A New Electronic System for Comparative Analysis of Verse and Prose

Evgeny Kazartsev
National Research University Higher School of Economics (HSE University), Faculty of Humanities / Moscow, 105066, Staraya Basmannaya St, 21/4
kazar@list.ru

Tatiana Zemskova
National Research University Higher School of Economics (HSE University), Faculty of Humanities / Moscow, 105066, Staraya Basmannaya St, 21/4
tatzem98@gmail.com

Abstract

This paper will focus on the development of a new computational system, *Prosimetron*, which enables comparative statistical studies of the rhythm of verse and prose in different languages (currently 10 languages are operative, with the possibility of adding more). The results of the analysis can be used not only for studying the processes for the genesis, expansion, and modification of various versification systems, but also for commenting on and interpreting the verse rhythm in different national poetic traditions in comparison with their foreign sources and language prosody. In addition, the possibility to model various processes of poetic speech generation and to analyze rhythmic vocabularies of prose allows hypotheses about the cognitive mechanisms of verse generation. This system operates in a semi-automatic mode and, by minimizing errors and enabling the processing of large amounts of data, provides a unique tool for computer research on the rhythm of different modes of speech.

Keywords: rhythm; prose and poetry; comparative statistical analysis; versification models
DOI: 10.28995/2075-7182-2021-20-378-384

Новая компьютерная система сравнительного анализа стиха и прозы

Евгений Казарцев
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), факультет гуманитарных наук / г. Москва, 105066, Старая Басманная, 21/4
kazar@list.ru

Татьяна Земскова
Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), факультет гуманитарных наук / г. Москва, 105066, Старая Басманная, 21/4
tatzem98@gmail.com

Аннотация

В работе описываются элементы новой компьютерной системы – Прозиметрон, которая позволяет осуществлять сравнительно-статистический анализ ритма стиха и прозы на разных языках (на настоящий момент доступны 10 языков, планируется расширение этого списка в дальнейшем). Результаты исследования могут быть использованы не только для изучения процессов становления, распространения и эволюции систем стихосложения, но и для изучения ритма стиха в разнообразных поэтических традициях в сравнении с их иностранными источниками и языковыми характеристиками. Кроме того, возможность моделирования процессов порождения стихотворной речи позволяет выдвигать гипотезы о когнитивных процессах, связанных с генерацией стиха. Предложенная система функционирует в полуавтоматическом режиме и, позволяя обрабатывать большое количество данных, представляет собой уникальный инструмент для компьютерного анализа ритма стиха и прозы.

Ключевые слова: речевой ритм; проза и стих; сравнительный статистический анализ; модели стихосложения

1 Introduction

This research is devoted to the study of poetic rhythm based on a new computational system for the analysis of prosodic structures in different languages. The system works on the so-called "Russian" or "linguo-statistical method"[1], and is called *Prosimetron*,[2] because it enables the analysis of rhythm both in metrically organized texts, and in texts that are free from metrical organization — thus both in verse and in prose. Currently, this system is still being developed; however, there are already some results from this project that correspond with the results and calculations in previous studies that were carried out manually or with the use of basic computational tools. This allows us to argue that the new system works well and provides reliable information. Some of its interesting findings are presented in this paper.

The development of this new computational system for verse analysis is an important step towards studying peculiarities in the organization of poetic speech against the background of prose rhythm in various languages [2: 155]. This system is comprehensive: it can store and process various texts in relation to their rhythm.

All texts entered into the system are pre-attributed. For each text, a separate data cell is created, which indicates such parameters as the author, year of creation, title, language of the work, type (prose or poetry), bibliographic description of the source, name of the person who puts the text into the system, verse parameters for poetic texts, and some others. The system is already a unique repository of both fairly common texts and rare ones, including several editions of the same text, which can be used to study changes in the rhythm of an idiolect or the development of rhythmic inertia in verse. To enable further rhythmic analysis special rhythmic markings are added to all texts: the border of rhythmic words and the position of the stress are indicated. For metrical texts, special phenomena such as caesura are also noted.

Currently there are two modes for marking texts: automatic and manual. Automatic indication of tresses and boundaries of rhythmic words has so far been implemented only for Russian, but in the future it will be expanded to all working languages of the system. The accuracy of this marking is now about 90%.[3] Whichever way the text has been marked, it then goes through a semi-automatic three-stage process of checking and correcting the markup. The first stage involves a search for technical errors — when the stress falls on consonants, spaces and other signs — and is carried out automatically. At the second stage all verse lines are checked for compliance with the meter of the entire poem (the meter is determined automatically). If a line does not match the meter of the entire text, either the text is designated as polymetric, or the line is marked as invalid and needs to be checked by a user.

The third and the most important stage in validating the markup involves the possibility of a manual search for all forms of a particular word in a rhythmic context and manually correcting the position of stress or other attribution, followed by the automatic recalculation of all related parameters. While a part of the validation phase, this toolkit also provides an important means for research. Seeing at what position in the poetic line a particular word or its form is most often found, we can draw conclusions regarding its stress status both within a given text and in general (for example, the controversial status of some rhythmic clitics can be revised). Automatic recalculation of all results allows users to quickly change the entire markup in case it has been decided to change the stress status of a particular word.

In the future, it is planned to create an algorithm that will be able to automatically search for texts on the network, mark them up and then enter them into the corpus with minimal human participation. This will expand the applicability of the techniques developed using the current material to an unlimited number of texts, enabling the Prosimetron system to carry out big data analysis.

As already mentioned, Prosimetron currently works with two categories of texts: prose and poetry. In total, 10 working languages have now been implemented: Ancient Greek, English, German, Dutch, French, Swedish, Russian, Belarusian, Ukrainian, and Polish. All of these are represented by samples of both types of texts. The total volume of the poetic corpus is about 250,000 lines, while the prose corpus

[1] That terms are introduced into the international scientific use by James Bailey [1].

[2] The authors of this article, together with Boris Maslov and Viktor Vashchenkov, as well as a group of HSE-University students led by Evgeny Kazartsev and Tatyana Zemskova, are actively participating in the development of that system, especially in the framework of a project supported by Russian Science Foundation in the period 2016–2020. The name *Prosimetron* was suggested in 2020 by Boris Maslov.

[3] Two algorithms are used to prosodic marking of verse, a preliminary one based on the Zaliznyak dictionary and a final one based on a specially trained recurrent neural network trained on a large corpus of metrical texts.

contains about 390,000 rhythmic (phonetic) words.[4] At the moment the poetic corpus consists primarily of clearly metrical poems, written in iambic, trochaic, dactylic, etc., but there are also several syllabic and polymetric poems. In principle, there are no restrictions on the type of versification.

2 Results

Let us show some analytic possibilities that the Prosimetron can provide. One of the results from processing a prose text is the compilation of rhythmical vocabularies. In other words, it is possible to find out which rhythmic words (taking into account their length and stress placement) are most often found in a particular text and, more broadly, in a particular language. It is important to note that in this context we are dealing specifically with rhythmic words: thus, when rhythmic clitics (prepositions, conjunctions, articles, particles etc.) combine with rhythmically independent words they form groups of syllables, united by a single stress vertex. For example, *the táble*, *in a tíme*, *dróp it*, or Russian combinations *like moi dóm*, *na dné*, *pered lésom* etc all comprise single rhythmic words.[5] With the help of this toolkit, it becomes clear that the non-fixed Russian stress generally tends to be in the middle of a word rather than closer to the ends.[6] This result — previously observed in some texts of Russian prose — is now confirmed and strengthened thanks to the computational analysis of big data.

Russian words look longer than English or German, and some scholars believe that the relatively low number of realized stresses in Russian metrical verse as compared with English or German is predetermined by the word length. Russian words seem to be longer, and therefore stress cannot be realized at as many strong positions of verse. Figure 1, which shows the average word length (in syllables) found in fiction samples from different languages, indicates that this is not the case. Granted, the average number of syllables for rhythmic words in the Slavic languages is slightly larger, but this difference is not so great that by itself it could cause the number of realized stresses in the Slavic poetic tradition to differ so sharply from, for example, the German: in German verse, about 75% of iambic tetrameter lines are fully-stressed, in Russian a little less than 30% are, and in Ukrainian still fewer. Most likely, the key factors are the author's intention and internal laws of the poetic tradition, while the language itself does not necessarily predetermine the verse rhythm. For example, in English iambic tetrameter verse, omissions of metrical stresses occur almost as often as in the Russian or Ukrainian iamb, although English words are shorter than Russian or Ukrainian.[7]

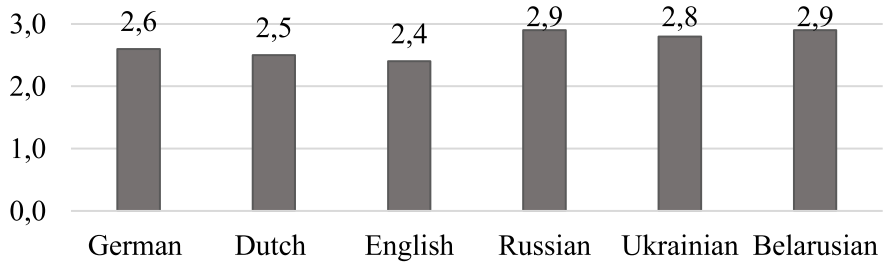


Figure 1: Average Length of Rhythmic Words[8]

[4] The corpus and the Prosimetron-system are in a state of formation, now the largest part of corpus, about 66%, is represented by Russian texts. An increase in the share of other languages is underway.

[5] Thus, a rhythmic word is identical to the concept of a phonetic word, for Russian and other East Slavic languages it is, as a rule, a complex of syllables united by one stress, for German, Dutch or English it is the same complex, but united by one main stress (Hauptton), for example *zur Strássenbahnlinie*. In this paper, speaking about the word, we will mean a rhythmic/phonetic word.

[6] For example, in Pushkins "The Captain's Daughter" among the three-syllable words, words with an accent on the second syllable prevail, their 16% of all words in the text, three-syllable words with an accent on the first syllable are only 6%, and three-syllable words with an accent on the third 9%. In Pasternak's "Doctor Zhivago" are practically the same figures: 15%, 7% and 8%. A similar picture is observed in longer words, with 4, 5 and 6 syllables, words with a non-extreme position of stress also prevail in them. And that is in all Russian texts.

[7] For more information, see Kazartsev's paper [3].

[8] The average length of rhythmic words in every language was calculated on the corresponding prose corpus in Prosimetron-system in the following way: the total number of syllables in prose corpus (of selected language) is divided by the number of rhythmic words in this corpus.

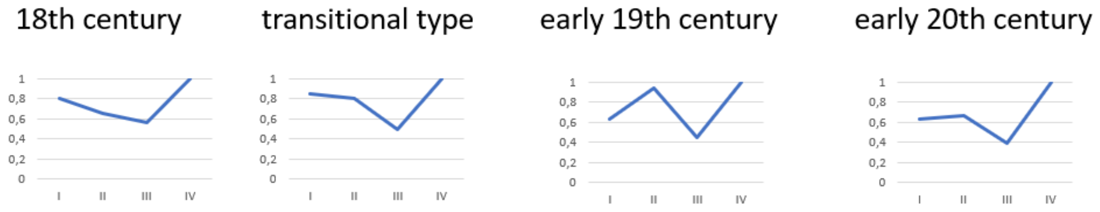
We have already said that the system allows one to consider individual words in all possible metrical and rhythmical contexts. For example, we can find all iambic tetrameter verse lines in which irregular, non-metrical stress on the odd syllables — first, third, and fifth — is present. It is known that, in principle, the continental model of metrical versification, unlike the insular English tradition, largely prohibits non-metrical stressing — that is, the replacement of iambic line fragments with, for example, trochaic ones. The Prosimetron allows us to see all cases when stress at the odd positions is present in a verse line of, for example, Russian and Belarusian iambic verse. Thus, if we look for all such cases in the Russian iambic tetrameter, we will see that, in general, examples of non-metrical stress are extremely rare around all strong positions, and only the first strong position is a place where a trochee may occasionally replace an iamb. In general, this is not surprising, because Russian versification, of course, inherits the continental model.

	Syllables in Iambic Tetrameter			
	1	**2**	**3**	**4**
Russian	4,7%	0,8%	0,7%	0,7%
Belorussian	4,9%	4,2%	3,8%	5,0%

Table 1: The Number of Violations on Weak Positions of Iambic Tetrameter

So, in the Russian iambic tetrameter of the 19th and 20th centuries, only about 5% of the lines have a stress on the first syllable, the rest - about 1% or even less. But further from the center of Europe through Russia to the south, this rule gradually loses its strength. If the same analysis is carried out for the Belarusian iambic tetrameter of the corresponding period, we will see that violations occur with approximately the same frequency — 4-5% — on all odd syllables. Weak positions in the iambic line here receive more stresses. That is, the purity in the realization of the continental model of iambic verse gradually dissipates.

Prosimetron allows us to identify, in the context of intercultural influences, rhythmic features inherent to a tradition, as well as stages of internal development. Thus, for the Russian iambic tetrameter, which was discussed above, one can show the evolution in how the number of stresses on strong positions changes; in other words, the evolution of the so-called verse stress profile (figure 2).


18th century

transitional type

early 19th century

early 20th century

Figure 2: The Evolution of Metrical Realization in Russian Iambic Tetrameter

These data, obtained automatically by Prosimetron, correspond to the data collected manually by Taranovskii in the 1930s and 1940s, and describe the most important stages in the evolution of the Russian iambic tetrameter (first published in 1953) [6]. They show the extent to which the rhythmical profile has varied. In some periods, it looks more like a frame. In some, it resembles a jagged streak of lightning: for instance, an alternating rhythm was typical for Alexander Pushkin's poems in the Russian tetrameter after 1825. Derived by analyzing many lines of verse, these statistics allow us to speak not only about the author's rhythmic originality, but also about some of the internal laws of the iambic tetrameter's development, including the tendency towards the distribution of strong ictuses (those heavily stressed). In addition, this technique can be used for the attribution of texts, since the rhythm of certain authors can be quite characteristic.

In the context of studying national literatures, Prosimetron makes it possible to obtain qualitatively new data. The more texts used in the analysis, the more unexpected results that can be obtained. Let's cite two specific examples. The complete rhythmic similarity of Lev Loseff's poem "Iosif Brodsky or an ode to 1957" and Alexander Pushkin's "Eugene Onegin" (1826) was completely unexpected. It would seem that if Loseff's text has a subtext, then it is necessary to look for it in Brodsky's work, but the rhythmic structure (extremely uncharacteristic for Loseff on average) coincides with the structure of Eugene Onegin and thus highlights a second important layer of sources for this text (see figure 3).

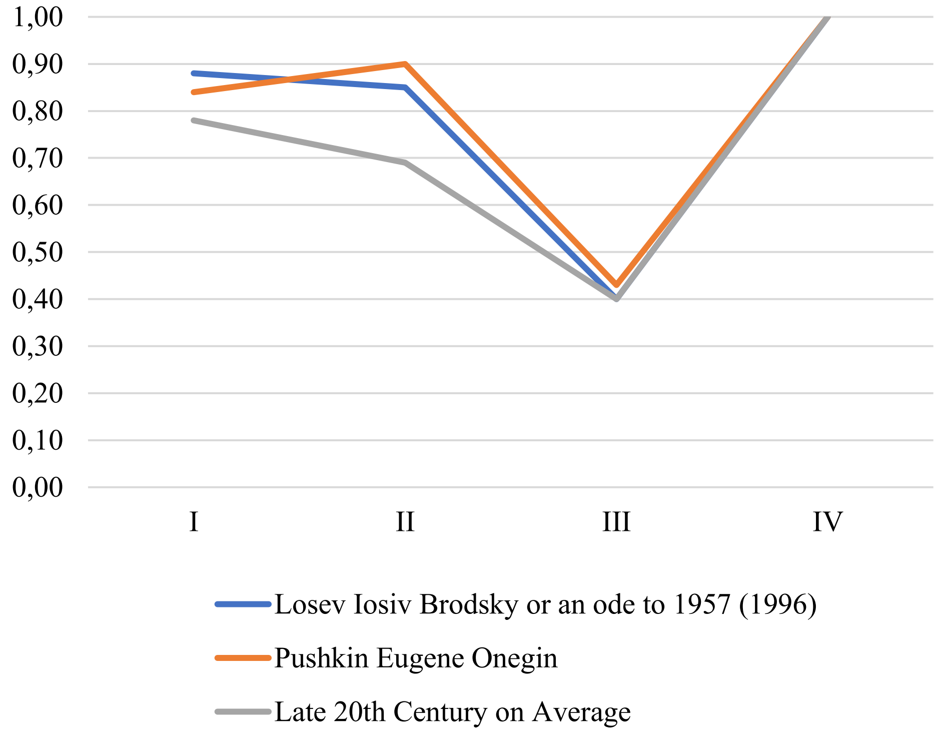


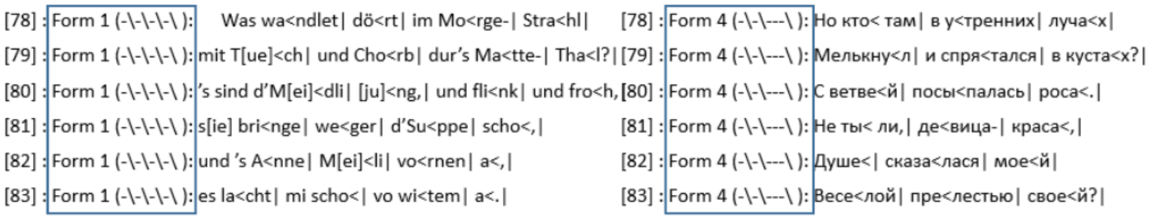
Figure 3: Three Stress Profiles of Russian Verse

These three graphs show that Lev Loseff's verse from the second half of the 20th century significantly diverges from the corresponding tradition of his era and turns out to be surprisingly close to Pushkin's verse from the mid-20s of the 19th century.[9] Of course, such parallels can be important both for philological interpretive work and for comparing the influence of different languages on each other, as well as for comparing different modes and styles of speech.

The next important example of a Prosimetron finding involves a comparative analysis of Vasilii Zhukovsky’s iambic pentameter in translations from Johann Peter Hebel. His translations of this particular Alemannic author are known to contain, for example, the first instance of Russian iambic pentameter without caesura, namely, the poem "Vergänglichkeit". It is this meter that will later be used in "Boris Godunov" and many other significant texts. But the rhythmic influence of Hebel on Zhukovsky actually runs much deeper. For example, in the text "Morning Star", Zhukovsky imitates the rhythm of Hebel directly, but functionally. What I mean is that where Hebel uses the same iambic rhythmic form for several lines, Zhukovsky also does not change the rhythmic pattern; where Hebel's form changes, Zhukovsky makes a change as well. This virtuoso game with rhythm is most likely not perceived by readers, but there is reason to believe that it was created by Zhukovsky, and it certainly influenced the developing Russian iambic verse tradition (a similar technique will appear later in the works of Pushkin and other authors).

[9] Despite the fact that the Loseff's poem is much shorter than Pushkin's verse, obviously, such a tendency could not have developed by chance, the centuries-old experience of Russian studies of metrics shows that, as a rule, the tendencies inherent in large poems are preserved in small texts. Lev Loseff copies Pushkin's manner not only in rhythm, but in the style of the text.

An Example of a Similar Rhythmic Structure of Hebel's Text and Zhukovsky's Translation:[10]



```
[78] : Form 1 (-\-\-\-\ ):    Was wa<ndlet| dö<rt| im Mo<rge-| Stra<hl|        [78] : Form 4 (-\-\---\ ): Но кто< там| в у<тренних| луча<х|
[79] : Form 1 (-\-\-\-\ ): mit T[ue]<ch| und Cho<rb| dur’s Ma<tte-| Tha<l?| [79] : Form 4 (-\-\---\ ): Мелькну<л| и спря<тался| в куста<х?|
[80] : Form 1 (-\-\-\-\ ): ’s sind d’M[ei]<dli| [ju]<ng,| und fli<nk| und fro<h, [80] : Form 4 (-\-\---\ ): С ветве<й| посы<палась| роса<.|
[81] : Form 1 (-\-\-\-\ ): s[ie] bri<nge| we<ger| d’Su<ppe| scho<,|              [81] : Form 4 (-\-\---\ ): Не ты< ли,| де<вица-| краса<,|
[82] : Form 1 (-\-\-\-\ ): und ’s A<nne| M[ei]<li| vo<rnen| a<,|                 [82] : Form 4 (-\-\---\ ): Душе<| сказа<лася| мое<й|
[83] : Form 1 (-\-\-\-\ ): es la<cht| mi scho<| vo wi<tem| a<.|                 [83] : Form 4 (-\-\---\ ): Весе<лой| пре<лестью| свое<й?|
```

At the moment, the Prosimetron system provides the ability to download all statistical data for a given subsample of texts sorted by date of creation, author, length, language or other parameter. If for Taranovsky it took half of his life to create such reference materials, this computer system makes it possible to process a larger amount of data over a predictable period of time. The possibility for constant rechecking, clarification, and expansion of data at any stage of the analysis minimizes the possibility for errors and their influence on the result.

This system also makes it possible to compare the rhythm of verse and prose. In this regard, the "rhythm of prose" means modeling the potentially possible rhythm of verse on the basis of the rhythmic variability of prose — that is, on the basis of a rhythmic dictionary, created through the use of the Kolmogorov model, also called the language model. Modeling, on the one hand, allows one to obtain comparable rhythmic data for two different modes of speech, that is, for the poetic and the prosaic, and on the other hand, it allows one to assess the differences that can rhythmically distinguish a poem from prose. One of the results of this work can be described as the confirmation of a previously stated hypothesis — the "prosaicization" of the Russian iambic tetrameter at the beginning of the 20th century — through analyzing a large volume of data from poems of the period. [6: 367–97]. The prosaicization of poetic rhythm in this context is defined as its gradual approaching the percentages in the language model; that is, the growing tendency to write rhythmically "non-poetic" poems.

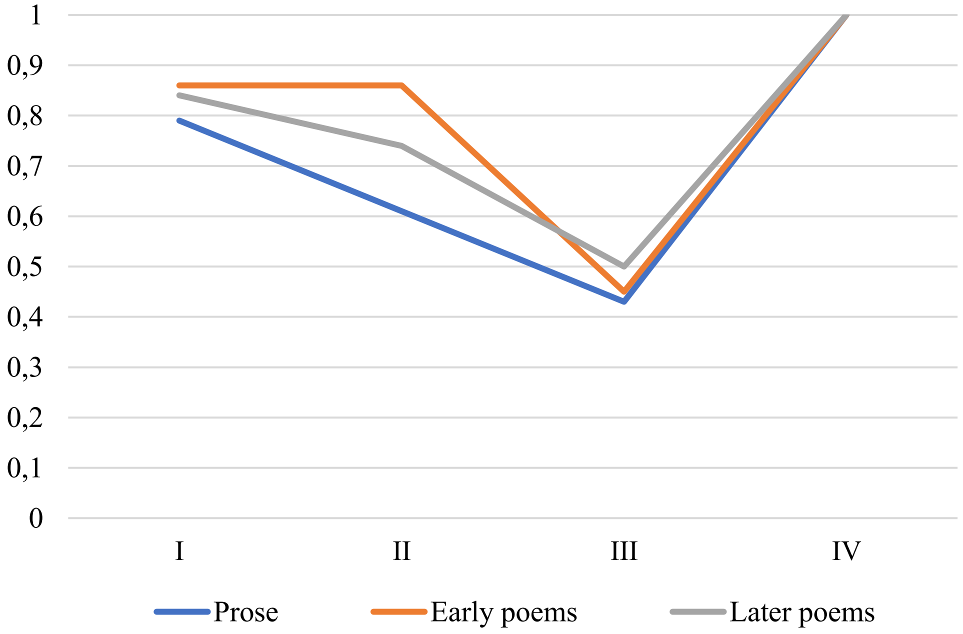


Figure 4: Rhythm of the Early and Late Poems of Boris Pasternak in Comparison with the Prose Prosody (Language Model)

Let us consider this phenomenon using as an example the rhythm of the poems of Boris Pasternak. Note the difference between the red and gray lines, which represent the rhythm of the early and late

[10] The sign "<" indicates in the Prosimetron the position of an accent, the sign “|” indicates a border between rhythmic words. Square brackets indicate that several graphemes denoting vowel sounds convey one complex phoneme, the case of diphthong, they should be calculated as a center of one syllable.

verses, respectively (see figure 4). While the rhythm of his early poetry is itself far from the typical verse of Pushkin's time and is an example of a transitional type, the rhythm of his later poetry is even farther from "poetic" rhythm and much closer to that of prose (this model is based on Pasternak's own prose). Interestingly, the model of speech based on the rhythm of occasional iambs in Pasternak's prose sharply moves away from his probability language model, and approaches the values found in his early verse and Pushkin's early poems [4: 64].

3 Conclusion

The Prosimetron-system, working with different languages, can test hypotheses about the proximity of verse rhythm to the linguistic preconditions, about the influence of one poetic tradition on another, and about the evolution of poetic forms within one of the traditions. The results presented here make it possible to see the utility of this new system for the comparative analysis of verse prosody. The system enables us to obtain data regarding the distribution of the rhythmical elements in different texts — in particular, the frequency of stresses on the metrically strong positions — as well as to simulate the rhythm of the verse according to the probability parameters for the distribution of rhythmical words. The analytical tools within this system allow for procedures of varying complexity — from the creation of dictionaries and stress profiles, to modeling and cross-language comparative analysis of rhythmic structures — and thus make it a truly comprehensive program that can be useful for both poetry theorists and linguists.

Acknowledgements

This publication was prepared as a result of research on the project no 20-04-023 supported by Research Foundation of HSE University in 2021, as a part of "5-100 Program" for leading universities of Russia.

References

[1] Bailey James. Toward a Statistical Analysis of English verse. — Lisse: Peter de Ridder Press, 1975.

[2] Kazartsev Evgeny. Computer Models of Verse Prosody. // CEUR Workshop Proceedings. — 2020. — P. 155–165.

[3] Kazartsev Evgeny. Language and Meter in the Early English, Dutch, German and Russian Iambic Verse. — Comparative Literature Studies, 2015. — Vol. 52(4).

[4] Kazartsev Evgeny. The Rhythmic Structure of «Tales of Belkin» and the Peculiarities of a Poet's Prose. — "A Convenient Territory": Russian Literature at the Edge of Modernity: Essays in Honor of B. P. Scherr. — Slavica, 2015.

[5] Kazartsev Evgeny. Comparative Study of Verse: Language Probability Models. — Style, 2014. — Vol. 48(2).

[6] Taranovskii K. F. (2010), Russian Iambic and Trochaic Verses. Articles about Verse [Russkie dvuslozhnye razmery. Stat'i o stikhe], Slavic Culture Languages [Iazyki slavianskoi kul'tury].