Development and Implementation of the Algorithm for Automatic Analysis of Metrorhythmic Characteristics of Russian Poetic Texts

Vladimir Barakhnin[1], Olga Kozhemyakina[1], and Irina Kuznetsova[2]

[1] Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
bar@ict.nsc.ru

Abstract. This paper presents the implementation of the program module responsible for the analysis of the structural level – the definition of the poem's metrorhythmics (meter, number of feet, and rhyme) in Russian poetic texts. The algorithm for determination of meter and number of feet takes into account the problem of the ambiguity of the placement of emphasis in homographs, possible omissions of schematic emphasis (pyrrhic), overlay of over schematic emphasis (spondee), which are solved by method "by analogy". The algorithm to identify the cases of shifting the emphasis from one part of speech to another (proclitic) is described. The algorithm of rhymes search is presented, the result of which is the definition of the stanzas of the poem.

Keywords: analysis of poetic texts, meter definition, metrorhythmic analysis, rhyme identification.

1 Introduction

The compilation of metric reference books to the corpus of poems is an important task of Russian literary researches. At the moment, the development of information technologies allows to automate the analysis of poetic texts, which in turn will reduce the amount of routine work of philologists. To solve this problem, it is necessary to develop algorithms for automating the analysis of the structural level of the poetic text, including the following metrorhythmic characteristics: meter, number of feet (foot – the basic unit of measurement of accentual-syllabic meter), rhyme.

Among similar works for other languages the system "SPARSAR" [1] can be identified, which provides automatic analysis of poetic texts in English and Italian. This system performs analysis at the quantitative, syntactic and semantic levels using NLP (Natural language processing). The obtained data are visualized in the form of developed schemes that allow comparing the works of one poet with each other and the works of different poets. The works of William Shakespeare, Thomas S. Eliot and Sylvia Plaza were used as a test sample. The program determined the size of their works with an accuracy of 90%. The system analyzed 500 poems, then the expert checked the accuracy of the analysis of 50 randomly selected poems. 5% of errors were detected.

In work [2] the static methods for the analysis, generation and translation of rhythmic poetry were used. The authors used machine learning without a teacher (unsupervised learning) to identify the patterns of stress placement in the number of poems to supplement in doubtful cases the proposed rhyme model. The authors also conducted experiments on the generation of English-language love lyrics and translations of Italian poetry into English with the preservation of the desired rhythmic scheme. The authors used 5 Shakespeare sonnets (70 lines) as a test sample; 81.4% of lines (57 lines) were correctly classified by metrorhythmics.

In [3], the authors have classified the texts according to the meter. They used an open source "the Scandroid" [4] to extract features (the program contains an algorithm for determining the stress in words and its own dictionary with accentuation of words – exceptions) and machine learning to classify poems by meter. Also in this work was implemented a module that defines the rhyme using the dictionary "The Carnegie Melon University Pronouncing Dictionary" [5] as a source of information about the pronunciation. To determine the accuracy of the work the sample of 205 poems was used, 88% of the words were correctly divided into syllables, the number of feet was determined with an accuracy of 99%.

In [6] the corpus of poems was analyzed using the "connectionist model" of poetic meter. It was shown that the prosodic picture of the poetic text is individual, and it is possible to determine the author by it, as well as that it reflects the aesthetics of the period of the author's creative work. As a test sample, a number of 1000 lines (100 lines by ten different authors) was used. The software package for statistical data processing SPSS Statistics was applied for the analysis [7].

In the work [8] it is described how to create software to determine the quantitative characteristics of the style of American poets and the visualization of a collection of poems in relation to each other. To visualize the obtained metrics, the authors used the principal component analysis and Classical Multidimensional Scaling.

It should be noted that it is impossible to design a universal system of automatic analysis of meter and rhythm, suitable at least for a group of more or less similar languages, because each language requires the development of its approaches, taking into account its structure. Such an experiment was conducted for similar in structure (Latin and Greek) languages, but even in this case, the study revealed features of languages, because of which their joint analysis is impossible [9, p. 52–54].

Finally, although in the analysis of the lower levels of the structure of the Russian verse the simplest mathematical approaches have been used for a long time – for example, numerous studies of the statistics of the types of Russian rhyme (including those applied to the temporal dynamics), generalized in [10], but often the collection of statistical information is still carried out almost manually (except for content analysis).

Some studies describing an integrated approach to automating the characteristics of Russian poetic texts (for example, [10]), affect, as a rule, very specific genres of poetry – for example, folk poetry, structural characteristics of which, such as metric, themes, etc., are significantly different from the corresponding structures in "literary" verse, or are of a rather private nature: the quantitative analysis of semantic associations of pentameter on the material of a number of Russian poets studied by M. L. Gasparov [11], the effect of metrorhythmic on semantics in the work of I. A. Brodsky – by M. Yu. Lotman [12], the metric halo's of "Black shawl" by A. S. Pushkin – by M. Wachtel [13], etc. The similar studies in relation to Czech poetry were conducted by M. Chervenka (see, for example, [14]).

For the analysis of the structural level of Russian poetic texts there are no practically implemented systems (at least in the open access), except for the pilot project of the system [15, 16], developed at the Institute of Computing Technologies of SB RAS.

The algorithm from [17] is at the core of this system, but it has several disadvantages, for example, does not take into account unequal meter of the poem, while in syllabic-tonic versification, there are meters, like the free amphibrach, choree, pentameter, characterized by a different number of feet in lines. In addition, the usage of the algorithm in its "pure" form, without taking into account the possible ambiguities of automatic accentuation, gives a relatively high percentage of errors in determining of the number of feet.

For these reasons, it was decided to implement the algorithm from [18], which includes a more strict classification of poems by meter. However, it does not affect such problems as ambiguous accentuation of homographs and clitics, so it also needs some modification. This study describes the development and implementation of the algorithm for analyzing the structural level: meter, number of feet, and rhyme; the modifications of the algorithm [18], features of its implementation and results are presented. The results of the work of two systems are compared: the one, which is developed on the basis of a modified algorithm from [18], and the one which is already existing on the basis of the algorithm from [17].

2 The Algorithm for Determining the Meter and Number of Feet

Let's describe the steps of the algorithm for determining the meter and number of feet, presented in [18]. The essence of this algorithm is to compare the rhythmic variants of the verse of the studied poetic text with a set of rhythmic patterns from a certain repertoire of metrorhythmic variants of the verse. This algorithm consists of five steps:

1. The pre-processing of a text. The lines of poetic text are numbered (St(n)), and PT={St(n)}, n=1, 2, …, N, where N is the total number of lines. All punctuation marks are deleted.
2. The accentuation is carried out on the basis of the dictionary A.A. Zaliznyak [19], which contains the accented word forms.
3. Each word of poetic text is divided into syllables and translated into a sequence of characters "c" and "C", denoting unstressed and stressed syllables respectively. The spaces between words are removed to display the syllabic scheme Sl:

$$\mathrm{Sl} =$$

$$= c_1 \dots c_{m(0)} C c_1 \dots c_{m(1)} \dots C_i c_1 \dots c_{m(i)} \dots C_{k-1} c_1 \dots c_{m(k-1)} C_k c_1 \dots c_{m(k)}, \quad (1)$$

where $c_{m(i)}$ is the unstressed syllable of the *i*-th word, *i*∈[0,*k*], *Ci* is the stressed syllable of the *i*-th word at *i*∈[0, *k*], *k* is the number of stressed syllables;

4. Scheme (1) is converted into syllabic rhythmic scheme Rs:

$$\text{Rs} = c^{r(0)} C_1 C^{r(1)} \ldots C_i c^{r(i)} C_{i+1} \ldots C_{k-1} c^{r(k-1)} C_k c^{r(k)}, \tag{2}$$

where *k* is the number of stressed syllables in the string, *R* is the number of all syllables in a line, *r(i)* is the interaccent interval, where *i*∈[1,*k*–1], *r(0)* and *r(k)* is the anacrusis and the clause. After that, the parameters *k, r(i), R-r(k)* are extracted, what further determine the type of rhythmic scheme of a poem.

5. The selection of the terms of classification of poetic text by metrorhythmic based on existing principles of versification. Depending on whether the parameters *k, r(i), R-r(k)* take constant or non-constant values for different foot lines, the authors formulate 5 classification conditions that can be implemented in Russian versification (Table 1).

Table 1. Classification of poetic text by metrorhythmic schemes

	R-r(k)	*k*	*r(i)*	Classification
1	≠const	≠const	arbitrarily	1. Metric verse in unequal feet 2. Dismetric verse
2	≠const	=const	arbitrarily	*k* - accentual verse in unequal/equal feet
3	≠const	=const	=const	*k* - foot syllable-tonic verse in equal feet
4	=const	=const	=const	*k* - foot syllable-tonic verse in strict equal feet
5	=const	≠const	≠const	isosyllabic verse in equal feet (proclitic, pyrrhic)

The first condition describes the often encountered type – syllabic-tonic versification with violation of the number of feet in the line – free syllabic-tonic versification. The second and third conditions of classification describe verses with specific metrorhythmics, and in this work will not be considered. The fourth condition classifies the verses with ideal meter without breaking the rhythm caused by omissions of schematic emphasis or overlay of over schematic emphasis. This kind of metrorhythmics are rare, because most poems contain alternating male and female rhymes, which automatically violate the condition of the constancy of the parameter *R - r(k)* (see Table 1). Most poetic texts contain disturbances in the rhythm; and are described in the algorithm of the fifth condition for classification as isosyllabic poems.

In the framework of this research the implementation and testing of the first and fifth conditions according to the classification of poetic text in metrorhythmic schemas was carried out.

3 Modification of the Meter and Number of Feet Detection Algorithm

This paper proposes a number of modifications to optimize the algorithm to improve the accuracy in the analysis of poetic texts.

Modification 1. The algorithm [18] assumes an "ideal" accentuation of words and completely ignores the existence of problems related to the omissions of schematic emphasis (pyrrhic). An example of pyrrhic can be shown in the quatrain of "Eugene Onegin":

Мой дядя самых честных правил,
Когда не в шутку занемог,
Он уважать себя заставил
И лучше выдумать не мог.

That quatrain will translate into:

cC cC cC cC c
cC cC cc cC
cc cC cC cC c

Only the first line of the scheme is strictly maintained (iambic tetrameter), and in others there are three accents, the one is missing. In the case of missing of the metric stress there is a special auxiliary foot of two unstressed syllables – pyrrhic (cc), which can replace the foot of iamb and of choree (e.g «Нет, не черкешенка она» from the "Answer to F.T." by A.S. Pushkin).

The algorithm [3] does not consider the overlay of over schematic emphasis (spondee), an example of which is illustrated by line «Швед, русский, колет, рубит, режет» from the "Poltava" by A.S. Pushkin with the scheme: cc cC cC cCc, the transfer of the emphasis from one part of speech to another (proclitic: «уронили мишку нА пол») and homographs («чЕстных», «честнЫх»).

These problems are solved by the method "by analogy", the idea of which in relation to this problem was expressed in [20]. The essence of the method is the following: lines and stanzas with ambiguous accent arrangement are compared with lines and stanzas, in the words of which the stress is placed unambiguously, and the choice of accent is made, providing the unity of metric characteristics for the whole poem.

To implement this method, when the poetic text is translated into a sequence of characters "c" and "C" (denoting unstressed and stressed syllables, respectively) in words with an ambiguous arrangement of accent, the positions of all possible variants of the accent are denoted by the symbol "x". Thus, the text is represented as a table of characters "C", "c", "x" of dimension n (the number of lines of poetic text) on m (the line with the maximum length).

Further, to eliminate the ambiguity of the accent arrangement in each line of the table, the element "x" is searched and the column of the table is taken by the index of this element. In this column, the most common single element is looked for and its value is assigned to the element "x".

The algorithm from [18] does not consider the proclitics (the pulling the accent on the preposition, for example, in the poem "Teddy Bear" by A. Barto – «уронили мишку нА пол»), so the database of proclitics on the basis of the dictionary of A. Zaliznyak [19] has compiled, To resolve the ambiguities associated with proclitics. It contains the information on the variants of accentuation of combinations of some words and prepositions. The text is analyzed for the presence of prepositions. If there is a preposition in the text, then the search for a combination of this preposition in conjunction with the word standing to the right of it is carried out. Upon detection of this combination in the database of proclitics, the information about the variants of accentuations in this combination is retrieved. In the case of ambiguous variants of the arrangement of accents, we again resort to the method “by analogy”.

Modification 2. The algorithm from [18] is sensitive to the parameters which it receives ($R – r(k), k, r(i)$), what leads to an incorrect definition of the meter and number of feet. Therefore, for the parameters the inaccuracies have been introduced: the parameters considered to be constant (=const), if they are constant for at least 90% of the lines of the poem.

Modification 3. The step of the algorithm from [18] is worked out in detail, specifically the fifth condition, the classification of poems according to metrorhythmic, which takes into account pyrrhic and spondee ($R – r(k)$=const, $k \neq$const, $r(i) \neq$const). If the text satisfies this condition, then further clarification to determine the meter and number of feet of the poetic text is made as follows. After each word is divided into syllables and translated into a sequence of characters “c” and “C”, the syllabic pattern is compared to the pre-compiled patterns, which are the foot of the intended meter. Using statistical evaluation the most suitable pattern from which the most suitable meter follows is revealed, that is the pattern with minimal difference from the syllabic scheme corresponds to a certain meter.

4 Approach to the Implementation of the Module Definition of the Rhyming Lines

In the article [18], in the algorithm for determining rhyme of poetic text it is suggested to seek rhyming lines in a poem, using the web app “Big rhyming dictionary” [21]. This app takes a word and returns a set of words rhyming with it.

Because the sending of requests to the web application for each word is a long process and the extraction of all the words and sets, and rhyming words to them with their following conservation in the database, is the process requiring big resources, in this paper we used an alternative way to search rhymed lines.

The rhyme search algorithm is implemented for reasons of the possibility of rhyme formation: the lines are rhymed if the last words in the line have the same position of the stressed syllable and the endings are phonetically coincided.

To identify the phonetically matched endings, we used data about the endings from the article [22]. It contains a pair of letter combinations, reflecting the sounds of rhyming verse endings from the literature of the 18–19 century:
[('и', 'ы'), ('и', 'ый'), ('ы', 'ый'), ('и', 'е'), ('и', 'ий'), ('у', 'уй'), ('ой', 'о'), ('кий', 'ки'), ('ей', 'е'), ('ай', 'о'), ('ой', 'а'), ('ей', 'и'), ('ий', 'е'), ('и', 'ьи'), ('и', 'ья'), ('ьи', 'ья'), ('и', 'ье'), ('е', 'ье'), ('к', 'г'), ('х', 'к',), ('г', 'х'), ('а', 'о'), ('е', 'и'), ('ья', 'ье'), ('ьи', 'ье'), ('ом', 'ым'),

('ит', 'ет'), ('ин', 'ен'), ('ий', 'а'), ('ой', 'а'), ('ый', 'а'), ('о', 'у'), ('уг', 'ок'), ('ах', 'ых'), ('е', 'ы'), ('ив', 'ов'), ('и', 'ой'), ('и', 'а'), ('я', 'и',), ('а', 'ы'), ('ы', 'у'), ('я', 'е'), ('ы', 'о'), ('ый', 'о'), ('ы', 'ой'), ('у', 'ой'), ('у', 'ый'), ('ы', 'ей'), ('ешь', 'ишь'), ('он', 'ен'), ('ел', 'ол'), ('ей', 'ой'), ('ом', 'ем'), ('ть', 'дь'), ('д', 'т'), ('ор', 'ер'), ('ом', 'им')]

With their usage, the algorithm for determining rhyme was developed, which consists of the following sequential steps:

1. The splitting of the text into stanzas.
2. The line numbering and extracting the last word from each line of the stanza.
3. The definition of accentuation for each word. The search for sets of words with the same accentuation.
4. The search for sets of words whose endings are rhymed (phonetically matched)
5. The search for the intersection of sets from p. 4 and 5 – the obtained rhyming lines.
6. The representation of rhymes in letter form: each set of rhymed lines is assigned a letter of the Latin alphabet, and the male – a lowercase letter, the female – a capital letter. On the basis of the received designations the letter sequence of lines in a stanza is made.

5 Obtained Results

As a test sample we have used a corpus of lyrical works of Alexander Pushkin (156 poems from the period of 1818–1825), pre-marked by meter, foot, rhyme with the help of reference book [23]. The algorithms for the definition of the meter and number of feet were tested (we modified the algorithm from [18] and implemented the algorithm from [17]).

Table 2. The comparing of algorithms for determining the metrorhythmic characteristics

Modified algorithm from [18]		Algorithm from [17]	
Meter	Number of feet	Meter	Number of feet
95,5%	95,5%	66,6%	57%

The accuracy of the determination of meter and number of feet in the modified algorithm [18] increased in comparison with the algorithm from [17] (Table 2). The difference in the work of the algorithms is due mainly to the fact that the first of the presented algorithms recognizes the metric versification with unequal feet, while the second incorrectly classifies it. Also, the sensitive parameters of the first algorithm made it possible to increase the accuracy in determining the number of feet in comparison with the already existing algorithm.

Among the disadvantages of this algorithm for determining the meter and stop we can underline: the limited database compiled on the basis of the dictionary A. Zaliznyak (a lack of some words), the replacing the letter “ё” by “е” in the test sample, what is allowed by the rules of the Russian language, but is critical for this algorithm, these problems are partially offset by the usage of the method “by analogy”. Further researches will address these shortcomings.

The main percentage of errors is due to the fact that there are unfinished poems in the corpus, in which one or more words are missing, they are replaced by supplemented ones, hereupon the standard algorithm that provides a full line does not work correctly. The further purpose of the study will be the identification of such lines to be excluded from the General analysis.

Also, this algorithm does not distinguish a meter with unequal feet and complex meter. This will also be done in a further study.

The module of the rhyme definition in modified algorithm of [18] has identified correctly 95% of the strophic patterns.

6 Conclusions

The article compares the work of two algorithms for the analysis of the structural level of Russian poetic texts: a relatively simple one from [17], which was used by us in the implementation of the pilot project of the system [16], and a more advanced algorithm from [18], modified by us, first of all, in order to eliminate possible ambiguities of automatic accentuation. It is shown that the modified algorithm from [18] gives a higher accuracy in determining the meter and number of feet, and also determines the strophic pattern with an accuracy of 95 %.

Further researches will be aimed at identifying typical problem situations that generate the errors in the work of the algorithm, and their subsequent elimination.

Acknowledgements

The study was carried with the support of the Russian Science Foundation (project No. 19-18-00466).

References

1. Delmonte, R.: A Computational approach to poetic structure, rhythm and rhyme. In: Proceedings of the First Italian Conference on Computational Linguistic in Pisa University Press, Vol. 1, P. 144–150 (2014).
2. Greene, E., Bodrumlu, T., and Knight, K.: Automatic analysis of rhythmic poetry with applications to generation and translation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. P. 524–533 (2010).
3. Tanasescu, C., Paget, B., and Inkpen, D.: Automatic classification of poetry by meter and rhyme. In: Proceedings of AAAI 2016. University of Ottawa (2016). Available at: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/viewFile/12923/12883 (Date accessed 06.07.2019).
4. The Scandroid. Available at: http://charlesohartman.com/verse/scandroid/index.php (Date accessed 06.07.2019).
5. The Carnegie Melon University Pronouncing Dictionary. Available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (Date accessed 06.07.2019).
6. Hayward, M.: Analysis of a corpus of poetry by a connectionist model of poetic meter. Poetics **24** (1), 1–11 (1996).

7. IBM SPSS Statistics. Available at: https://www.ibm.com/ru-ru/products/spss-statistics (Date accessed 06.07.2019).
8. Kaplan, D.: Computational analysis and visualized comparison of style in American poetry. Unpublished undergraduate thesis (2006). Available at: https://faculty.missouri.edu/~kaplandm/pdfs/KaplanBlei2007_ComputationalPoetryStyle_long.pdf (Date accessed 06.07.2019).
9. Mittmann, A.: Escansão automático de versos em português. Tesis (Doctorado), Universidade Federal de Santa Catarina. (2016).
10. Samoilov, D.: Book about Russian rhyme. M.: Hudozhestvennaya literatura (1982) (in Russian).
11. Gasparov, M.L.: Meter and meaning. About one of the mechanisms of cultural memory. M.: RSUH (1999) (in Russian).
12. Lotman, M.: "On the death of Zhukov" (1974). How does the poem of Brodsky work. In: The researchs of the Slavists in the West. M.: Novoe literaturnoe obozrenie. P. 64–76 (2002) (in Russian).
13. Wakhtel, M.: "Black shawl" and its metric halo. In: Russian verse: metric, rhythm, rhyme, stanza. M.: RSUH. P. 61–80 (1996) (in Russian).
14. Chervenka, M.: Meaning and verse. In: Works on poetics. M.: Yazyki slavyanskoy kultury (2011) (in Russian).
15. Barakhnin, V. and Kozhemyakina, O.: About the automation of the complex analysis of Russian poetic text. CEUR Workshop Proceedings **934**, 167–171 (2012) (in Russian).
16. Analysis of the poetic texts online. Available at: http://poem.ict.nsc.ru/ (Date accessed 06.07.2019) (in Russian).
17. Kozmin, A.V.: Automatic verse analysis in Starling system. Computer linguistics and intellectual technologies: Proceedings of the international conference "Dialogue 2006". M.: Publishing center of RSUH. P. 265–268 (2006) (in Russian).
18. Boikov, V.N., Karyaeva, M.S., Sokolov, V.A., and Pilschikov, A.I.: About automatic specification of the verse in the information-analytical system. CEUR Workshop Proceedings **1563**, 144–151 (2015) (in Russian).
19. Zaliznyak, A.A.: Grammatical dictionary of the Russian language. The changing word forms: about 10,000 words. 2-e ed. M.: Russian language (1980) (in Russian).
20. Barakhnin, V.B., Kozhemyakina, O.Yu., and Zabaykin, A.V.: The algorithms of complex analysis of Russian poetic texts for the purpose of automation of the process of creation of metric reference books and concordances. CEUR Workshop Proceedings **1536**, 138–143 (2015) (in Russian).
21. Big rhyming dictionary. Available at: http://rifmovnik.ru/docs.htm (Date accessed 06.07.2019) (in Russian).
22. Zhirmunsky, V.M.: Rhyme, its history and theory. Petrograd: Academia (1923) (in Russian).
23. Lapshina, N.V., Romanovich, I.K., and Yarkho, V.I.: Metrical Handbook for Pushkin's poems. M.; L.: Academia (1934) (in Russian).