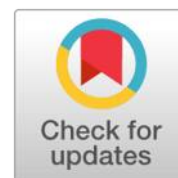# EVALUATING THE PERFORMANCE OF A NEW TEXT RHYTHM ANALYSIS TOOL

Elena Boychuk[1], Ksenia Lagutina[2],
Inna Vorontsova[3], Elena Mishenkina[4], Olga Belyayeva[5]

[1, 3, 4, 5] *K. D. Ushinsky* Yaroslavl State Pedagogical University, Yaroslavl, Russia,
[2] *P. G. Demidov* Yaroslavl State University, Russia

Abstract

The paper assesses and evaluates the performance of the ProseRhythmDetector (PRD) Text Rhythm Analysis Tool. The research is a case study of 50 English and 50 Russian fictional texts (approximately 88,000 words each) from the 19th to the 21st century. The paper assesses the PRD tool accuracy in detecting stylistic devices containing repetition in their structure such as *diacope, epanalepsis, anaphora, epiphora, symploce, epizeuxis, anadiplosis,* and *polysyndeton.* The article ends by discussing common errors, analysing disputable cases and highlighting the use of the tool for author and idiolect identification.

Keywords: text rhythm analysis, diacope, epanalepsis, anaphora, epiphora, symploce, epizeuxis, anadiplosis

Corresponding author:
Elena Boychuk, Doctor of Philological Sciences, is an Associate Professor with the Department of Romance Languages, K.D. Ushinsky Yaroslavl State Pedagogical University, Russia. She teaches courses in linguistics, stylistics, intercultural communication, and French as a foreign language. Her research interests include computer linguistics, phonetics, cognitive linguistics, psycholinguistics, and communication theory.
E-mail: elena-boychouk@rambler.ru   http://orcid.org/0000-0001-6600-2971

Other authors' notes at the end

Rhythm figure analysis

This research aims to assess and evaluate the performance of the ProseRhythmDetector (PRD) tool (Larionov et al., 2020) in terms of relevant automated identification of rhythm figures in 50 English and 50 Russian fiction texts (approximately 88,000 words each)[i] from the 19th to the 21st century when contrasted with manual search results.

The PRD tool has been designed to perform a quick and accurate search producing a quantitative analysis of rhythm figures containing repetition in their structure (diacope, epanalepsis, anaphora, epiphora, symploce, epizeuxis, anadiplosis, polysyndeton). These rhythm figures are examples of repetition determined by the position of repeated units (beginning, end or junction of sentences or clauses, etc.).

Rhythm figure analysis is instrumental in identifying authors' idiolects and making conclusions about the uniqueness of their style and language. This is directly related to the problem of linguistic uniqueness and author identification (e.g. Lagutina et al., 2019; Boychuk & Belyaeva, 2019). The tool has demonstrated encouraging results in this respect.

Other stylistic devices containing various forms of repetition in their structure (chiasmus, polyptoton, derivation, syntactical parallelism etc.) will be considered at a later stage of the tool performance assessment.

Existing tools: state-of-the-art

Few researchers have addressed the problem of using automated tools for text rhythm analysis. There are several works on text attribution, where the following rhythm analysis parameters are considered: rhyme, syllabification, accentuation, and word repetition. Dumalus and Fernandez (2011) regard text rhythm as a valid author's style marker using a simple Naive Bayesian Classifier. Plecháč et al. (2018) apply rhythm parameters to establishing the authorship of poetic texts. These parameters include frequencies of stressed syllables at particular metrical positions and frequencies of particular sounds. Hou and Huang (2019) propose to leverage the phonological information of tones and rimes in Mandarin Chinese automatically extracted from unannotated texts. Balint and Trausan-Matu (2016) consider eight features: numbers of syllables per word, word deemed frequent; normalized numbers of sentence anaphora,

punctuation unit anaphora and commas; percentage of falling word length patterns, frequent words at the end of sentences and at the beginning of punctuation units. Dubremetz and Nivre (2018) assess features based on such rhythm figures as epanaphora, epiphora, and chiasmus. They apply a binary logistic regression classifier to a combination of words and achieve decent extraction quality: over 50% of F-score for all rhythm features.

The authors referred above consider rhythm as a manifestation of one or two parameters rather than a complex phenomenon revealing itself at the level of grammar and lexis. Modern computational linguistics obviously lacks systems capable of both efficiently extracting rhythm features and presenting them in such a way that would make it possible for a researcher to analyse the rhythm of a fictional text in its entirety as well as study its particular aspects.

The Prose Rhythm Detector (PRD) tool

When searching for diacope, epanalepsis, anaphora, epiphora, symploce, epizeuxis, anadiplosis, the PRD filters out words from a stop word list. Each figure can have its own list of stop words with the exception of polysyndeton that refers to a set list of conjunctions.

The search for epanalepsis is based on an algorithm that reviews each sentence for a match of its beginning and ending. If the match is found and the matching units are not on the stop word list, the case is attributed to epanalepsis.

The tool uses two algorithms for detecting epizeuxis. The first compares the neighbouring sentences and registers the aspect as epizeuxis if the sentences repeat. The second checks a single sentence: if it contains words that are repeated in a row, the aspect is also identified as epizeuxis. In neither case are the matching units identified as epizeuxis if they contain stop words.

The algorithm for the search of diacope is based on detecting the repetition of words in a particular sentence. If a word is repeated in a position non-relevant to epizeuxis or epanalepsis and is not on the stop word list, the aspect is registered.

Finally, when all aspects have been identified, the tool displays their full list, as well as the text with the highlighted aspects, and a list of figures with the number of their aspects.

English-language text analysis

The initial assessment of the efficiency of the PRD tool (Boychuk, et al., 2020. p. 107-119) was performed with the use of randomly selected English fiction texts. This research has a more structured approach with 50 texts covering a three-century span. The underlying idea was to see whether texts differ in the use of rhythm figures from century to century. Another interesting point discussed is how the results obtained for the English texts compare with those acquired for the Russian texts.

The total number of words in the English texts in this research is about 1,500,000 per century, i.e. approximately 4,500,000 in total.

The analysis algorithm involved the following steps. The text was uploaded in the text box and processed by the application, which resulted in the generation of an aggregate rhythm figure list (Fig. 1). Selecting a particular figure, the researchers then assessed its use in context discriminating between the proper and the improper automated identification of the figure. In case the tool misidentified the figure, the context was removed from the list and was not accepted for analysis.
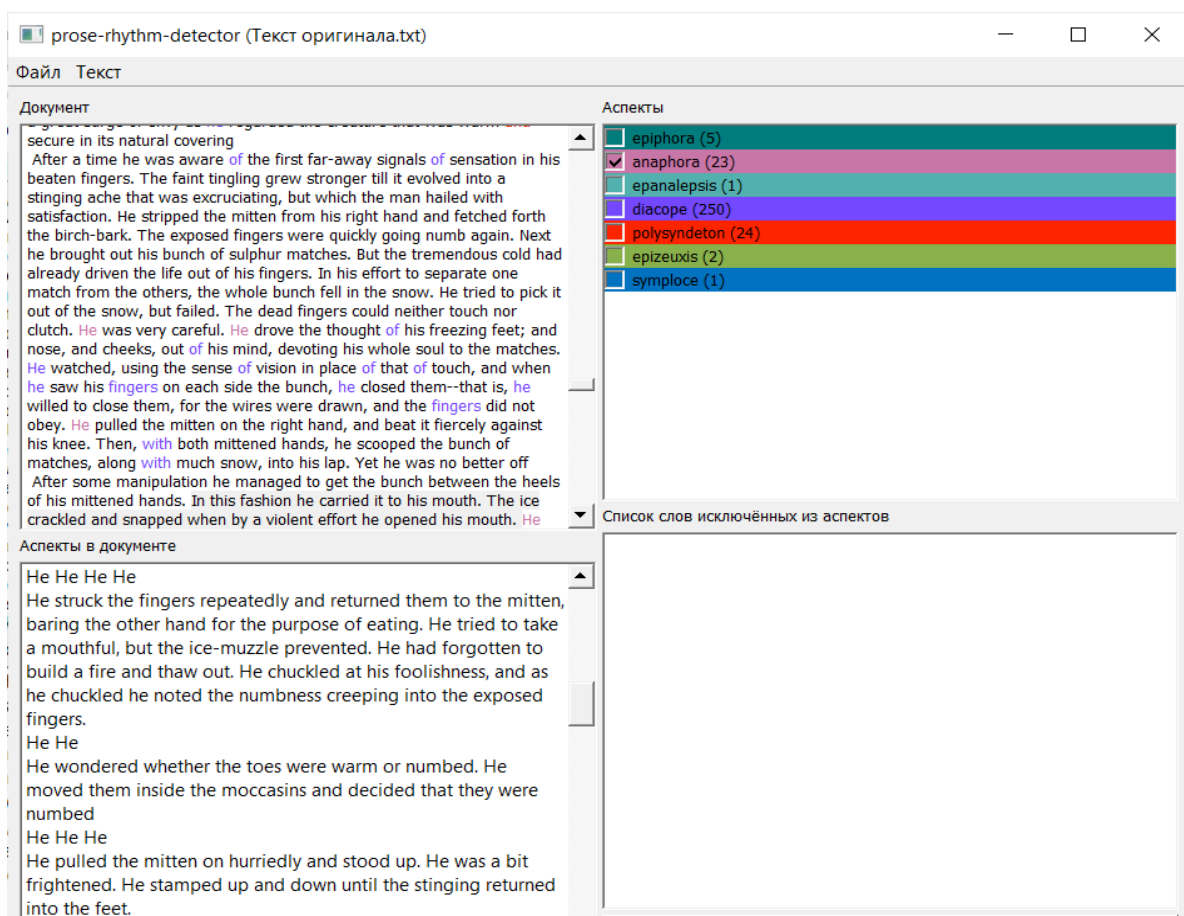

Figure 1. Screenshot of the PRD tool output interface

The findings were organized in tables reflecting the rhythm figure statistics for each text. Since the data are very extensive, they cannot be presented at large in the body of the article, so we describe them in the form of text supplemented with short summary tables (Table 1 and Table 2).

Diacope is the most frequent rhythm figure in English fictional texts of 19th – 21st centuries, ranging between 800 and 9 000 units per text, which depends on the text size and the peculiarities of the author's style:

(1) *It may **hate him** who dares to scrutinise <...> but **hate** as it will, it is indebted to **him** (Ch. Bronte "Jane Eyre").*

The PRD tool demonstrates 87% accuracy of diacope identification (Table 1), which we undoubtedly consider high. The errors introduced by PRD mainly stem from the use of stop words which may prove to be a decisive factor for determining the type of repetition. As has been mentioned previously, all contexts undergo manual verification for errors as well as cross-identifications:

(2) *I thought of course you'd **want to see** her - I don't **want to see** her!* (*I. Murdoch "The Black Prince"*).

The given context contains a case of epiphora rather than a diacope recognized by the tool as such, with the "her" form filtered out.

Polysyndeton is second only to diacope in relation to the frequency of use:

(3) *In fact, he's alert **and** empty-headed **and** inexplicably elated (I. McEwan "Saturday").*

In terms of the accuracy, its level is neither high nor low constituting 77%. Some errors occur due to the misidentification as no difference is detected between, for example, preposition *'for'* and conjunction *'for'*:

(4) *<...>**for** Jay Strauss, **for** there was a possibility of <...> (I. McEwan "Saturday").*

Some inaccuracy of the identification can be explained by the length of the sentences where the conjunction is repeated not to achieve an artistic effect, but to connect clauses in one sentence:

(5) *Don't you really know, Durbeyfield, **that** you are the lineal representative of the ancient and knightly family of the d'Urbervilles, <...> **that** renowned knight <...> (Th. Hardy "Tess of the D'Urbervilles").*

Based on the results of the automated processing of all English texts considered**, anaphora** ranks third for the frequency of use in English fiction:

(6) ***Many*** *strange arms were twined round strange bodies.* ***Many*** *liaisons, some permanent, were formed in the night <…>. (M. Spark "The Girls of Slender Means").*

The accuracy of anaphora identification is very high amounting to 92.5%. The errors are mainly related to cross-identification of anaphora, epizeuxis and simploce which the tool attributes to two classes simultaneously, e.g. epizeuxis and anaphora:

(7) ***Only Bradley. Only Bradley****. (I. Murdoch "The Black Prince").*

A few cases of misidentification are connected to semantic heterogeneity of the repeated units associated with different denotata and included in different types of speech (direct and indirect):

(8) *"****She*** *told me." [end of dialogue, new paragraph]* ***She*** *appraised him a moment, then stood <…>. (J. Fowles "The Ebony Tower").*

Epiphora runs fourth in frequency after diacope, polysyndeton and anaphora:

(9) *Parallel to this, but further from the fire, is a table with Madame's work-box; her two pots of flowers, <…> and her books of* ***devotion****. But Madame reads more than books of* ***devotion****. (E. Gaskell "French Life").*

Compared to diacope and anaphora, the accuracy for epiphora is significantly lower and constitutes 69.9%. A large part of errors is associated with cross-identification of epiphora, epizeuxis and simploce (similarly to anaphora). Errors stemming from the isolated location of the repeated units are not uncommon either. There are a few overlaps with epanalepsis and diacope and a number of misdetections of commas, hyphens, dashes and speech marks.

Epizeuxis is thoroughly used in English fiction, but less frequently than anaphora or epiphora:

(10) *<…> and I walked along it through* ***valleys and plateaus, valleys and plateaus*** *(N. Gaiman "M is for Magic").*

The accuracy of detection attains 72.8% on average, although there might be from 4 to 219 examples of use per text. Having analyzed them, we would like to

highlight that there is a considerable proportion of adverbs such as *yes*, *no*, *all right*, *well* and exclamation *ok* used for emphasis mainly in the dialogues rather than narrative:

(11) **'All right, all right**,*' he says querulously (I. McEwan "Saturday").*

The inaccuracy of detection can be justified by the fact that the PRD tool sometimes identifies a simple repetition of words as epizeuxis, whereas the author uses negative and positive forms with a different intent:

(12) *He may be in denial, knowing and **not knowing**; **knowing** and preferring not to think about it (I. McEwan "Saturday").*

What is more, the repetition of pronouns 'you' or 'it' is also identified as the above-mentioned figure of speech:

(13) *Let me reconstruct a scene for **you**: **You** were out in the garden <...> (N. Gaiman "M is for Magic").*

Epanalepsis is among the least frequent rhythm figures being in advance of only anadiplosis and simploce:

(14) **Everyone** *was going to be a great writer, but **everyone!** (D. Lessing "The Golden Notebook").*

The number of units per text ranges from 4 to 76 and does not allow for spotting any particular trends in terms of its dependence on the time period the text belongs to, the author's gender or individual style*.* The tool accuracy is relatively low constituting 56.01%. The errors are related to its being confused with epizeuxis and positional remoteness of the repeated units (see anaphora, epiphora). A new type of errors is tied to the homonymy of forms recognized as epanalepsis:

(15) **There** *were a great many words **there**. (I. Murdoch "The Black Prince").*

Anadiplosis comes seventh in terms of the frequency of use, although it is a very important literary device that helps writers to draw readers' attention to central characters, their feelings, and the most significant events, etc.:

(16) *And then, <...>, **I'm falling. I'm falling** into a black tunnel, the same black tunnel<...> (S. Thomas "The End of Mr Y").*

One of the most common cases is the use of proper names:

(17) *What's he on about, **Baxter**? **Baxter** shoves the broken wing mirror <...> (I. McEwan "Saturday").*

According to the statistics, anadiplosis accounts for 71.5%, so we can see the level of accuracy is relatively low. The main issue with its identification is that the PRD tool detects anadiplosis when there is a repetition of personal pronouns, auxiliary verbs, question words and demonstrative pronouns:

(18) *So I fooled **you**. **You** were out of position. (I. McEwan "Saturday").*

Symploce is the least frequent rhythm figure of speech found in our corpus of English fictional texts:

(19) ***Maybe** it was **too late. Maybe** we got her **too late**. (R. Galbraith / J.K. Rowling "The Cuckoo's Calling").*

In 1/8 of the texts the PRD tool did not detect any examples of it at all. In the rest of the texts the number of symploce varies from 1 to 12 per text. The accuracy of the identification of this figure is rather low reaching only 48.6%. There are quite many overlaps with anaphora and epiphora as the PRD tool regards the repetition of the whole sentence as symploce:

(20) ***Get out and run. Get out and run.** (S. Thomas "The End of Mr Y").*

Table 1

Accuracy of automated rhythm figure detection in 50 English texts

Devices	Devices quantity		Accuracy (%)
	found by the instrument	real quantity	
diacope	137 958	120 023	87.00
epanalepsis	1 105	619	56.01
epiphora	3 090	2 160	69.90
anaphora	9 808	9 072	92.50
symploce	183	89	48.60
epizeuxis	3 288	2 396	72.80
anadiplosis	1 029	736	71.50
polysyndeton	53 984	41 567	77.00
Sum total of devices	210 445	176 662	83.94

As could be seen from Table 2 below, the rhythm figure pattern of English fictional texts changes throughout the centuries. A steady decline in the use of diacope and polysyndeton is among the most notable trends. Although no objective evidence has been collected so far, we can hypothesize that such a tendency could be explained by the 20th -21st century authors expressing less interest to the narrative development and focusing their effort on the unfolding and improvement of dialogues which are intended

to serve an artistic mould of spontaneous speech. Dialogue (speech)-centred texts are likely to witness an increase in the use of anaphora, which is another trend indicated by the research data. The fact is that anaphora is one of the most powerful rhetoric means capable of producing a strong and convincing impression and thus frequently resorted to by the speakers to reach their audience. Interestingly, many of the authors analyzed are (were) university professors or lecturers, which offers ample evidence of their remarkable speaking skills. The accelerating trend in the use of anaphora in written texts, as well as the dramatic rise in the use of epizeuxis and epiphora in the 20th century fiction, could also be (have been) inspired by the employment of these rhythm figures in the audio and audio-visual media – radio, TV and cinema, in the first place. Finally, a connection could be established between the increase in the use of the above figures and the growing complexity of the genres and plots of modern fiction, whereby the clarity as well as the persuasive effect could be achieved through an enhanced role of rhetoric figures.

Table 2

Rhythm figure distribution statistics for English texts

Devices	XIXc.	XXc.	XXIc.
diacope	49 432	38 803	31 788
epanalepsis	206	210	203
epiphora	457	965	738
anaphora	2 380	3 164	3 528
symploce	19	31	39
epizeuxis	806	923	667
anadiplosis	240	250	236
polysyndeton	16 638	13 403	11 526

Russian-language text analysis

Russian-language texts also cover the period from the 19th to the 21st century[i]. As is the case with the English texts under analysis, the total number of words in the Russian texts in this research is around 1,500,000 per century, i.e. approximately 4,500,000 in total.

Polysyndeton. The frequency of its use is very high reaching 86.6%. The most common conjunction for polysyndeton is the conjunction *u*, which can be repeated in the text from 2 to 5 times depending on the author:

(21) *В другом случае характер его был чрезвычайно мрачен, **и** когда напивался он пьян, то прятался в бурьяне, **и** семинарии стоило большого труда его сыскать там. (N. Gogol "Viy").*

However, for example, in A. Terekhov's work "The Germans", the repetition of this conjunction is 9 times within one phrase.

Diacope comes second in Russian texts making 3000 cases per text on average:

(22) *Староста расчесал себе бороду и важно упирается на **палочку** из соседней рощи, **палочку**, известную многим в деревне. (V. Sollogub "Serezha").*

The tool accuracy in detecting diacope is 72.06% (Table 2), the error being quite large and arising out of cross-identification of diacope, anaphora, epiphora, epanalepsis, syntactical parallelism, epizeuxis and chiasm.

Anaphora ranks third for the frequency of use. The level of accuracy in identifying anaphora is very high reaching 90.13%. As has been mentioned, the errors mainly occur due to its cross-identification with diacope:

(23) ***Бабушка** до сих пор любит его без памяти <…> **Бабушка** знала, что Сен-Жермен мог располагать большими деньгами. (A. Pushkin "The Queen of Spades")*

or epizeuxis:

(24) ***Где доктор? Где доктор**, я вас спрашиваю! (A. Strugatsky, B. Strugatsky "Hard to be a God").*

It should be noted that pronominal anaphora prevails over other types making 90% of cases.

Epizeuxis. In terms of its detection by the tool, the degree of accuracy is 87.89%.

(25) ***Прощайте, прощайте,** храни вас господь! (F. Dostoevsky "Poop Folk").*

In some cases, when the number of repeated elements is greater than two, only the first and last elements are defined by the tool, attributing this example to epanalepsis, for intstance:

(26) ***Мой, мой, мой!** (I.Turgenev "Annouchka").*

The tool sometimes detects repetition as epizeuxis, although in the following cases there is anadiplosis. This is because the comma is between both homogeneous elements and parts of the sentence.

Epiphora. The tool ensures 87.49% accuracy in detecting epiphora:

(27) *А я буду **плясать**. Жену, детей малых брошу, а пред тобой буду **плясать**. (F. Dostoyevsky "The Idiot").*

The errors here are reminiscent of those described above and include mistaking epizeuxis for epiphora:

(28) *Я игра-ать мной **не позво-олю! Не позво-олю** (A. Chekhov "The Duel"),*

and misidentification of repeated initials and abbreviations consisting of repeated letters:

(29) *<...> Харитонов **А. А.** (O. Slavnikova "The Immortal").*

Anadiplosis. The use of words at the junction of the parts of the sentence and sentences is detected by the tool very well achieving a high level of accuracy which is 89.21%:

(30) *<...> он прошел в **кабинет**. **Кабинет** медленно осветился внесенной свечой (L. Tolstoy "Anna Karenina").*

Regarding the improvements that should be made to the tool, abbreviations with punctuation ought to be taken into consideration:

(31) *<...> при своем превосходном уме и положительном знании жизни **и пр. и пр.,** <...> (F. Dostoevsky "The Idiot").*

In the following example, the repeated elements are identified as epizeuxis, although according to the meaning and structure of the sentence this repetition corresponds to anadiplosis:

(32) *Это был наш общий **язык, язык,** подаренный мне ею, <...> (E. Vodolazkin "The Abduction of Europa").*

Epanalepsis. A relatively high level of accuracy for epanalepsis – 70.79% – speaks for the correct laydown of the tool specifications:

(33) ***Аглая** мне урок дала; спасибо тебе, **Аглая**. (F. Dostoyevsky "The Idiot").*

The tool misdetects epanalepsis confusing it with epizeuxis in the following examples:

227

(34) **Секунда… секунда…** *(V. Pikul "Requiem for Convoy").*

In order to avoid such errors, the presence of intermediate components between repeated units should be included in the tool specifications.

Simploce. The frequency of its use is very low in the texts. The level of accuracy of symploce detection by the tool is quite high constituting 72.84%.

(35) **Он** *никак не ожидал того, что он увидал и почувствовал у* **брата**. **Он** *ожидал найти то же состояние самообманыванья, <…> во время осеннего приезда* **брата**. *(L. Tolstoy "Anna Karenina").*

In some cases, the tool detects symploce as a repetition of the conjunction *и* at the beginning of the sentence, considering it as a content word at the end of the sentence:

(36) **И** *тогда мать* **заплачет**. **И..,** *может, он тоже* **заплачет** *(V. Pikul "Requiem for Convoy").*

These examples are ambiguous, because, on the one hand, the repetition of the conjunction *и* can be anaphoric and can bear a certain meaning, and, on the other hand, the roles of the link-word and the content word are not equal.

Table 3

Accuracy of automated rhythm figure detection in 50 Russian texts

Devices	Devices quantity		Accuracy (%)
	found by the instrument	real quantity	
diacope	30 701	22 123	72.06
epanalepsis	493	349	70.79
epiphora	2 542	2 224	87.49
anaphora	4 033	3 635	90.13
symploce	81	59	72.84
epizeuxis	3 855	3 388	87.89
anadiplosis	760	678	89.21
polysyndeton	40 852	35 376	86.60
Sum total of devices	83 317	67 832	81.41

The century-based findings recorded for the Russian literary texts are summarized in Table 4 and reveal a decline in the use of rhythm figures from the 19th to the 21st century. It is an observation so far. Still, the figures allow for an assumption that the above tendency may testify to changes in the literary language quality or other important processes. However, it undoubtedly requires further comprehensive research

which will focus on other linguistic parameters of the text structure along text rhythm exploration.

Table 4

Rhythm figure distribution statistics for Russian texts

Devices	XIXc.	XXc.	XXIc.
diacope	7 838	7 277	7 008
epanalepsis	163	128	49
epiphora	994	797	433
anaphora	1 307	1 278	1 050
symploce	24	18	16
epizeuxis	1 759	852	6 94
anadiplosis	250	220	208
polysyndeton	18 287	8 095	8 774

Discussion

With regards to the research results we consider it essential to address the causes of the discrepancies noticed when testing the tool:

1. Lower than expected accuracy in detecting diacope, epanalepsis, epizeuxis and simploce resulting from their cross-identification and automatic attribution to several classes: the solution to the problem is seen in the introduction of new stop words and word units ("had had", "was (.,;)was", "that that", "you (.,;) you", etc.) as well as accounting for intermediate words between repeated units;

2. Misdetection of punctuation marks (commas, hyphens, dashes and quotations) preventing the tool from accurately identifying certain rhythm figures, diacope, anaphora and epanalepsis in the first place;

3. Misrecognizing of initials (with a full stop) as full-fledged sentences: the above problems can be solved by defining specifications for such cases, e.g. listing the relevant punctuation marks as stop words;

4. Confusion of rhythm figures, e.g. epiphora and mimesis (the latter is currently not on the list of rhythm figures available for the tool) which calls for the necessity of formulating a set of specific rules for the case.

Conclusions

The PRD tool has demonstrated a rather high level of accuracy in detecting rhythm figures— 83.94% for English texts and 81.41% for Russian texts.

Some of the statistical errors discovered in the course of the research can be rectified by compiling more comprehensive as well as better targeted stop-word lists, reducing text portions intended for automated analysis and making other rhythm figures available for the tool, which is almost certain to improve its accuracy level. However, not all statistical uncertainties can be eliminated. This is particularly true for misidentifications stemming from homonymy (polysemy) and other content-based phenomena.

The different quantity of rhythm figures in the texts of different authors allows for an assumption that each author has their own bank of rhythm-based stylistic devices. Thus, this tool can also be used for the identification of specific features of authors' idiolect and style.

The quantitative research of rhythm figures in English and Russian fictional texts covering a time span of three centuries has demonstrated a more extensive use of the above figures in English fiction as compared to Russian. This can be explained by the peculiarities of the language morphologic, lexical and semantic structures as well as their principles of clause and sentence construction. The accuracy of automated rhythm figure identification is high for both languages: over 83% for English and over 81% for Russian.

The quantitative data concerning the distribution of rhythm figure show a downward temporal trend in rhythm figure use in the Russian fiction. The English fiction witnesses a steady decline in the use of diacope and polysyndeton along with an appreciable rise in the use of anaphora. Further comprehensive research is required to conclude on the statistics obtained.

References

Balint, M., & Trausan-Matu, S. (2016). A critical comparison of rhythm in music and natural language. *Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information*, *9*(1), 43–60.

Boychuk, E., & Belyaeva, O. (2019). La téchnique de stylométrie réalisée à la base de l'analyse informatique du rythme du texte. *10-ièmes Journées Internationales de Linguistique de Corpus (JLC)*. Université Grenoble-Alpes, 26-28.11.19. 163-167.

Boychuk E., Vorontsova I., Shliakhtina E., Lagutina K., Belyaeva O. (2020) Automated Approach to Rhythm Figures Search in English Text. In Wil M. P. van der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, V. Kuskova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, P. M. Pardalos, M. Pelillo, A. V. Savchenko, E. Tutubalina (Eds.), *Analysis of Images, Social Networks and Texts. AIST 2019*, *Communications in Computer and Information Science, vol 1086* (pp. 107-119). Springer. https://doi.org/10.1007/978-3-030-39575-9_11

Dubremetz, M., & Nivre, J. (2018). Rhetorical Figure Detection: Chiasmus, Epanaphora, Epiphora. *Frontiers in Digital Humanities, 5*(10). 1-16. https://doi.org/10.3389/fdigh.2018.00010

Dumalus, A., & Fernandez, P. (2011). Authorship attribution using writers rhythm based on lexical stress. *Proceedings of the 11th Philippine Computing Science Congress*. 82–88

Hou, R., & Huang, C. (2020). Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering, 26*(1), 49-71. https://doi.org/10.1017/S135132491900010X

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I. (2019). A Survey on Stylometric Text Features. *25th Conference of Open Innovations Association (FRUCT)*, Helsinki, Finland, 2019, 184-195. https://doi.org/10.23919/FRUCT48121.2019.8981504

Larionov, V., Petryakov, V., Poletaev, A., Lagutina, K., Manakhova, A., Lagutina, N. and Boychuk, E., (2020). *ProseRhythmDetector*. K.D. Ushinsky Yaroslavl State Pedagogical University, Yaroslavl, Russia. https://github.com/text-processing/prose-rhythm-detector

Plecháč, P., Bobenhausen, K., Hammerich, B. (2018). Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry. *Studia Metrica et Poetica*, *5*(2), 29–54. https://doi.org/10.12697/smp.2018.5.2.02

English-language text analysis is based on works by:

- 19th century - Charles Dickens**,** Charlotte Bronte, Elizabeth Gaskell, Jane Austen, Thomas Hardy
- 20th century - Robert Lewis Stevenson, Virginia Woolf, James Joyce, Iris Murdoch, Muriel Spark, Daphne du Maurier, John Fowles, David Herbert Lawrence, Doris Lessing
- 21st century - Ian McEwan, Neil Gaiman, Scarlett Thomas, Joan K. Rowling, Sebastian Faulks, Jenny Colgan, Kazuo Ishiguro, Paula Hawkings, Sarah Perry, Ruth Hogan, Tony Parsons

Russian-language text analysis is based on works by:

- 19th century - Nikolay Gogol, Fyodor Dostoevsky, Alexander Pushkin, Vladimir Sollogub, Leo Tolstoy, Ivan Turgenev, Anton Chekhov
- 20th century - Ivan Bunin, Alexander Grin, Mikhail Bulgakov, Maxim Gorky, Vasily Aksenov, Valentin Pikul, Sergey Dovlatov, Victor Pelevin, Alexander Prokhanov
- 21st century - Eugeny Vodolazkin, Vladimir Mikushevich, Zakhar Prilepin, Alexander Terekhov, Dmitry Bykov, Olga Slavnikova

Authors' note

Ksenia Lagutina, MSc in Computer Science, is a Postgraduate Student and Assistant Professor with the Department of Theoretical Informatics, P.G. Demidov Yaroslavl State University, Russia.
E-mail: lagutinakv@mail.ru   http://orcid.org/0000-0002-1742-3240

Inna Vorontsova, PhD, is an Associate Professor with the Department of Translation and Interpretation, K.D. Ushinsky Yaroslavl State Pedagogical University, Russia. She teaches courses in lexicography, legal translation, and English as a foreign language. Her research interests lie in the field of lexicology and lexicography, discourse analysis and theory of translation.
E-mail: arinna1@yandex.ru   http://orcid.org/0000-0001-5897-9299

Elena Mishenkina, PhD, is an Associate Professor with the Department of Translation and Interpretation, K.D. Ushinsky Yaroslavl State Pedagogical University, Russia. She teaches courses in translation, intercultural communication, and English as a foreign language. Her research interests include linguistics, psycholinguistics, ethnosociolinguistics, communication theory and theory of translation.
E-mail: vitalt@mail.ru   https://orcid.org/0000-0002-1314-4156

Olga Belyayeva is an Assistant Professor with the Department of Translation and Interpretation, K.D. Ushinsky Yaroslavl State Pedagogical University, Russia. She teaches courses in translation and English as a foreign language. Her research interests are related to applied linguistics, corpus linguistics and English literature.
E-mail: olbelyaeva@yandex.ru   https://orcid.org/0000-0003-3658-7336