretation [11; 12]. The novelty of the ProseRhythmDetector tool lies in its ability to search and process stylistic figures based on repetition, as well as to visualize them, providing the opportunity for a linguist to study both the text rhythm as a whole and its individual aspects.

In this work, the rhythm figures of the text are determined on the basis of the repetition of words and punctuation marks in a certain configuration, in a certain position, with a certain number of repeating elements, in accordance with their definitions in classical linguistics. ProseRhythmDetector finds the following rhythm figures:

1. Anaphora "— a repetition of sequence of words at the beginning of neighboring sentences. For example, “**I wanted** a miracle job advertisement. **I wanted** someone to come along and say”.

2. Epiphora "— a repetition of the same word or words at the end of neighboring sentences (also called epistrophe). For example, “Frank **knew**. And Maxim did not know that he **knew**”.

3. Symploce "— a repetition of the beginning and the end of two or more neighboring sentences, combination of anaphora and epiphora. For example, “**I'm** wanting **to tell you**. **I'm** waiting **to tell you**”.

4. Anadiplosis "— a repetition of the same word at the end of a clause and at the beginning of the following clause. For example, “It was right to do **it, it** was kind to do **it, it** was benevolent to do it, and he would do it again”.

5. Epanalepsis "— a repetition of the initial part of a sentence at the end of the same sentence. For example, “**The king** is dead, long live **the king**”.

6. Polysyndeton "— a repetition of the same conjunction within one sentence (simple and pair conjunctions and conjunctive adverbs can be repeated). For example, “There were frowzy fields, **and** cow-houses, **and** dunghills, **and** dustheaps, **and** ditches”.

7. Diacope "— a repetition of a word or phrase with intervening words within one sentence. For example, “**Help**, Charmian, **help**, Iras”.

8. Epizeuxis "— a repetition of a word or phrase in immediate succession within one sentence. For example, “**Weak**! **Weak**! **Weak**!”.

9. Chiasmus "— a reversal of grammatical structures in successive phrases or clauses with the repetition of words. For example, “**You forget** what **you** want to **remember**, and **you remember** what **you** want to **forget**”.

10. Aposiopesis "— a figure of speech in a sentence which is deliberately broken off and left unfinished. For example, “She resurrected nothing but the cat ... but the cat ...”.

11. Repeating interrogative sentences "— a repetition of the interrogative point at the ending of neighboring sentences. For example, “Where’s my car? Where’s my house?”.

12. Repeating exclamation sentence "— a repetition of the exclamation point at the ending of neighboring sentences. For example, “Jeepers! You scared the life out of me!”.

The choice of these figures for analyzing rhythm, especially for their automated search and quantitative processing, is due to the fact that these are rhythm figures used in prose texts most often, and they that stand out as rhythm features at the lexical and grammatical level by most linguists conducting research in areas of the text rhythmization.

In order to analyze the rhythm of a prose text, a complex of algorithms was developed that automatically find rhythm figures in the text, namely, lexical and syntactic ones.

The input data for all algorithms is a plain text. Each text is presented as a set of ordered sentences, consisting of words and punctuation marks. In the process of running the algorithms, sentences are sequentially sorted out, repetitions of words and punctuation marks are highlighted. Repetitions of these elements that fit the definitions of rhythm figures are included into the aspect lists. These lists are returned as output.

The precision of the search algorithms was computed by experts in classical linguistics manually. Four researchers processed a total of 24 texts of different authors, randomly selected from the corpus. Each expert worked 16 hours. She manually evaluated precision of search for all rhythm figures. An exception is the diacopa, because ProseRhythmDetector found several thousand rhythm aspects for it, so the experts checked only random 10 % of them. The experts concluded that the accuracy of figure search reaches 80-95 % for all rhythm figures.

The ProseRhythmDetector application is implemented in the Python programming language using the Stanza text processing library. It is available online: `https://github.com/text-processing/prose-rhythm-detector`.

Thus, this tool allows to identify rhythm features quickly, accurately, and completely automatically, even for large texts. This significantly speeds up the work of an expert-linguist when comparing the author's style of texts and allows large-scale experiments to analyze the rhythm of large text corpora, which would be practically impossible without such automatization.

The **third chapter** is devoted to the study of how a complex of rhythm features can be used for automated experiments with the analysis of the author's style in prose. The researcher carries out an automatic search for these features in fiction and a statistical analysis of their appearance in the XIX–XXI centuries. The rhythm features are compared with the standard features of the word and character levels. Experiments are carried out with the tool search algorithms implemented in the ProseRhythmDetector tool.

For rhythm figures, the following numerical stylometric features were chosen:

- the number of occurrences of a particular figure (anaphora, epiphora, etc.) divided by the number of sentences in a text;

- the number of all rhythm figures divided by the number of sentences in a text;

- hapax legomenon—the fraction of unique words among all words that appear in rhythm figures, in this case, those that are repeated only once;

- the fractions of words of a particular part of speech: noun, verb, adverb, and adjective—among all words that appear in rhythm figures.

The choice of these features for analyzing the rhythm, namely, for their automated search and quantitative processing, is due to the fact that they stand out at the lexical and grammatical level as rhythmic means by most linguists conducting research in the field of text rhythmization.

The following characteristics were selected as stylometric characteristics at the character and word levels.

Character-based features:

- the number of letters, both individual and their total number;
- the number of punctuation marks, both individual and their total;
- average length of a sentence in characters.

Word-based features:

- average sentence length in words;
- average word length;
- frequencies of top-40 n-grams for $n = 1, 2, 3$. For each unigram, bigram, or trigram we calculate the number of occurrences in a text corpus, then we choose the most frequent 40 unigrams, bigrams, and trigrams. For each text we also compute their numbers of occurrences and divide them by the total number of occurrences of these 120 n-grams in the text.

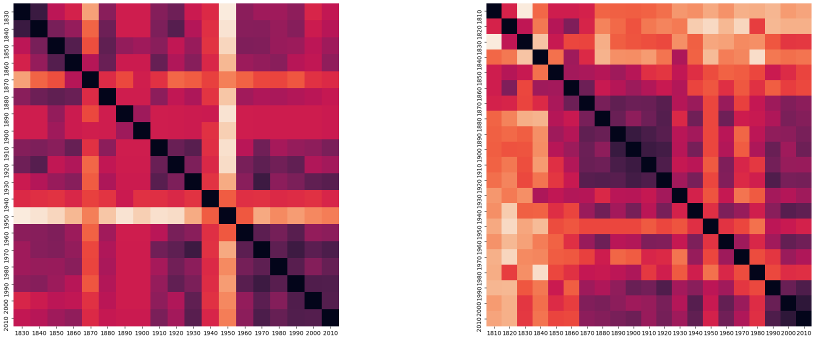
The choice of these stylometric features at the levels of characters and words was due to the fact that they are the most indicative in determining the author's style during the study of the text.

Stylometric characteristics of three different levels are calculated and visualized automatically. Experiments with these characteristics were set up as follows.

- At first, rhythm figures were identified in the texts.
- Stylometric features were calculated for the identified rhythm figures.
- In parallel with the calculation of the rhythm features for the texts, the stylometric features of the levels of words and characters were calculated.
- Stylometric features of the texts were aggregated by decades, the decades were compared with each other.
- In the last step, the comparison results were visualized using heatmaps and graphs.

Experiments were carried out with text corpora in English, Russian, French, and Spanish.

Each corpus includes 240 texts by more than 90 famous authors. Each of the texts is marked by the publication date from 1815 to 2019. Each text contains up to 425 000 words.



а)

б)

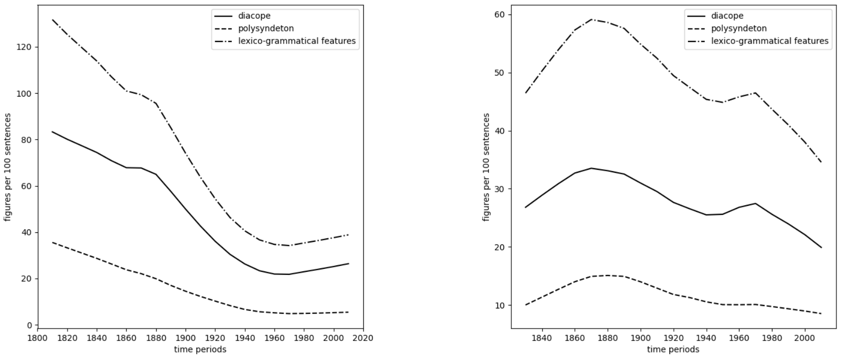
Fig. 1 — All rhythm features by decades, Chebyshev distance, for a) Russian, b) English languages

Heatmaps show large clusters of similar rhythm with texts of 19th century and the end of 20th–beginning of the 21st century. Besides, we can see small clusters with 2–3 decades with close rhythm features.

If we compare plots and heat maps for decades, we can conclude that the Chebyshev distance works well and highlight clusters when the quantity of the figures is quite large. For the 21st century, when figures appear more rarely, this measure is not useful.

Thus, the heat maps and the plots reveal the tendencies in the figure use over decades and centuries, so rhythm figures can be helpful indicators of style changes.

Quantitative analysis allows to divide the set of rhythm figures into two groups by their frequencies of occurrence: frequent (diacope, polysyndeton) and rare (anaphora, epiphora, anadiplosis). The identified decrease of the total number of figures corresponds to the first group. The amount of rare figures does not show such a result. Therefore, we can conclude that the most common rhythm figures are the most useful for determining the time of writing a text.



а) б)

Fig. 2 — Rhythm features by decades: all, diacope, polysyndeton in a) English, b) Russian languages

Experiments have shown that, although decades can be successfully clustered by proximity to each other, each of them is unique in terms of the combination of rhythm and simple stylometric features. This means that the model based on these features can be successfully used to classify texts by centuries and decades of creation/publication.

The **fourth chapter** is devoted to the two research tasks: (1) the automatic classification of fiction of the XIX–XXI centuries by periods of their publication using rhythm features and (ii) the comparison of classification quality of three types of stylometric features: character-based, word-based, and rhythm-based. Such classification can provide the explanation of changing and evolving writing styles [13; 14].

The model of the text, that is, the set of its stylometric features, was taken the same as in the previous chapter: features of the levels of characters, words and rhythm.

Stylometric features form a text style model that can be used to classify texts. Texts are classified into three classes according to the date of their publication: XIX, XX, and XXI centuries.

First, rhythm figures are extracted from natural language texts using the algorithms from the second chapter. Search precision is 80–95 %. Then the rhythm features are calculated separately for each text.

In parallel features of the levels of characters and words are calculated for texts. The algorithm finds the most frequent n -grams in the corpus, then independently for each text calculates the frequency of occurrence of the top n -grams and other low-level features.

After calculating the features for each text, the results are combined into a common matrix. Thus, each text is presented as a vector of numerical stylometric features.

The vectors of stylometric features are given as inputs to four supervised classifiers: AdaBoost, RandomForest, Bidirectional LSTM, a neural network GRU. These four algorithms often demonstrate the high quality of the text classification [15; 16], they were chosen for experiments.

All the algorithms are trained on a half of a text corpora. The text corpora sizes are not large, that is why classifiers are tested on the significant fraction of samples. For neural networks training the author applies categorical cross-entropy as a loss function and Adam as an optimization algorithm.

The results of the test phase of multi-class classification were evaluated with four common measures: accuracy, macro-average precision, recall, and F-score.

The author experimented with four corpuses of texts in English, Russian, French, and Spanish languages from Chapter 3.

For all languages neural networks outperform Random Forest and AdaBoost meta-classifiers. The accuracy and F-measure for meta-classifiers is less than 80 %, while neural networks provide from 82 to 89 % in the best cases.

Tables 1, 2, 3, and 4 present the results of the classification by centuries for individual feature types and their combinations with the best results. Columns A, P, R, F contain the values of the following quality metrics: accuracy, precision, recall, and F-measure. “Character” means character-level features, “Word” — word-level ones, “Rhythm” — rhythm ones, + marks the combination of two feature types, “All” — the combination of three feature types. Precision, recall, and F-measure are calculated as the averages for all authors. Bold marks the lines with best quality and best F-measures.

Table 1

Classification of English-language prose by century

Classifier	Feature type	Accuracy	Precision	Recall	F-measure
LSTM	Character	74.1	75.4	73.5	74.4
LSTM	Word	70.7	69.2	69.2	69.2
LSTM	Rhythm	70.0	70.5	70.9	70.7
LSTM	Word + Rhythm	**86.0**	**85.9**	**85.7**	**85.8**
LSTM	All	**89.5**	**89.8**	**89.5**	**89.6**

Table 2

Classification of Russian-language prose by century

Classifier	Feature type	A	P	R	F
GRU	Character	66.1	63.3	65.9	64.6
GRU	Word	74.6	74.3	74.3	74.3
GRU	Rhythm	68.3	69.1	70.6	69.8
GRU	All	**88.1**	**88.1**	**88.7**	**88.4**

Table 3

Classification of French-language prose by century

Classifier	Feature type	A	P	R	F
LSTM	Character	78.9	75.5	75.2	75.4
LSTM	Word	76.3	78.4	74.9	76.6
LSTM	Rhythm	65.8	61.2	60.3	60.7
LSTM	Character + Rhythm	**86.8**	**86.9**	**85.6**	**86.3**
LSTM	All	**84.2**	**82.6**	**82.6**	**82.6**

Table 4

Classification of Spanish-language prose by century

Classifier	Feature type	A	P	R	F
LSTM	Character	94.1	91.5	94.2	92.8
LSTM	Word	92.6	89.4	93.2	91.3
LSTM	Rhythm	92.6	89.7	91.2	90.4
LSTM	Character + Rhythm	**95.6**	**90.6**	**97.3**	**93.8**
LSTM	All	**95.6**	**90.6**	**97.3**	**93.8**

For all languages there are discovered the same tendencies for the classification by centuries. The more feature types we use, the greater classification quality we reach. The only exception is the French language where the combination of character- and rhythm-based features is slightly better than others. Besides, among single feature types character- and word-based features perform better than rhythm-based ones. But rhythm-based features themselves achieve quite good results of the accuracy 65–70 %.

Summing up, there are discovered the same tendencies for the classification by centuries. For Russian and English, the more feature types we use, the greater classification quality we reach. For French and Spanish, the combination of character-level and rhythm features gives better results than other pairs of feature types. Besides, among single feature types character- and word-based features perform better than rhythm-based ones. But rhythm-based features themselves achieve quite good results of the accuracy from 65 to 91 % for different languages.

The **fith chapter** is devoted to the authorship verification of literary texts. It analyzes rhythm features and popular low-level feature based on the statistics of text elements. The comparison is carried out on the corpora of literary texts in English, Russian, French, and Spanish.

The analysis of the state-of-the-art papers shows the lack of comparison of different feature types with linguistic ones, especially for artistic texts [17]. The authors usually rely on standard statistical features based on words and characters and try to extend them by relatively small number of syntactic, topical, or other linguistic features. Deep linguistic features remains under-researched, most probably, because of their complexity in search. Although such features are directly identify the author's style [4] and can be the most interpretable ones.

The authors compares three feature types: levels of characters, words and rhythm, the same as in the previous chapters.

After style features extraction from texts there are the matrix where rows are texts of particular authors, columns are particular features. Each author is verified separately using the whole matrix for the author's language. His/her texts are labeled as belonging or not belonging to him/her. Then the binary classification is performed.

Two classifiers are compared: AdaBoost and Bidirectional LSTM. They have already show their quality in solution of state-of-the-art text classification tasks, as shown in the previous chapter. The five-fold cross-validation technique is used to estimate the stability of classifiers. The texts are divided into five parts, 80 % of texts are the training samples, 20 % are the test ones.

The author compares literary texts of four languages: English, Russian, French, and Spanish. The corpora were created manually collecting famous works of famous authors written in their native language.

In order to make texts equal in size, the author extracted 1–4 fragments with the size about 50 000 characters including spaces from each prose text. In such a way each author is presented by 40 text fragments. English, Russian, and French corpora contain texts of 20 famous authors of 19th–21st centuries, 800 texts per corpora. The Spanish corpus has texts of 8 authors of 19-th–20th centuries, 320 texts in total.

Comparison of two classifiers discovered that AdaBoost outperforms the neural network by 10–15 % of precision, recall, and F-measure. Most probably, it happens due to the fact that the training sample has the insufficient size for better performance of the LSTM network. So the tables in this section contains classification quality for the AdaBoost algorithm.

Table 5

Mean measure values of the authorship verification

Language	Feature type	Precision	Recall	F-measure
English	Character	87.8	80.7	84.1
English	Word	85.8	78.2	81.8
English	Rhythm	82.0	74.2	77.9
English	All	**94.7**	**85.4**	**89.8**
Russian	Character	91.2	81.4	86.0
Russian	Word	92.0	81.9	86.7
Russian	Rhythm	84.7	76.7	80.5
Russian	All	**96.9**	**87.4**	**91.9**
French	Character	93.7	86.5	90.0
French	Word	91.8	80.1	85.6
French	Rhythm	83.5	75.9	79.5
French	All	**97.5**	**90.0**	**93.6**
Spanish	Character	89.9	85.0	87.4
Spanish	Word	92.3	87.9	90.1
Spanish	Rhythm	88.5	86.3	87.4
Spanish	All	**94.1**	**90.0**	**92.0**

Table 5 describes authorship verification quality for all feature types and their combinations. Rhythm features provide the good classification quality. It is lower by 3–11 % of F-measure in the most cases, but has quite high values of 78–87 %. Besides, the number of rhythm features is several times less than character- and word-level ones, so the relatively small number of specific style parameters allow to achieve significant authorship verification quality. Any combination of feature types improve quality by 2–14 %.

Thus, all feature types can provide good verification quality. The specific linguistic features — rhythm features — achieve in many cases high precision, recall, and F-measure with small standard deviations. So they are as useful and stable style markers as standard statistical features: character and word level ones.

Verification of particular authors shows that the many authors have the same style in different fragments. They can be successfully separated from others using only one feature type or the combination of standard and rhythm features. Nevertheless, texts of several authors are verified with very high standard deviations, so there are needed other linguistic features to verify reliably their texts.

The **conclusion** contains the main results of the work:

1. Algorithms and the software tool for automatic search and visualization of lexical and grammatical rhythm features in prose texts for Russian, English, French, and Spanish have been developed. The tool made it possible to automate the work of an expert linguist to analyze the rhythm of the author's text, as well as to carry out experiments on the automatic detection of the text rhythm to build a model of the text rhythm.

2. It is shown that rhythm features can be the indicators of the style of the era based on statistical experiments with large-scale prose texts of the XIX–XXI centuries.

3. The high effectiveness of the text rhythm model for solving the problem of classifying literary texts of the XIX–XXI centuries by centuries and half a century of their publication has been demonstrated. It is shown that rhythm features in combination with standard features increase the quality of solving this task.

4. It is shown that rhythm features are independent markers of an individual author's style and are close to standard features in terms of the quality of the authorship verification, and in combination with them, they increase the effectiveness of authorship verification.

Bibliography

1. *Bergman D. J.*, *Bergman C. C.* Elements of stylish teaching: Lessons from Strunk and White // Phi Delta Kappan. — 2010. — Vol. 91, no. 4. — P. 28–31.

2. *Neal T.* [et al.]. Surveying stylometry techniques and applications // ACM Computing Surveys (CSUR). — 2018. — Vol. 50, no. 6. — P. 86.

3. *Rudman J.* The state of authorship attribution studies: Some problems and solutions // Computers and the Humanities. — 1997. — Vol. 31, no. 4. — P. 351–365.

4. *Boychuk E.* [et al.]. Automated approach for rhythm analysis of French literary texts // Proceedings of 15th Conference of Open Innovations Association FRUCT. — IEEE. 2014. — P. 15–23.

5. *Freyermuth S.* Poétique de la prose ou prose poétique? Le rythme contre le prosaısme // Questions de style, Vous avez dit prose? — 2009. — P. 67–80. — (in French).

6. *Stamatatos E.* A survey of modern authorship attribution methods // Journal of the American Society for information Science and Technology. — 2009. — Vol. 60, no. 3. — P. 538–556.

7. *Lagutina K.* [et al.]. A Survey on Stylometric Text Features // Proceedings of the 25th Conference of Open Innovations Association FRUCT. — IEEE. 2019. — P. 184–195.

8. *Kestemont M.* [et al.]. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection // Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings. Vol. 2125. — 2018. — P. 1–25.

9. *Plecháč P.*, *Bobenhausen K.*, *Hammerich B.* Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry // Studia Metrica et Poetica. — 2018. — Vol. 5, no. 2. — P. 29–54.

10. *Robinson J.* General and individual style in literature // The Journal of aesthetics and art criticism. — 1984. — Vol. 43, no. 2. — P. 147–158.

11. *Balint M.*, *Trausan-Matu S.* A critical comparison of rhythm In music and natural language // Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information. — 2016. — Vol. 9, no. 1. — P. 43–60.

12. *Niculescu I.-D.*, *Trausan-Matu S.* Rhythm analysis in chats using Natural Language Processing // Proceedings of the 14th International Conference on Human-Computer Interaction RoCHI'2017. — 2017. — P. 69–74.

13. *Rubino R.* [et al.]. Modeling Diachronic Change in Scientific Writing with Information Density // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — 2016. — P. 750–761.

14. *Rowlett J. L.* Ralph Cohen on Literary Periods: Afterword as Foreword // New Literary History. — 2019. — Vol. 50, no. 1. — P. 129–139.

15. *Kowsari K.* [et al.]. Text classification algorithms: A survey // Information. — 2019. — Vol. 10, no. 4. — 150 (1–68).

16. *Nowak J.*, *Taspinar A.*, *Scherer R.* LSTM recurrent neural networks for short text and sentiment classification // International Conference on Artificial Intelligence and Soft Computing. — Springer. 2017. — P. 553–562.

17. *Lim C.-G.*, *Jeong Y.-S.*, *Choi H.-J.* Survey of Temporal Information Extraction. // Journal of Information Processing Systems. — 2019. — Vol. 15, no. 4. — P. 931–956.