# Lexis meets meter: attraction of lexical units in Russian verse

Boris Orekhov[1][0000-0002-9099-0436]

[1] National Research University Higher School of Economics, Russia
nevmenandr@gmail.com

**Abstract.** The presented work deals with the quantitative parameters of the statistically calculated attraction of words to a number of meters in Russian poetry. This attraction tendency can be detected by means of Fisher's exact test. At the level of large corpora, where the meters are not differentiated by the number of steps, we failed to detect any semantic causes of the attraction. As for the prosodic structure of tokens, it does not explain the majority of the cases for which the statistics suggest attraction between the words and meters.

However, if we take the small subcorpora composed of lines of a particular meter differentiated by the number of steps, we are able to check whether the detected trends relate to the historical and literary reputation of a meter (for example, if this meter is often used for folkloristic styling, or for the theme of remembrance).

**Keywords:** Verse studies, Meter, Poetry, Statistics.

## 1 Introduction

Formal and quantitative approaches to the studies of poetry have a long history, particularly in Russian academic community. A poetic line can be efficiently described with numeric parameters, and mathematical methods of analyzing poetry have been used since the beginning of the 20th century. The Russian poet and a scientist Andrei Bely [Bely 1910] pioneered this research area in 1910 with his work on poetic symbolism. This methodology was further developed by B. Tomashevsky, G. Shengeli, the famous mathematician A. Kolmogorov and others. The final state of this methodology was reached in the works of M. Gasparov, who constructed the system of poetic terms and ideas and introduced linguistic methods into poetic studies. Thus, he laid the cornerstone of a new discipline, which is now called "linguistics of poetry". As a matter of fact, combining traditional linguistics with poetics is a nontrivial phenomenon in the world science. This approach has been called "the Russian method" since J. Baley (see [Korchagin 2011: 90]). Therefore, it would be fair to say that the recent trends in metrical analysis that are currently being employed worldwide merely follow in the footsteps of the Russian researchers and the work done by them in the last hundred years.

However, within the Russian body of research, there still remain less researched areas in linguistic poetics and metric studies. For example, the patterns of meter observed in Russian poetry are well-known [Gasparov 2000] (in addition [Smith 1985]); thorough analyses are available of poetic syntax and the morphological behavior of words in poetic lines [Gasparov & Skulacheva, 2004], as well as the description of different grammatical categories (e.g. verbal aspect and tense, agreement in person, etc.) [Kovtunova, 2005]. Due to the active development of lexicology in Russian linguistics, a large number of frequency dictionaries were produced that describe the quantitative parameters of lexis used in poetry [Shestakova, 2011]. Moreover, semantic phenomena in poetry, such as metaphors, metonymy and periphrases, are also well described (though not quantitatively) [Grigorieva & Ivanova, 1985]. At the same time, there is hardly any research which statistically explains the interaction of linguistic and poetic phenomena, since such investigations require expertise in computational methods.

## 2 Data

A corpus of poetic texts was recently compiled as a part of the Russian National Corpus project. This is a quite representative subcorpus, created by professional linguistics and literary scholars. More information on its creation can be found in [Korchagin 2015]. The texts in the corpus are manually annotated with specifically designed metric markup. The markup denotes the meter for every line and contains details about the number of stress foots found in the line and the clausula. The use of the corpus for the study of poetic language is advantageous in several ways. Firstly, it allows to analyze word distributions throughout a large collection, with works of different authors and from various epochs taken into account.

Secondly, the corpus makes it possible to study lexical distributions not only from the perspective of time and authorship, but also in relation to different meters.

We took a dump of this corpus for our research; the dump consisted of 9,693,341 word tokens and included a representative collection of Russian poetic texts dating from the 18th to the 20th (up to the 1930s) centuries. During this time period, Russian poetry mainly belonged to accentual-syllabic versification. This system of versification exploits the interchange of stressed and unstressed syllables as its main source of rhythmic organization. Thus, a poetic line can be described as a repetition of a number of steps, or syllabic groups, bound to the stressed syllable. There are five common combinations of syllabic groups, or meters – iamb, trochee, dactyl, amphibrach and anapest. The first two meters are called disyllabic because they consist of one stressed and one unstressed syllable. The three other meters are called trisyllabic, as they are made from one stressed and two unstressed syllables.

In the beginning of the 20th century, new rhythmic organization principles emerged in the Russian poetry. The most popular versification system in that time period was dolnik, a non accentual-syllabic form. This poetic form, although present in the corpus, was not included in the data for our research.

All in all, our data can be viewed as five independent subcorpora (we call it large corpora opposed to small corpora, where take into account the number of feets in each line), each one featuring only lines of a particular meter. These subcorpora are not of equal size: iamb is the most popular meter, comprising the largest part of the corpus, namely, 5,480,538 word tokens. The trochaic subcorpus consists of 1,593,554 word tokens, the dactylic of 402,434, the amphibrachic of 651,032, and the anapestic of 632,234 word tokens. Thus, the size of the syllabic-accentual part of the entire poetic corpus totals 8,759,792 word tokens, and we can notice that the majority of texts included into the poetic corpus exploit this versification system.

## 3 Goals and methodology

We aim to find the connections between meters and lexical units as manifested in the Russian poetry in the course of its history. Obviously, such connections are not binary: we cannot say that a random word can be found only in lines of a particular meter and will not be present in all other meters. Almost any word can be found in poetic lines of different meters. For example, the word *огонь* ‘flame, fire’ can be found in iamb:

*Свято`й* ***ого`нь*** *гори`т у вас в оча`х* [F. N. Glinka. Греческие девицы к юношам (1821)] ‘Holy fire burns in your eyes’

and in trochee:

*Дожига`й после`дние оста`тки*
*Жи`зни, бро`шенной в* ***ого`нь****!* [N. A. Nekrasov. «Ничего! гони во все лопатки...» (1854)]

‘Burn the last remnants of the life thrown into fire!’

and in all trisyllabic meters:

dactyl:

*В мо`ре не то`нет, в* ***огне`*** *не гори`т…* [L. A. Mei. Оборотень (1858)]

‘In the sea does not sink, in the fire does not burn’

amphibrach:

*И со`лнце пыла`ло на не`бе огне`м* [L. N. Trefolev. Маргаритка (1866-1889)]

‘And the sun blazed in the sky with fire’

anapest:

*В них ого`нь неземно`й*
*Жа`рче со`лнца гори`т!* [A. V. Kol'cov. Глаза (1835)]

‘Their ethereal fire burns hotter than the sun!’

As for the words with low frequency, the peculiarities of their functioning in texts cannot be revealed with quantitative approaches due to the insufficient amounts of data. The statistical methods we apply (see further) to our data cannot ensure reliable results on such amounts of word occurrences.

When we discover connections between certain lemmas and meters, that does not mean that a lemma invariably belongs to a particular meter. We rather consider this regularity as the “attraction” of the lemma towards some meter. In other words, a lemma can occur in lines of different meters, but still it demonstrates a clear preference for a certain meter.

Such tasks as finding connections between lemmas and meters, or a more general task of exploring relationships between several variables can be solved with Fisher’s exact test [Fisher 1922]. This test is successfully applied in linguistic research, namely, in collostructional analysis [Stefanowitsch & Gries 2003] which explores a similar phenomenon – the degree of connectivity between words, on the one hand and constructions, on the other. As in the case of meters, slots in constructions can be filled in with various words; however, Fisher’s test shows that the word distribution is not random, and we can say that some words tend to be attracted to certain constructions (for example, the verb сказать ‘to say/tell’ is attracted to the past tense [Rakhilina 2010: 37]). In the present research we will apply Fisher’s exact test to our task.

In particular, we aim to reveal statistically significant connections between lemmas and meters as well as between certain tokens and meters. The connections are discovered on the data from the large metric subcorpora (iambic, trochaic, dactylic, amphibrachic and anapestic, see “Data” above) and from smaller corpora of particular meter varieties (for example, the iamb with four steps = the iambic tetrameter, the iamb with five steps, etc.). Exploring meter varieties may be of interest to historians of Russian poetry and for philological research in general, because it is widely known that meters with different number of steps have different functions and belong to different genres. At the same time, the proposed approach is not entirely accurate from the point of view of statistics, as smaller corpora do not yield reliable results. Therefore, we take into consideration only words with 10 or more occurrences in the corpus. One more our goal is to find the factors underlying the attraction of words to certain meters.

## 4 Tokens and large corpora

We call a large corpora texts collections written in iamb or trochee or other meter without taking into account how many feet meter has in each line. After excluding the words below the frequency threshold of 10 items in the corpus we obtained the list of words for exploration containing 65,000 tokens. As many as 34,864 tokens among them did not show any signs of attraction to any specific meter. That means that p-value of their attraction to any meter was lower than 0.05. The attraction of other words to meters is shown in Table 1.

**Table 0.** Distribution of tokens with attraction to meters.

| Meter | Number of tokens | Size of the large corpus | % of all tokens |
|---|---|---|---|
| iamb | 8,735 | 5,480,538 | 13.4 % |
| trochee | 7,258 | 1,593,554 | 11.1 % |
| dactyl | 3,799 | 402,434 | 5.8 % |
| amphibrach | 4,454 | 651,032 | 6.8 % |
| anapest | 3,268 | 632,234 | 5.0 % |

Iamb has shown the largest number of tokens attracted to it.

Here you can see the examples of the different words, which differ one from another by part of speech, semantics, prosodic structure (see the acute sign on vowels) and the power of attraction (p-value).

пройду`т (Verb) number of occurrences in iambic lines: 178; number of occurrences in trochaic lines: 28; number of occurrences in amphibrachic lines: 21; p-value: 0.00031; кра`йней (Adjective), number of occurrences in iambic lines: 110; number of occurrences in trochaic lines: 23; p-value: 0.00026; серде`чно (Adverb), number of occurrences in iambic lines: 101; number of occurrences in trochaic lines: 26; p-value: 0.02383

All these tokens are found in iambic lines as well as in lines of other meters. However, these words have a statistically higher frequency in iamb.

Each meter has a different list of tokens that are attracted to it. These tokens belong to different parts of speech and vary in their prosodic structure. In the next sections of this paper we will analyze in more detail the prosodic structure of words attracted to specific meters.

# 5 Tokens and their prosodic structure

One of our goals is to find the factors underlying the attraction of words to certain meters, as we claimed earlier. These factors can be various in their nature: semantic, or structural, or other. It doesn't seem virtually possible to single out the common semantic feature for thousands of words that belong to different parts of speech. This means that if semantic factors worked, we would see the attraction to a specific meter of a specific semantic groups (e.g. verbs of motion or a qualitative adjectives).

At the same time, the structural factors seem to be the most natural explanation to the phenomena we investigate. This idea arises from the fact that Russian poetry itself is organized prosodically. Therefore, we formulate the following hypothesis: “the words in which the stress falls on the same syllable as in the line of a certain meter, are strongly attracted to this meter.” On the one hand, such a hypothesis is quite intuitive; on the other hand, cases when a word in a poetic line doesn't fit into its predicted position can easily be found in the corpus, e.g.:

***А`дский*** *| вихрь, что | пла`мя, | жгу`чий* [A. N. Maikov. «Воплощенная, святая...» (1888)]

‘The **hellish** whirlwind, like a flame, burning’

This line is written with trochee (the steps are denoted with vertical bars), and the token *адский* is attracted to this meter according to Fisher’s exact test. The null hypothesis was that the word *адский* is neutral in relation to all meters. But it was rejected with p-value 0.03045. It's instance in iambic lines is 66, in trochee 12, less than 2 for dactyl, amphibrach and anapest.

*Был* ***а`д|ский*** *зно`й, | но он | строчи`л | с утра…* [S. Ya. Nadson. Июль (1886)]

‘it was the heat of hell, but he had been writing from the morning’

This line is written with iamb (the steps are denoted with vertical bars), and it is obvious that the first and the second steps tear the token *адский* 'hellish' apart.

The poetic corpus contains the annotation of stressed syllables (more precisely, the annotation of ictus – the places with the predefined metric accentuation). This annotation allows us to verify the hypothesis.

We divided all tokens into the two groups: the words that match a certain meter in their prosodic structure, and those that don’t. Thus, the words *истле`вшие* (p-value 0.0163), *истле`ют* (p-value 0.00307), *исто`к* (p-value 0.02204), which are attracted to iamb, fall into the first group, and *и`стиной* (p-value <0.00001), *и`стину* (p-value <0.00001), *и`стины* (p-value <0.00001) fall into the second group.

In Fig 1 we show the total number of tokens that match and mismatch the prosodic structure of a meter.

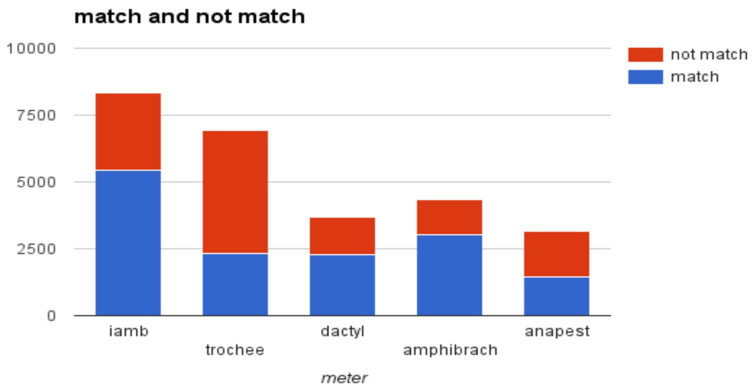

**Fig. 0.** A prosodic structure of a meter.

The comparison of different meters has yielded contradictory results. On the one hand, the hypotesis was supported by the data fro iamb and amphibrach, as for these meters the majority of the attracted words match with the prosodic structure of these meters. This is due to the fact that there is a lot of the words in the corpus that have with the stress on the second syllable are well represented in the corpus. On the other hand, for every meter there is at least a thousand cases in wich the prosodic structure of a word does not explain the attraction of this word to the given meter. Unlike iamb and amphibrach, in the trochee and anapest the hypotesis explains less than half of all the cases of attraction

## 6 Lemmas and large corpora

As the next step, we explored the behaviour of lemmas. We used the morphological analyzer mystem to bring all word forms of one lemma. In this case, we used it's option of disambiguation. We have got a sample of 25,900 lemmas. Approximately a half of all the lemmas were attracted to some meter.

This distribution is different from the analogous distribution for tokens: a larger number of lemmas lemmas are attracted to trochee than to iamb. It is possible that the distribution of tokens was biased due to the disproportion in the sizes of the subcorpora. The Pearson's correlation coefficient between the size of a corpus and the number of attracted tokens is equal to 0.873, whereas the correlation between the size of a corpus and the number of attracted lemmas is lower and equals to 0.643. The correlation coefficients for different meters are presented in Figure 2.

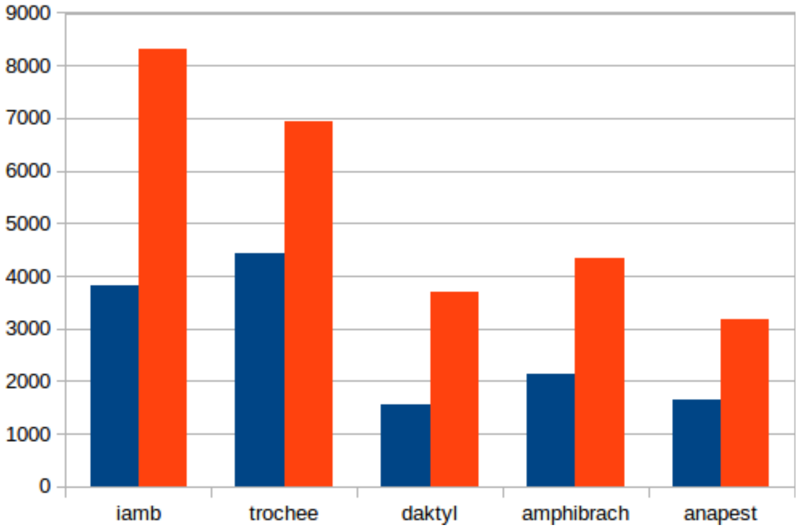

**Fig. 2.** A correlation coefficients for different meters.

## 7 Lemmas and subcorpora of metric variations

Despite the fact that splitting the corpora into smaller parts decreases the significance of statistical tests, it is necessary to compile the subcorpora of metric variations, because the four-step and the five-step variations of trochee are completely different from each other from the historical and the philological perspectives.

We applied Fisher's exact test to the meters with 3, 4, 5 or 6 steps. Only lemmas with the frequency above 10 were included in the analysis. As a result, we received 20 lists of lemmas attracted to a particular metric variation. The average number of lemmas in a list was 976.3, with the minimum of 298 (trochee 6) and the maximum of 2,427 (anapest 3). As many as 5,922 lemmas exhibited no attraction to meters.

## 8 Testing metric attraction on folkloristic tradition

It is not an easy matter to come up with an evaluation criterion in our research. What we explore is not a binary characteristic of a word, but a vague parameter - a word can be attracted to a certain meter, but it also can be encountered in other meters. There is no gold standard for this task and we have to invent our own way to evaluate the results.

Our data is Russian poetry, and our goal is to assess whether the identified tendency of words attracion correlated with trends in literary history. Some genre characteristics of the meters can be helpful in the process. It is widely known that trochee is strongly connected with the folkloristic tradition [Orlitsky]. Folkloristic texts are the texts that feature specific words inducing ethnic associations in readers' minds. Such words are annotated in dictionaries with the label “folkloristic, poetic”. We extracted all the words with this label from the Dictionary of the Russian language in 4 volumes – a total of 130 words out of the 82,266 are labelled as folkloristic in this dictionary.

If the distribution of attracted lemmas was random for all the meters, we would expect the folkloristic terms to be distributed uniformly. In fact, they appeared to be predominantly attracted to trochee 3 and trochee 4. This fact is illustrated by Table 2.

**Table 2.** The attraction of folkloristic terms towards various meters.

| An3 | An4 | Am3 | Am4 | Am5 | D3 | D4 | D6 | T3 | T4 | T5 | T6 | I3 | I6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 2 | 2 | 4 | 2 | 2 | 3 | 1 | 15 | 12 | 1 | 1 | 1 | 1 |

The only exception is anapest 3, which attracts more folkloristic terms than trochee. As the search in the corpus demonstrates, anapest 3 is also used to copy the folkloristic style and is underestimated by researchers in this respect, e.g.:

*Мне хоте`лось бы спе`ть про* ***кручи`ну,***  
*Чтоб* ***кати`лися*** *сле`зы из глаз.*  
[N. A. Klyuev. «Я поведаю миру былину...» (1907)]  
‘I would like to sing about **grief**  
so that tears **streamed** down from the eyes.’

Bold highlights a specific folkloristic terms.

The case of folkloristic terms shows that it is not impossible to find the system that governs the distribution of attracted lemmas. Such a system is more likely to have semantic and stylistic reasons underlying it, whereas structural features do not fully explain the observed distribution. This corresponds to the theory of semantic halo by M. Gasparov [Gasparov 2012]. The semantic halo is a property of a meter, when certain topics are anchored to this meter, but can be employed in other meters too.

M. Gasparov described several semantic halos, “remembrance”, which is traditional for amphibrach 3, being among them. Indeed, as our research has proven the lemma *помнить* `to remember’ is attracted to this meter. However, the detailed analysis of such cases lies beyond the scope of our work.

## 9 Conclusion

The presented work deals with the quantitative parameters of the statistically calculated attraction of words to a number of meters in Russian poetry. This attraction tendency can be detected by means of Fisher's exact test. At the level of large corpora, where the meters are not differentiated by the number of steps, we failed to detect any semantic causes of the attraction. As for the prosodic structure of tokens, it does not explain the majority of the cases for which the statistics suggest attraction between the words and meters.

However, if we take the small subcorpora composed of lines of a particular meter differentiated by the number of steps, we are able to check whether the detected trends relate to the historical and literary reputation of a meter (for example, if this meter is often used for folkloristic styling, or for the theme of remembrance).

These relations turn out to be rather strong, and the distribution of words in the case of meter attraction has semantic and stylistic reasons.

## References

1. K. Korchagin (2011). Sovremennye zarubezhnye issledovanija metriki [Contemporary metrics research]. In *Voprosy jazykoznanija*. № 4. pp. 90—115.
2. A. Bely (1910). Simvolism [Symbolism]. Moscow
3. Current trends in metrical analysis / [edited by] Christoph Küper, Wilfried Kürschner, Volker Schulz. Frankfurt am Main ; New York : Peter Lang, 2011
4. Gasparov, M. & T. Skulacheva (2004). Statji o lingvistike stikha [Articles on the linguistics of poetry] Moscow.
5. Fisher, R. A. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P". Journal of the Royal Statistical Society 85 (1): 87–94.
6. I. Kovtunova (eds.) (2005). Poeticheskaya grammatika [The poetic grammar]. Moscow
7. Stefanowitsch, A., & S. Gries (2003). Collostructions: Investigating the interaction of words and constructions. International journal of corpus linguistics, 8 (2), 209-243.
8. M. Gasparov (1979). Semanticheskij oreol metra. K semantike russkogo trekhstopnogo iamba [The semantic halo of a meter. Towards the semantic of Russian iamb 3]. In *Linguistics and poetics*. Moscow. pp. 282-308.
9. M. Gasparov (1980). Semanticheskij oreol trekhstopnogo amfibrakhija [The semantic halo of a amphibrach 3]. In Problems of structural linguistics. Moscow. 174-192.
10. M. Gasparov (1990). Semanticheskij oreol pushkinskogo chetyrekhstopnogo khoreja [The semantic halo of Pushkin's trochee 4]. In Pushkinskie chtenija: sbornik statej (eds. S. Isakov). Tallinn.
11. M. Gasparov (2000). Ocherk istorii russkogo stikha: metrika, ritmika, rifma, strofika [Essay on the history of Russian verse: metrics, rhythm, rhyme, stanza] Moscow.
12. M. Gasparov (2012). Metr i smysl [The meter and the meaning].
13. L. Shestakova (2011). Russkaja avtorskaja leksikografija. Teorija, istorija, sovremennost' [Russian author's lexicography: theory, history, contemporary situation]. In *Jazyki slavianskih kultur*.
14. Grigorieva, A & N. Ivanova (1985). Jazyk poezii XIX-XX vv. Fet. Sovremennaja lirika [The language of poetry of 19-20 centuries. Fet. Contemporary lyrics]. Moscow.
15. G. S. Smith The Metrical Repertoire of Russian Émigré Poetry, 1941-1970 // The Slavonic and East European Review, Vol. 63, No. 2 (Apr., 1985), pp. 210-227
16. K. Korchagin (2015). Poezija XX veka v poeticheskom podkorpuse Nacional'nogo korpusa russkogo jazyka: problemy reprezentativnosti [The poetry

of the 20th century in the poetic corpus of RNC: the representativeness problem]. In *Trudy Instituta russkogo jazyka*. Moscow. pp. 234-255.

17. E. Rakhilina (eds.) (2010) Lingvistika konstrukcij [The linguistics of constructions]. Moscow
18. Yu. Orlitsky (2009). Stikh "Kalevaly» v russkoj poezii i proze ["Kalevala" in the Russian poetry and prose]. In *ART* №3. http://www.artlad.ru/magazine/all/310/397/425/431