Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries

Ksenia Lagutina, Anatoliy Poletaev, Nadezhda Lagutina
P.G. Demidov
Yaroslavl State University
Yaroslavl, Russia
ksenia.lagutina@fruct.org, anatoliy-poletaev@mail.ru
lagutinans@gmail.com

Elena Boychuk
Yaroslavl State Pedagogical
University named after K.D.Ushinsky
Yaroslavl, Russia
elena-boychouk@rambler.ru

Ilya Paramonov
P.G. Demidov
Yaroslavl State University
Yaroslavl, Russia
ilya.paramonov@fruct.org

***Abstract*—The paper is devoted to automatic detection of rhythm in fiction and investigation of how rhythm of prosaic texts changed over 19th–21st centuries, based on results of such detection. The authors developed algorithms, which extract rhythm figures related to word repetitions (anaphora, epiphora, polysyndeton, etc.), and visualized their statistical features in plots and heat maps by decades on the material of British and Russian literature. The experiments allowed to find rhythm changes over periods and give interpretation of their reasons from a linguistic point of view.**

I. Introduction

Determining quantitative and qualitative features of a text and its structure allows to identify an individual author style and use it for authorship attribution, authorship verification, text classification, author profiling, and other Natural Language Processing (NLP) problems [1]. We analyzed use of different stylometric features in solving such tasks [2] and concluded that semantic and linguistic features are rarely used in research in NLP and Computer Science fields, unlike classical linguistics, despite the fact that they could improve quality of a problem solution. One of the reasons is that such features are difficult to determine automatically.

Rhythm features of a text are examples of complex linguistic features. Rhythm is defined as a regular repetition of similar and commensurable units of speech [3]. Our goal is to investigate how rhythm figures affect an author's style in prose. We perform automatic identification of these figures in fiction and statistical analysis of their occurrences during 19th–21st centuries. We have developed a software tool ProseRhythmDetector to analyze the rhythm figures in texts and identify tendencies in their use over time. In this article we describe results of experiments with algorithms of figures search and ProseRhythmDetector.

The paper is structured as follows. Section II describes state-of-the-art research concerning text rhythm and an author's style. In Section III we provide our algorithms for rhythm figures detection. Section IV describe design of experiments with algorithms that compute statistical features of text rhythm and visualize them. In Section V we experiment with the proposed algorithms and the ProseRhythmDetector tool and reveal main tendencies in rhythm change over decades and centuries. Section VI analyzes these tendencies and propose possible interpretations. Conclusion summarizes the paper.

II. State-of-the-art

Among all rhythm features state-of-the-art research pays most attention to phonetic ones. Hou et al. [4] use phonetic features to determine authorship of Chinese texts. Accuracy of the method exceeds 85–90 %. Keeline and Kirby [5] describe a series of automatic algorithms that generate data about phonetic rhythm of prosaic Latin texts. They apply statistical criteria to determine the similarity of documents and conclude that different authors usually choose different rhythm.

Niculescu and Trausan-Matu [6] consider the structure of stressed and unstressed syllables as the rhythm basis. Their paper describes an application that analyzes the rhythm of English, Romanian, and French texts of various styles: poems, fiction, and political speech. Applying the tool to rhythm analysis of various works, the authors conclude that it is possible to determine the style of a text by its rhythm features.

Li [7] offers an algorithm for an automated pronunciation assessment system for learning English based on analysis of a structure of sentences and their phonetic rhythm.

Besides, a number of publicly available software applications have been developed to search for phonetic rhythm features. For example, the SPARSAR system analyzes poetry texts to find their rhythm and determine emotional coloring of phrases [8]. The Web application Metricalizer [9] determines metric features of German verses by stressed and unstressed syllables, rhymes, and performs phonetic analysis of words.

Lexical and grammatical rhythm figures based on repetitions of words and phrases are much less studied. Dubremetz and Nivre [10] use the binary logistic regression classifier to extract chiasmas, anaphora, and epiphora from political texts. For most figures the accuracy and F-measure are about 55–65 %. Toldova et. al [11] compare anaphora search engines. The best results are shown by linguistic algorithms that apply rule-based and ontological approaches. Balint et. al [12], [13] define various rhythm features including lexical, grammatical, phonetic, and quantitative. Using statistical algorithms they verify that a text belongs to a particular genre. The accuracy of their method is quite high: 80 %.

A significant number of works that investigate rhythm, is devoted only to development and quality analysis of algorithms for searching lexical and grammatical rhythm figures. A smaller number of works is devoted to application of these features in various NLP tasks. Probably, this is due to complexity of development of algorithms for automatic search of rhythm features that requires deep understanding of the subject area. Moreover, linguistic definitions of rhythm features are not usually formalized and contain many details and rules. Therefore, the development of a software tool with a sufficient search accuracy can be considered as an important task for rhythm research.

Significance of rhythm features for determining the style has been revealed in investigations of texts belonging to different time periods. A style change over time characterizes both individual authors [14] and the language as a whole [15]. In the latter work Kumar et al. search words and phrases that explicitly mark the particular time period. They achieve the median error about 30 years for text classification by dates.

The organizers of the Semeval 2015 contest [16] also noted that the language changes over time, even over relatively short periods. They proposed participants to solve the problem of automatic determining a time period of articles from newspapers published between 1700 and 2010. Interestingly that out of seven teams, only four managed to get a solution. The best results (the accuracy from 60.5 % to 86.8 % for the system based on linguistic and meta-features) were obtained using a wide range of various text features: meta-properties of the document, stylistic, grammatical, lexical functions, and even the search for a direct mention of the document date [17]. The contest results show that the problem of changing of stylistic features of texts over time is very little studied in computer linguistics.

The recent works devoted to automatic determining a specific time period of a text, usually build a text model based on very simple features, mainly word n-grams, sentence lengths, quantitative features of parts of speech [18]. However, Jatowt et al. propose to add more complex stylometric features to the model to improve results. In particular, Gopidi [19] emphasizes that the rhythm and grammar features allow to assess similarity and difference of works of different time periods.

Thus, the identification of a set of complex rhythm figures in works of different centuries and their quantitative analysis would allow to make a significant contribution to the study of the influence of rhythm on the author's style of texts.

III. Algorithms

In this paper the following rhythm figures are automatically extracted from texts and examined:

anaphora: a repetition of sequence of words at the beginning of neighboring sentences;

anadiplosis: a repetition of the same word at the end of a sentence and at the beginning of the following sentence;

diacope: a repetition of a word or phrase with intervening words within one sentence;

epanalepsis: a repetition of the initial part of a sentence at the end of the same sentence;

epiphora: a repetition of the same word or words at the end of neighboring sentences (also called epistrophe);

epizeuxis: a repetition of a word or phrase in immediate succession within one sentence;

polysyndeton: a repetition of the same conjunction within one sentence (simple and pair conjunctions and conjunctive adverbs can be repeated);

symploce: a repetition of the beginning and the end of two or more neighboring sentences, combination of anaphora and epiphora.

The algorithms for searching anaphora, anadiplosis, and polysyndetons with a single conjunction and conjunctive adverbs are taken from our previous work [20]. In that paper there is also an algorithm for a polysyndeton with pair conjunctions, but due to producing fake results for some sentences it has been rewritten for the purpose of this paper.

Algorithms for searching a diacope, epanalepsis, epiphora, epizeuxis, and symploce have been written from ground.

All the given algorithms work with a text previously split into sentences, which are, in turn, split into words. In addition, to search a figure, the algorithms use stop words (unique for every figure). Besides, the polysyndeton searching algorithms use the lists of simple conjunctions, pair conjunctions, and conjunctive adverbs.

As a result, all the algorithms produce lists of figures, every figure specified with the words from a text that form the figure, and a context—a sentence or sentences in which the figure has occurred.

The details of each algorithm including their implementations in pseudo-code are provided below.

A. Diacope searching algorithm

The algorithm steps through unique words in the given sentence and, for each of them, searches for all their positions that are not adjacent. If a word repeats two or more times, it forms a diacope. After that the algorithm steps through the found diacopes and merges those of them that are in nearby words. This allows to find diacopes with multiple word repetition.

```
Require: sentence as list of words $S$, list of excluded words
  (which cannot be part of diacope) $E$
Ensure: list of diacopes in the sentence $D$
  $D := \varnothing$
  $S_u := unique(S) - E$ {unique words in $S$, which can be part
  of diacope}
  for $word$ in $S_u$ do
    $word\_positions := []$
    for $i := 1, \ldots, len(S)$ do
      if $S[i] = word$ and $i - 1$ not in $word\_positions$ then
        append $i$ to $word\_positions$
      end if
    end for
    if $len(word\_positions) >= 2$ then
```

```
      append $diacope(words = word\_positions)$ to $D$
    end if
  end for
  {merging diacope in nearby words}
  $i := 1$
  while $i < len(D)$ do
    $power := 0$
    for $position_a$ in $D[i].words$ do
      for $position_b$ in $D[i+1].words$ do
        if $position_a = position_b$ then
          $power := power + 1$
        end if
      end for
    end for
    if $power >= 2$ and $power = len(D[i+1].words)$ then
      extend $D[i].words$ with $D[i+1].words$
      delete $D[i+1]$
    else
      $i := i + 1$
    end if
  end while
```

B. Epanalepsis searching algorithm

The algorithm steps through the first half of the given sentence and checks if a sentence ends with examined words. A repetition with the maximum length forms an epanalepsis.

```
Require: sentence as list of words $S$, list of stop words $W$
Ensure: epanalepsis in the $S$ if it occurs
  $repeat\_length := 0$
  for $length := 1, \dots, \lfloor len(S)/2 \rfloor$ do
    if $S[1 : length] = S[len(S) - length : len(S)]$ and
    $S[1 : length] \cap W = \varnothing]$ then
      $repeat\_length = length$
    end if
  end for
  if $repeat\_length > 0$ then
    return $epanalepsis(words = S[1 :$
    $repeat\_length] \cup S[len(S) - repeat\_length : len(S)])$
  end if
```

C. Epiphora searching algorithm

The algorithm steps through the sentences of the given chapter and checks if last words of neighboring sentences are the same. If so, then these sentences are a part of a chain repetition; if not, then the repetition chain breaks. The repetition chain forms an epiphora.

```
Require: chapter as sequence of sentences as sequences of
  words $C$, set of stop words $W$
Ensure: list of epiphora in the chapter $E$
  $last\_epiphpra := None$;
  $E := \varnothing$
  for $i = 1, \dots, len(C) - 1$ do
    if $C[i][len(C[i])] = C[i+1][len(C[i+1])]$ and
    $C[i+1][-1]$ not in $W$ then
      if not $last\_epiphora$ then
        $last\_epiphora := epiphora(context = C[i], C[i+1])$;
      else
        expand $last\_epiphora$ context into $C[i+1]$;
      end if
    else if $last\_epiphora$ then
      append $last\_epiphora$ to $E$
      $last\_epiphora := None$;
    end if
  end for
  if $last\_epiphora$ then
    append $last\_epiphora$ to $E$
  end if
```

D. Epizeuxis searching algorithm

1) Searching for an epizeuxis between sentences. For example, "**Weak**! **Weak**! **Weak**!".

The algorithm steps through the sentences of the given chapter and checks if the next sentence repeats the examined sentence. If so, then these sentences are a part of a chain repetition; if not, then the chain repetition breaks. The Repetition chain forms epizeuxis.

```
Require: chapter as list of sentences $C$, list of stop words $W$
Ensure: list of epizeuxis between sentences in the chapter $E$
  $current\_epizeuxis := None$
  $E := \varnothing$
  for $i = 1, \dots, len(C) - 1$ do
    if $C[i] = C[i+1]$ and $C[i+1] \cap W = \varnothing$ then
      if $current\_epizeuxis$ then
        expand $current\_epizeuxis$ context into $C[i+1]$
      else
        $current\_epizeuxis := epizeuxis(context =$
        $C[i], C[i+1])$
      end if
    else if $last\_epizeuxis$ then
      append $last\_epizeuxis$ to $E$
      $last\_epizeuxis := None$;
    end if
  end for
  if $last\_epizeuxis$ then
    append $last\_epizeuxis$ to $E$
  end if
```

2) Searching for an epizeuxis inside a sentence. For example, "**Pretty**, **pretty** good!"

Algorithm steps through the first half of the given sentence and checks if examined part repeats twice. If so, then algorithm examines the remainder part of the sentence and searches for additional repeats of the repeating part. The repetitions forms an epizeuxis.

```
Require: sentence as list of words $S$, list of stop words $W$
Ensure: list of epizeuxis in the sentence $E$
  $E := \varnothing$
  $i := 1$
  while $i < len(S)$ do
    $repeat\_length := 0$
    $n\_repeats := 0$
    for $length = 1, \dots, \lfloor len(S)/2 \rfloor$ do
      if $S[i : i + length] = S[i + length : i + length \cdot 2]$ and
      $S[i : i + length] \cap W = \varnothing$ then
        $repeat\_length = length$
        $n\_repeats = 2$
        break
      end if
    end for
    if $repeat\_length \neq 0$ then
    for $repeats = 3,\ldots,\lfloor len(S[i+repeat\_length \cdot 2]/repeat\_length)\rfloor$ do
      if $S[i:i+repeat\_length] \neq S[i+repeat\_length \cdot (repeats-1) : i+repeat\_length \cdot repeats]$ then
        break
      end if
      $n\_repeats := n\_repeats + 1$
    end for
    append $epizeuxis(context = S, words = i : i+repeat\_length \cdot n\_repeats)$ to $E$
    $i := i + repeat\_length \cdot n\_repeats$
  else
    $i := i + 1$
  end if
end while
```

E. Pair conjunction polysyndeton searching algorithm

The algorithm searches for words forming a given pair conjunction in the given sentence. If the pair conjunction repeats two or more times, its repetitions form a polysyndeton.

```
Require: sentence as list of words $S$, pair conjunction
  $C = C_1, C_2$
Ensure: pair conjunction polysyndeton if it is in the
  sentence $S$
  $positions := \varnothing$
  $first\_word\_position := -1$
  for $i = 1,\ldots,len(S)$ do
    if $S[i] = C_1$ then
      $first\_word\_position := i$
    else if $S[i] = C_2$ and $first\_word\_position \neq -1$ then
      append $(first\_word\_position,\ i)$ to $positions$
      $first\_word\_position := -1$
    end if
  end for
  if $len(P) >= 2$ then
    return $polysyndeton(words = Positions)$
  end if
```

F. Symploce searching algorithm

The algorithm searches an anaphora and an epiphora in the given chapter, which contexts intersect. These anaphora and epiphora forms a symploce.

```
Require: lists of anaphoras $A$ and epiphoras $E$ found in the
  text. For each anaphora and epiphora context (scopes of
  sentences) and indexes of repeating words are given
Ensure: list of symploces in the text $S$
  $S := \varnothing$;
  for $i = 1,\ldots,len(A)$ do
    for $j = 1,\ldots,len(E)$ do
      if $A[i].context \cap E[j].context$ then
        if $A[i].indices[len(A[i].indices)] > E[j].indices[1]$
        then
          append
          $symploce(context = A[i].context \cup E[j].context)$
          to $S$
        end if
      end if
    end for
  end for
```

IV. Design of Experiments

A. General structure of experiments

The experiments performed during our research include the following main stages:

- At the first stage we search for eight rhythm figures: anaphora, epiphora, symploce, anadiplosis, epizeuxis, diacope, and polysyndeton using the algorithms described in the previous section.
- At the second stage we calculate 14 statistical features that describe rhythm figures' occurrences and words in these figures.
- At the third stage we construct plots and heat maps that demonstrate how text rhythm changes over time.

The quality of algorithms of figures search was measured by an expert in linguistics. The methodology of expert analysis and quality of previous versions of algorithms was described in more detail in our paper [21]. Four researchers processed a total of 24 texts of different authors, randomly selected from the corpus. Each expert worked 16 hours. She manually evaluated accuracy of search for all rhythm figures. The exception is diacope, because for it ProseRhythmDetector found several thousands of rhythm figures, so the expert checked only random 10 % of them.

We have measured accuracy of previous and new algorithms on the same corpus of British texts. New algorithms allow to increase the search accuracy from 62–89 % to 80–95 % for most rhythm figures. Thus, at the first stage we get a high-quality model of text rhythm.

The second and third stages are considered in more detail below.

B. Computation of statistical features

We chose 14 statistical features that describe text rhythm in two ways: 9 features indicate occurrences of rhythm figures, and 5 features indicate occurrences of word types in figures. These features allow, on the one hand, to estimate rhythm figures as independent units and, on the other hand, to take into account a structure of figures.

Here is the list of features:

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

Each feature is computed independently for each text. Thus, we get a vector of 14 features for each text.

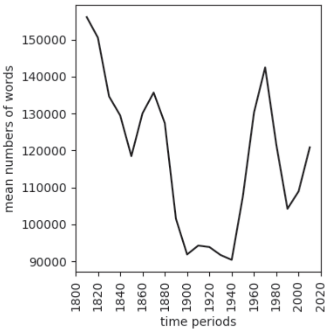

Fig. 1. Mean numbers of words in British texts by decades

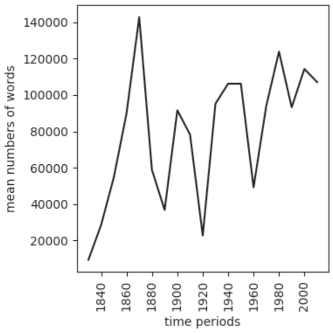

Fig. 2. Mean numbers of words in Russian texts by decades

C. Visualization of statistical features

For all texts from a corpus we compute statistical features. In such a way we get a table where columns are 14 features and rows are prosaic texts. Then, rows are sorted by the year of text publication and arithmetic means are computed for all pairs of a feature and a decade from 1810 to 2010. These data are visualized in two ways:

- As plots with decades on the x axis and feature values multiplied by 100 on the y axis. So we can see numbers of figures per 100 sentences for features of occurrences or percentage of words of a particular type in figures.
- As heat maps of decades similarity. Decades are compared in pairs using a similarity measure: correlation coefficient, Chebyshev distance, Euclidean distance, or Minkowski distance. The result matrix of distances is displayed as a square heat map where rows and columns are decades. Tints in cells mark decades similarities: lighter tints denote more different rhythm.

Both visualization types allow to analyze changes of rhythm figures over decades and centuries and reveal time periods that are similar or different by text rhythm. The results of visualization are presented in the following section.

V. Experiments

A. Corpora and tooling

We conducted experiments with two text corpora in English and Russian languages.

Each corpus contains 150 fiction novels. The English corpus has texts of 43 famous authors of British prose, for example, Jane Austen, William Somerset Maugham, Sebastian Faulks, etc. The Russian corpus has texts of 51 famous authors of Russian prose, for example, Alexander Pushkin, Mikhail Bulgakov, Victor Pelevin, etc. We took from 1 to 5 novels of each author. Each text is marked by a publication date from 1815 to 2019 for British literature and from 1843 to 2019 for Russian literature. Texts contain from 10 000 to 160 000 words. Mean numbers of words by decades are presented in Fig. 1 and Fig. 2.

The algorithms for text processing and scripts for visualization constitute the tool ProseRhythmDetector. It was implemented in Python and use StanfordNLP 0.2.0 and TextBlob 0.15.2 NLP libraries that provide models of English and Russian languages and API for text processing. The tool is available at https://github.com/text-processing/prose-rhythm-detector.

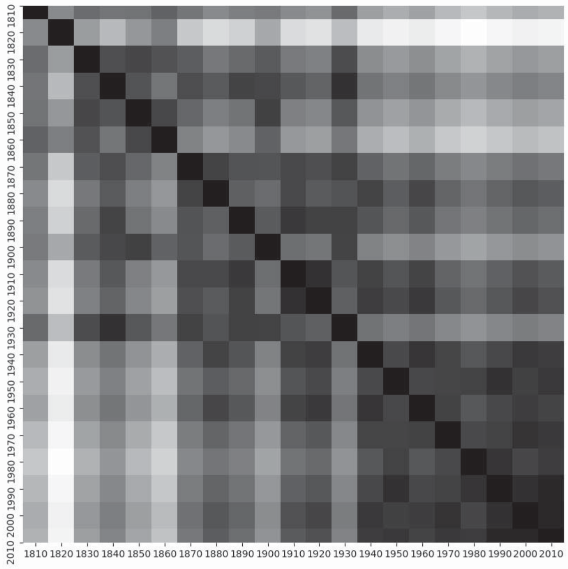

Fig. 3. Heat map for British texts by decades with the Chebyshev distance

B. Heat maps

During the experiments we constructed several heat maps for both corpora. We compared texts by their statistical features using four measures of similarity described in Subsection IV-C. The degree of similarity is shown by a color: the darker tint shows the denser rhythm.

The Euclidean distance turned up as unable to differ decades, so we do not provide its heat map here.

Chebyshev and Minkowski distances provide similar results. Heat maps for British and Russian texts are shown in Fig. 3 and Fig. 4 respectively.

For British texts the algorithms highlight three clusters of literature periods that are close by rhythm: 1830–1860 (the

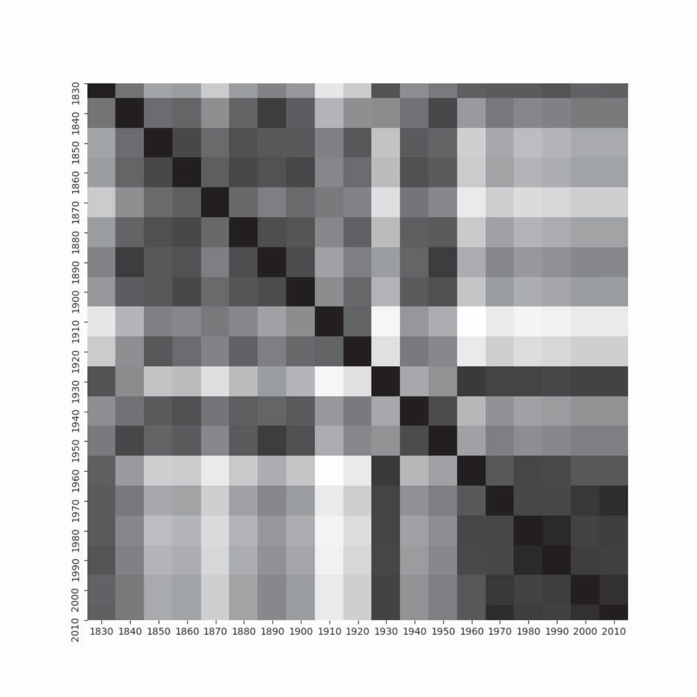

Fig. 4. Heat map for Russian texts by decades with the Chebyshev distance

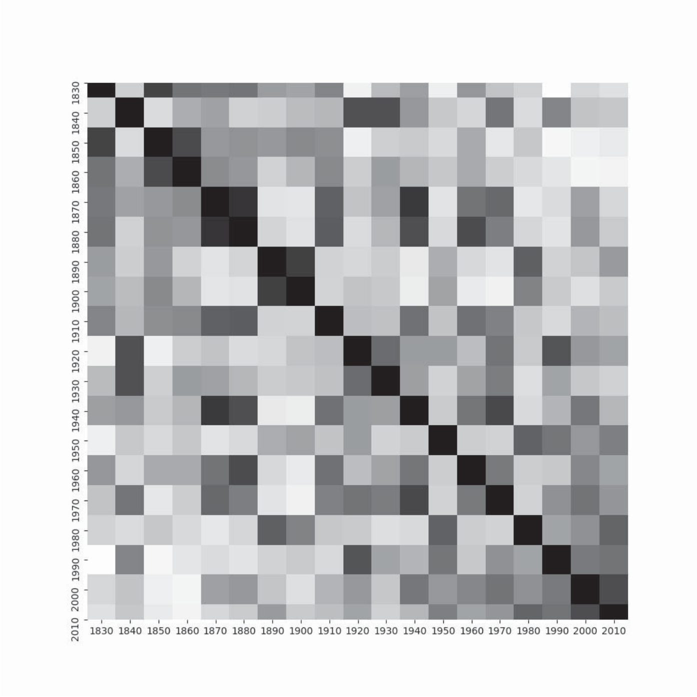

Fig. 5. Heat map for British texts by decades with the correlation coefficient

middle of 19th century), 1870–1930, and 1940–2020. For Russian texts they discover the cluster of the 19th century (1830–1900), two small clusters in the 20th century: 1910–1920 decades and 1940–1950 decades. Decades from 1960 to 2010 are also united into one cluster. For both languages the Chebyshev distance marks the decade of 1930s different from others.

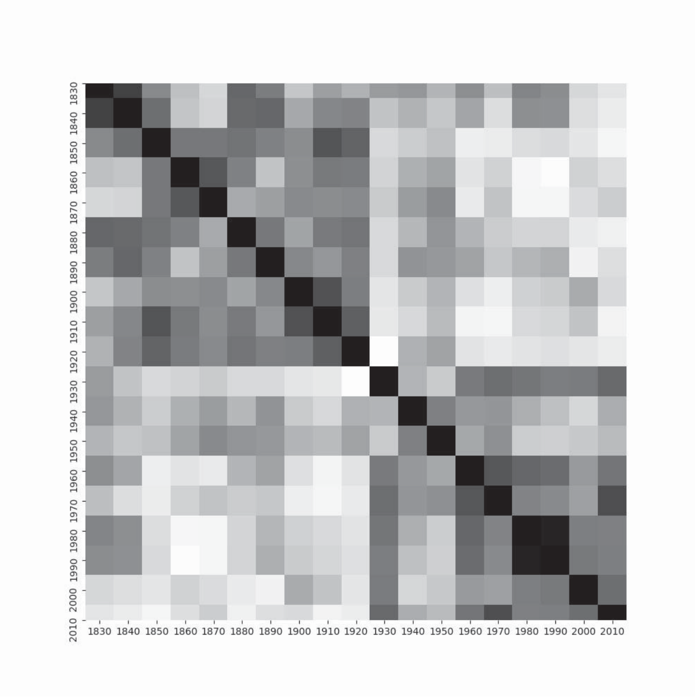

Fig. 6. Heat map for Russian texts by decades with the correlation coefficient

The correlation coefficient is able to split decades into clusters of lesser size than the other measures. Heat maps for British and Russian texts are shown in Fig. 5 and Fig. 6 respectively.

For British texts this coefficient separates decades of the end of the 20th century and the beginning of the 21st century. It also highlights the cluster of the 19th century from 1810 to 1880. For Russian texts the algorithm with this coefficient extracts two large clusters of 1830–1920 and 1960–2010 with small clusters inside. Thus, for both languages this measure splits 19th and 21st centuries into smaller periods and shows the 20th century as a set of periods significantly different by rhythm.

To sum up, heat maps 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.

C. Plots

The plots show changes of particular features over decades. We can see the numbers of all the figures, diacopes, polysyndetons per 100 sentence, and also percentage of unique words in rhythm figures in Fig. 7 for British literature, in Fig. 8 for Russian literature, and Fig. 9 for both languages together.

In the plot for British texts the number of figures has decreased significantly since 19th century: from 160 to slightly more than 60 per 100 sentences. Most frequent figures are diacope and polysyndeton, and their quantity shows the same tendency of a fall. The percentage of unique words has slightly increased, but remains between 60 % and 65 % over all time periods.

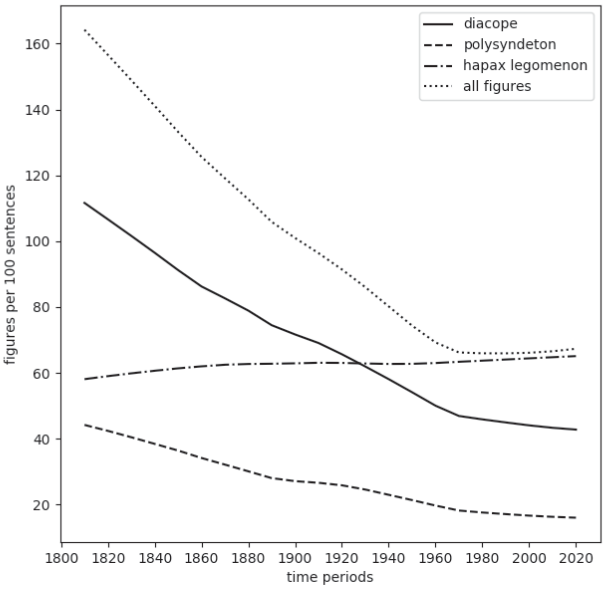

Fig. 7. Rhythm figures for British texts by decades: all figures, hapax legomenon, diacope, polysyndeton

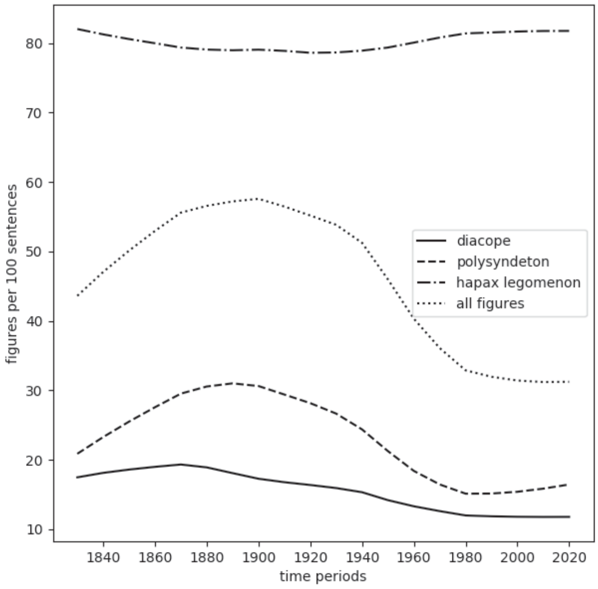

Fig. 8. Rhythm figures for Russian texts by decades: all figures, hapax legomenon, diacope, polysyndeton

Another tendency takes place for the figures in Russian texts: the quantity of rhythm figures was increasing until 1900, but in the 20th century it fell, and in modern literature these statistical features of rhythm are stable. Polysyndeton is more popular than diacope, and the fraction of unique words is higher: about 80 %.

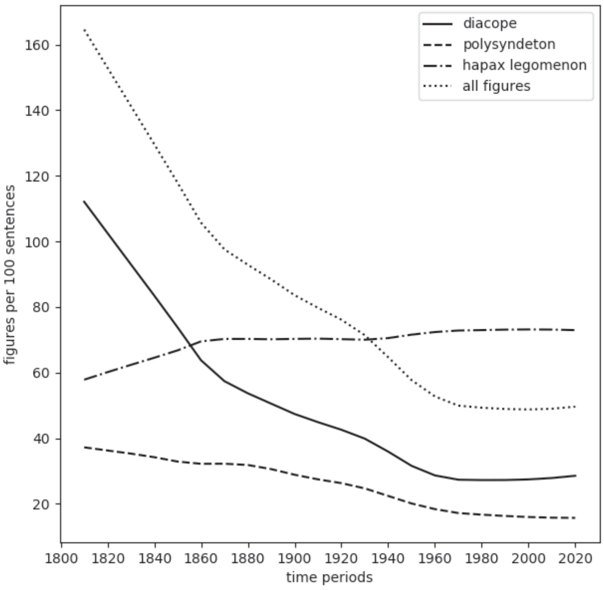

Fig. 9. Rhythm figures for all texts by decades: all figures, hapax legomenon, diacope, polysyndeton

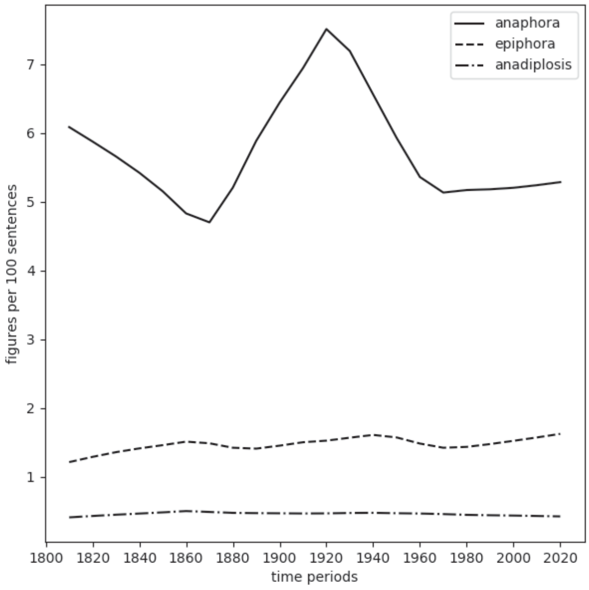

Fig. 10. Rhythm figures for British texts by decades: anaphora, epiphora, anadiplosis

Combining text corpora in both languages, we observe that the tendencies for all texts coincide with ones for British texts (Fig. 9): the fall from 1800 to 1960–1980 and stability for the end of the 20th century—beginning of the 21st century. The most common figures are polysyndeton and diacope.

Other rhythm figures appear too rarely: less than 10 times per 100 sentences. Among them we visualize three most frequent figures that show us the tendencies of their use: anaphora, epiphora, and anadiplosis.

In British texts (Fig. 10) anaphora appears 5–8 times per 100 sentences and have the peak in 1920s. Most probably, this peak is a peculiarity of the style of a particular author's novels, because the change is very small: only by 3 figures per 100 sentences. Epiphora and anadiplosis are rare and stable. Thus, the least frequent figures do not show significant tendencies of changes over decades.

D. Comparison

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. Otherwise, the correlation coefficient enables to distinguish time periods independently from feature values, but only for Russian texts.

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.

VI. Discussion

The results of the experiments demonstrate tendencies in rhythm changes for literature as a whole. In particular, we can see the decrease of the total number of used rhythm figures by the end of the 20th—beginning of the 21st centuries. Besides, the heat maps revealed that literature of the 19th century differs from the other time periods by its rhythm.

In addition, 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. Probably, rare figures can be used in determining the author's style of the text in tasks of authorization and verification. This assumption requires additional investigations of a large number of works of different authors. The software tool developed by the authors allows to conduct such large experiments.

It is very important to note that the results of our experiments that identify complex stylometric features, are useful not only for classical problems of computer linguistics, but also for other areas of linguistics in general. The automation of rhythm figures' search and statistical processing of results allow to change the scale of linguist's work and make interesting conclusions from the point of view of the history of literature.

The obtained results must be interpreted in close connection with historical events and with the history of literary process. An interesting fact is that in plots (see Fig. 7 and Fig. 8) for English-language and Russian-language literature processes of development of a literary text do not coincide from the point of view of rhythm figures implementation. It happens due to different conditions of formation and evolution of the literary language and the novel tradition.

We also see an evident uniform downward tendency in the use of rhythm figures in English literature between 1800 and 2020.

In England, the realism of the second half of the 19th century established the interaction of ancestral and poetic principles: prose genres were dramatized, the growing role of the subjective principle made the lyrical element in the prose obvious, new forms of poetic manifestation appeared. Literature of 1830-1840 years was one of the brightest pages in the world culture: that was the era of development of realism and the coexistence of romanticism with it. The hungry forties were marked by an exacerbation of social conflicts in England, a surge of workers' activity that spilled over into Chartism. In the era of Dickens, Thackeray, Bronte, Gaskell, English literature took a huge step forward, focusing on a person, his role in society, independence, aspirations, desires, ambitions. A gradual decrease in the use of rhythm figures in texts allows to talk about simplification of the language of literature, about its lesser poetization, maybe, about its degradation and a lower degree of emotionality. The general tendency to use diacope and polysyndeton is downward, however, with regard to anaphora, epiphora, and anadiplosis, in English literature from 1870 to 1960 there is a sharp rise (by 1920) and a sharp decline. This peak is not accidental: from 1918 to 1939, Great Britain was in the interwar period that favorably affects the creative mood of writers. English literature of the 1920s is characterized by an increase in modernist tendencies.

Soviet and Russian literature is also characterized by a gradual decrease of figures appearances, but with some options. Until 1900, there was a gradual increase of figures quantity (diacope and polysyndeton). During this period, such authors as F.M. Dostoevsky, L.N. Tolstoy, I.S. Turgenev, I.A. Bunin, A.P. Chekhov and many other Russian classics created their masterpieces. In addition, this is the period of the Silver Age of Russian literature. It was the great rise of Russian culture and beginning of its tragic fall. The beginning of the Silver Age is usually attributed to the 90s of the 19 century, when the poems by V. Bryusov, I. Annensky, K. Balmont, and other remarkable poets appeared. The heyday of the Silver Age is considered 1915—the time of its highest rise and end. Between 1900 and 1940 there is a gradual decrease in figures' appearances that is associated with historical events (World War I, the 1917 revolution, the overthrow of the monarchy); from 1940 to 1980 we see a sharp decline (World War 2). However, from 1980 to 2020 the number of figures is stabilizing. As for such lexical repetitions as anaphora and epiphora, the crisis falls in 1970, after which there is a decline in figures' use.

The obtained results provide an excellent opportunity to estimate the role of rhythm figures not only in determining specifics of the author's style, but also in characterizing the literary process as a whole, in establishing relationships between eras, historical events, literary directions, the author's idiolect, and to help to solve the problem of text attribution.

VII. Conclusion

In our research we developed algorithms that automatically determine a set of rhythm figures in prose. Using these algorithms allowed to process 300 works of English and Russian writers of the 19th–21st centuries in a short time. A statistical analysis of numerical features of texts showed the influence of rhythm figures on the style of fiction of different eras and revealed common and distinctive features of English and Russian texts. It is important to emphasize that the results of automatic text processing allowed to provide meaningful interpretation from the point of view of the history of literature.

Implementation of the developed algorithms for text processing and scripts for visualization are available at https://github.com/text-processing/prose-rhythm-detector.

The obtained results and developed tools of conducted experiments open up a wide range of directions for further research. We plan to continue the analysis of text rhythm and style, including works in other European languages. Another task is to study influence of rhythm figures on the individual author's style and the use of these figures for verification and attribution tasks.

Acknowledgment

The reported study was funded by RFBR according to the research project No. 19-07-00243.

References

[1] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 86, 2018.

[2] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, "A survey on stylometric text features," in *Proceedings of the 25th Conference of Open Innovations Association FRUCT*. IEEE, 2019, pp. 184–195.

[3] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina, "Automated approach for rhythm analysis of French literary texts," in *Proceedings of 15th Conference of Open Innovations Association FRUCT*. IEEE, 2014, pp. 15–23.

[4] R. Hou and C.-R. Huang, "Robust stylometric analysis and author attribution based on tones and rimes," *Natural Language Engineering*, pp. 1–23, 2019.

[5] T. Keeline and T. Kirby, "Auceps syllabarum: A digital analysis of Latin prose rhythm," *The Journal of Roman Studies*, vol. 109, pp. 161–204, 2019.

[6] I.-D. Niculescu and S. Trausan-Matu, "Rhythm analysis in chats using Natural Language Processing," in *Proceedings of the 14th International Conference on Human-Computer Interaction RoCHI'2017*, 2017, pp. 69–74.

[7] X. Li, "English sentence evaluation method using text clustering and semantic ontology." *International Journal of Simulation–Systems, Science & Technology*, vol. 17, no. 42, pp. 271–275, 2016.

[8] R. Delmonte and A. M. Prati, "Sparsar: An expressive poetry reader," in *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 73–76.

[9] K. Bobenhausen and B. Hammerich, "Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer2," *Langages*, no. 3, pp. 67–88, 2015, [in French].

[10] M. Dubremetz and J. Nivre, "Rhetorical figure detection: Chiasmus, epanaphora, epiphora," *Frontiers in Digital Humanities*, vol. 5, p. 10, 2018.

[11] S. Toldova, I. Azerkovich, A. Ladygina, A. Roitberg, and M. Vasilyeva, "Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language," in *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes*, 2016, pp. 74–83.

[12] M. Balint and S. Trausan-Matu, "A critical comparison of rhythm in music and natural language," *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information*, vol. 9, no. 1, pp. 43–60, 2016.

[13] M. Balint, M. Dascalu, and S. Trausan-Matu, "Classifying written texts through rhythmic features," in *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 2016, pp. 121–129.

[14] Y. Liu and T. Xiao, "A stylistic analysis for Gu Long's Kung Fu novels," *Journal of Quantitative Linguistics*, vol. 27, no. 1, pp. 32–61, 2020.

[15] A. Kumar, M. Lease, and J. Baldridge, "Supervised language modeling for temporal resolution of texts," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2069–2072.

[16] O. Popescu and C. Strapparava, "Semeval 2015, task 7: Diachronic text evaluation," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 870–878.

[17] V. Niculae, M. Zampieri, L. P. Dinu, and A. M. Ciobanu, "Temporal text ranking and automatic dating of texts," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, 2014, pp. 17–21.

[18] A. Jatowt and R. Campos, "Interactive system for reasoning about document age," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2471–2474.

[19] A. Gopidi and A. Alam, "Computational analysis of the historical changes in poetry and prose," in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 2019, pp. 14–22.

[20] N. S. Lagutina, K. V. Lagutina, E. I. Boychuk, I. A. Vorontsova, and I. V. Paramonov, "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," *Modelirovanie i Analiz Informatsionnykh Sistem*, vol. 26, no. 3, pp. 420–440, 2019, [in Russian].

[21] E. Boychuk, I. Vorontsova, E. Shliakhtina, K. Lagutina, and O. Belyaeva, "Automated approach to rhythm figures search in English text," in *International Conference on Analysis of Images, Social Networks and Texts, Communications in Computer and Information Science*, vol. 1086. Springer, 2019, pp. 107–119.