

Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts

Vladimir B. Barakhnin^{1,2,*}, Olga Yu. Kozhemyakina^{1,2} and Alexey V. Zabaykin^{1,2}

¹Institute of Computational Technologies SB RAS, Novosibirsk, Russia

²Novosibirsk State University, Novosibirsk, Russia

Abstract. In this article we present the algorithms of the automated analysis of metrical, strophic and concordance characteristics of Russian poetic texts, realized in the form of processing software of the poetic texts, which can be an important learning tool for philologists in comprehensive analysis of the poems. The results of the such of analysis will allow to expand essentially the possibilities of the philologists who study as the listed levels of verses, as their semantic and pragmatic characteristics, and also to free the philologists from routine work, to expand the range of analyzed works by reducing the dependence of the quality of the comparative analysis on the personal knowledge of the researcher, and also to apply the different methods of intellectual analysis of data.

1 Introduction

In the learning process of the philologists in higher education institutions there is an obvious fact of using of large amounts of text. Also a multifaceted analysis of these texts that meets educational objectives is required. In this case the need for quantitative analysis of various components of the poetic space of the text often occurs. From this point of view indispensable instruments the tools in the form of special programs, which allows analysis of arrays of text for the purpose of obtaining results with a high degree of objectivity, are necessary.

Multi-component structure of the literary text implies the complexity of the objective analysis, due to the need of considering multiple data belonging to different levels of poetics. The text semantic level is definitely complicated for automated analysis, although it can be viewed as consisting of available automated analysis components. But information technologies are working quite objective when it comes to the other categories.

According to the traditional classification of V.M.Zhirmunsky [1], the main sections of practical poetics are the metrics, the stylistics and the theory of literary genres. The theory of literary genres, as shown in [1], is associated with semantics very closely: «the signs of the genre deals with all aspects of poetic work. These include the features of composition, of structure of poetic work, but also the features of the theme, i.e. original content, specific properties of poetic language (stylistics), and sometimes features of the verse... The concept of genre is always the historical concept and... the relationship of content elements (the subject) with elements of composition, language, and verse is... the typical and traditional unity which is historically established...» But «the classification of genres has not

always the logical character», and therefore the automated analysis of genre characteristics of the text, because of such criteria difficulties, is complicated (but not impossible).

But metrics and stylistics, «the components of a theory of poetic language» are another matter: «The artistic image in the poetry is created through the language. Therefore, we can say that the first division of poetics, its lower floor, should be focused on linguistics. Linguistic categories should form the basis of how we classify the phenomenas of poetic language...». The linguistic categories (poetic phonetics, poetic vocabulary, poetic syntax) are objectively studied using automated analysis, as the components of these categories, unlike semantic and thematic fields of text, carry a share of subjectivity, which tends to zero, and, accordingly, are reasonably analyzed automatically.

The levels of the structure of the verse represent a certain hierarchy (see e.g. [2]): meter, rhythm, phonetics, vocabulary, grammar, speech genre (composition-speech unity), theme, literary genre. Besides this the process of analysis of the verse provides for initial review of each level as an independent semantic unit with their subsequent mutual connection.

The purpose of this paper is to present the algorithms for the automated analysis of metrical, strophic and concordance characteristics of Russian poetic texts, as well as the software tools for processing of poetic texts [3] implemented on the base of these algorithms, which can serve as an important tool for philologists in learning process of comprehensive analysis of the poems.

* Corresponding author: bar@ict.nsc.ru

2 Background

Although some work in the field of research of influence of the lower levels of the structure of the verse at higher levels appeared in the first half of the twentieth century (for example, in the book of K.I.Chukovsky [4] among other things, there is a discussion about the influence of the vowel sounds in the poetry of A.Blok on their emotional characteristics), but the systematic study of such influence has begun, apparently, with works of K.F.Taranovsky, who made the report «On the interaction of poetic rhythm and subject matter» in 1963 at the Fifth Congress of Slavists, in which the interaction of rhythmic features of the genre and the usage of iambic chorea based on the analysis of several dozens of Russian poetic texts. It has been shown that in many poems written by this size (starting with «I go out alone on the road...» by Mikhail Lermontov), «the dynamic motif of the way is contrasted to the static motif of life» (see [5]). In this work a method of definition of the semantics of one or another poetic dimension, consisting not in study of its individual consumption but in study of the traditions of its genre and thematic use that involves the analysis of array of poetic texts, is presented.

Systematic studies in this direction were continued by M. L. Gasparov, which, in particular, showed [6] that «the number of meters in the verse culture is usually relatively small, the number of typical structure of the content is much more, so the same meter can be a sign of several and even many thematic sets. <...> In such cases, when we come to the poem, then, perceiving meter, we guess at once some set of its conventional thematic expectations, and perceiving vocabulary, we determine which option from this set elected by the author. <...> First of all the vocabulary creates for us the semantics of this particular poem, the metric – the general background of the semantic tradition on which it is perceived».

So, the study of the impact of the lower levels of the structure of verse on its higher levels is a very important problem of Russian Philology. One of the main difficulties in the decision of this problem is the need of analysis of the array of poetic texts of great volume. This task is extremely laborious, so often the researcher gets only a relatively small circle of the works of classical poets, what, without doubt, significantly reduces the completeness of the analyzed material and, accordingly, the reliability of the results. Thus, there is a task of the automation of the analysis of the different levels of structure of the verse, what should relieve the researchers from routine work and also dramatically expand the range of the analyzed authors.

The above-described correlation between the levels of structure of an unordered message and of a verse shows that many technologies and mathematical methods that used in Informatics can be used in the process of the automation of the analysis of poems.

Of course, the simplest mathematical approaches are using in the philological analysis of Russian poems quite a while ago. The frequency dictionaries of the language of classic poets are widely known. The numerous researches of statistics of the types of Russian rhymes (including, in relation to temporal dynamics) were

realized and summarized in [7]. Often, however, the collection of statistical information is still compiled almost manually (the only exception is content analysis).

Practically the only work in which a large program of the researches of metric, rhythmic and phonetic (including rhyme) characteristics of Russian poetic texts was outlined, is the article [8], based on the usage of the system STARLING [9]. This system contains, in particular, the web application for morphological analysis [10], created on the basis of the Grammatical dictionary of A.Zaliznyak. This web app is a morphological analyzer, that provides, in particular, the full-accentuated paradigm of each word presenting in the dictionary (unfortunately, the system does not allow to generate the paradigm of a randomly given word, and there is no phonetic analysis in it).

This programme of the studies of the characteristics of the verse was a part of the project «Automated lingvo- and-poetry- investigating analysis of Russian poetic texts», which was directed by S.A.Starostin, but after his death in 2005, the work on this project was practically discontinued.

Thus, the work of V.B.Barakhnin and O.Yu.Kozhemyakina [11], devoted to elaboration of approaches to automation of the complex analysis of Russian poetic texts, was of pioneer character. Meanwhile, we don't know the publications, which describe the usage of the relevant algorithms and of software tools, based on them for the purpose of linguists' education.

3 Algorithms of the analysis of the metric and strophic characteristics of poetry

At the analysis of metric and strophic characteristics of poetic texts, it is advisable to consider the following twelve parameters used in the preparation of metrical handbooks and concordances:

1. The quantity of verses excluding blank.
2. The metric of the poem.
3. The quantity of metric feet.
4. The rhyme scheme of strophics.
5. The quantity of the masculine endings of the last words in poetic lines.
6. The quantity of the feminine endings of the last words in poetic lines.
7. The quantity of the dactylic and other endings of the last words in poetic verses.
8. The quantity of the masculine endings without rhyme.
9. The quantity of the feminine endings without rhyme.
10. The quantity of the dactylic and other endings without rhyme.
11. The quantity of verses without the final words.
12. The type of strophic form:
 - the poems consisting of one stanza (eight verses or less);
 - right repeated stanzas;
 - free stanzas;

- paired rhymes;
- free rhymes.

Characteristics 1–4 are considered in accordance with the metric guide [12], the characteristics of 5–12 with concordance [13]. All listed guides are created on the verses of A.S. Pushkin, therefore we tested the following algorithms on them.

Apparently, the simplest option for automatic calculation is the number of verses (characteristic 1). However, there are some pitfalls: for example, in the poem «When outside the town, thoughtful, I walk...» 17th verse, for semantic reasons, is printed in the form of two half-verses (and, of course, an electronic version of the poem has exactly this form), but because of rhythmic reasons in all reference books this verse is considered to be a single that gives a discrepancy with automatic and with manual counting of verses.

It is possible to identify such features in the graphic reproduction of the verses in the subsequent analysis of the rhymes (a half-verse structure violates the metric and rhythm of the poem), but this situation (fortunately, very rare) would require the manual intervention of an expert.

A key challenge in the analysis of poetic texts is the definition of syllabic-tonic meters (characteristics 2 and 3). To do this it's necessary to select a poetic foot consisting of one accented syllable in a strong position and one or more unaccented.

In dependence of the position of the accent in the foot we differentiate for the two-syllable size – the pentameter (the accent on the even-numbered position) or the chorei (the accent on the odd-numbered positions), for the three-syllabic sizes – the dactyl (the accent on the 1st syllable), the amphibrach (on the 2nd syllable) and anapest (on the 3rd syllable).

For automatic determination of the metric structure of the poetic text we have used the algorithm described in [8]. The algorithm involves the construction of a numerical vector as follows: character 1 denotes the unaccented syllables, 2 – the accented syllables of monosyllabic words, 3 – the accented syllables which occupy the first position in two-syllable word, 4 – the accented syllables, which occupy the second position in two-syllable word, 5 – the accented syllables of words that are longer than two syllables. The derivable vector is parsed according to the following rules:

1. If there are on the odd-numbered positions the symbols 1 or 2? If there are – this is the pentameter.

2. If there are on the even-numbered positions the symbols 1 or 2? If there are – this is the chorei.

3. If there are on the positions 2, 5, 8... only the symbols 1, 2 or 3, on the positions 3, 6, 9... only the symbols 1, 2 or 4?

If there are – this is the dactyl.

4. If there are on the positions 1, 4, 7... only the symbols 1, 2 or 4, on the positions 3, 6, 9... only the symbols 1, 2 or 3?

If there are – this is the amphibrach.

5. If there are on the positions 1, 4, 7... only the symbols 1, 2 or 3, on the positions 3, 6, 9... only the symbols 1, 2 or 3?

If there are – this is the anapest.

6. If 1-5 are not done, and there is no sequence 111, this is the accentual verse with the number of unstressed syllables from 0 to 2.

The characteristic 4 identifies the type of the verse rhyming. For this purpose it is already required to obtain the phonetic information. The phonetic transcription is more necessary for a precise definition of rhyming verses, than the literal pairwise comparison (such rhymes which are called graphically exact, are represented only a small part of all rhyming). The first stage of phonetic transcription - the accentuation - is decided by us using the tools of automatic processing of the texts in natural language (Project AOD) [14], which was developed in the process of the creation of a system of automatic translation DIALING. Its dictionary contains about 3.5 million accentuated word forms, but, of course, this dictionary is still not complete.

The presented algorithms are implemented in the computer language Python 2.7 in the form of processing software of the poetic text [3]. The actual scheme of alternations of accented and unaccented vowels, which determines the rhythm, is shown on the display, what allows the learner to understand the difference between the rhythm and the «perfect meter» and also to study visually such concepts as pyrrhic, spondee, enclitics, proclitics, clause, etc.

Besides, in the poem processing, the log-file is created which shows the emergence of all the described above cases of ambiguity; along with this a separate table is written for the words that were not found in the dictionary of accents or for the words where the accent is ambiguous. On the basis of this table the student philologist can produce the addition of the word in used thesaurus or to choose the desired form of homograph.

For the phonetic analysis, we developed the module of phonetic analysis of words, which is based on sequential (order is important!) applying of the known rules of phonetics and orthography [15]. It should be noted that the phonetic transcription strongly depends on the accent in the word, so it is important to know the correct accent. Unfortunately, this is not always achieved because of the above-mentioned natural incompleteness of the dictionary of accents. However, the accuracy of phonetic analysis in these cases can be increased as follows. If the analysis of other strophes of the poem (in which there were no problems with the accentuation of the words) allows us to set its metro-rhythmic characteristics, then on the basis of this characteristics it's often possible to determine the accent in a word accent of a word which is not included in the dictionary and to make its phonetic analysis.

In the purpose of the definition of the type of rhyming when we divide a poetic text into quatrains, we use as basic variants of verification the enclosed rhyme, the couplet rhyme, the alternate rhyme and monorhyme. In the case of absence of the above-mentioned stanzas, the algorithm looks for a repeating structure of length up to 16 strings. So, in the case of Pushkin's poetry the maximum length of such a structure – 14 verses with the rhyme ababccddehhekk (the Onegin stanza).

The characteristics 5–7 noted in the reference book – the quantity of the endings of various types of rhyme

(masculine, feminine and other) for each of the poetic text. To determine the type of rhyme in automatic mode it is necessary to determine the accented vowel what is carried out using the above-mentioned dictionary AOT. A known problem of automation is the inability of choosing the correct homograph in case of different types of homography (case, between the parts of speech, etc.). If at the end of the verse is the word, for which there are different kinds of accents, we do not consider this verse and mark it as incorrectly defined. It is assumed that the linguist can manually choose the correct shape of homograph or, in the case of absence of the word in the dictionary, to add the word in used thesaurus.

The characteristics 8–10 (the quantity of the endings without rhyme of the last words in the verses of different types) are defined similarly to the characteristics 5–7 considering the type of the rhyming. If the structure of the verse is set, it is not difficult to find the quantity of non-rhymed endings. A more complicated situation is when the analyzing poetic text is belongs to the category of free stanzaic structure. In this case, the binding of the rhyming endings is searched in a certain range, usually no more than 7.

The quantity of verses without the end words (characteristic 11) is determined by identifying of the verses that stand out from the overall metrical structure by a smaller number of syllables.

Finally, a type of strophic form (characteristic 12) follows from the rhymes of the stanza structure (characteristic 4).

The using in the software of the listed algorithms allows the learners to learn on practical examples the approaches to the classification of rhymes in Russian poetry.

The building in the automatic mode of the concordances is rather trivial. The main problem is to separate the homonyms (homographs) and to relate them to the correct sets of lexemes. Now while solving this problem we see no alternative to the work of the linguist (in practice, quite a competent native speaker, in the first place – the student-philologist) in manual mode using a convenient software interface.

4 Analysis of lexical and grammatical characteristics of poems

The lexical analysis of the poem provides [2] the creation of its lexical dictionary, which is used, in particular, to identify the dominant parts of speech, thematic (semantic) fields and poetic phraseology (primarily used metaphors).

For the identification of thematic fields it is advisable to use an electronic dictionary of synonymous arrays. Many dictionaries of this kind, placed on the Internet and used by specialists in the field of information technologies, unfortunately, are usually not approved by professional linguists, this makes them impossible to use them in philological researches. Therefore, we developed and implemented an algorithm of conversion of the arrays of synonyms of «The dictionary of the synonyms

of Russian language in two volumes», compiled by the staff of the dictionary sector of the Institute of Russian language of the Academy of Sciences of the USSR headed by A.P. Evgenyeva [16], in a relational database; as a result the database with over 4,000 synonymic series (with a total of about 10,000 words) was developed.

Grammatical analysis of the text includes the definition of its possible belonging to nominal or verbal styles (respectively only the denominative sentences or the listing of actions), as well as of the temporal plan and subjective structure of the poem (what requires the studies of the usage of the categories of tense, voice and person).

Nominal or verbal style is determined by direct analysis of the lexical dictionary. For the definitions of usage of the categories of tense, voice and person is also required to use fairly simple morphological rules of the Russian language, allowing to establish the category of tense, voice or person was used.

5 Unresolved problems of automation of analysis of the highest levels of the poetic texts

The direct identification of the theme of a poem is a task that is very difficult for an automated solution, because it requires semantic analysis of the texts at a level, which is close to the perception of natural-language texts by humans. However, the research of the correlation of the theme with the lower levels of the structure of the verse is one of the least explored areas of philological analysis. In this area there is a number of unsolved problems, some of them are formulated in [4]:

«The question about the correlation of metro-rhythmic level of the text with its subject matter, is still debatable...

The method of identification of the semantic connotation of rhythm have not been enough developed until today...

This question [about theme, image and emotional associations connected with those or other sounds – ed.] is under development, and while we can't give a brand indisputable characteristics of the semantics of each sound».

The application of methods of statistical analysis of large arrays of poetic texts could be an effective method of resolving of these and similar problems of philological analysis.

An important area of the researches is the usage of multivariate analysis of semantic, emotional, etc. associations, a large-scale use of which is practically impossible without the application of methods of the automation.

Exactly the problem statement, which is formulated above, allows the students to see the wide perspectives of the usage of computer and mathematical methods for the complex analysis of Russian poetic texts.

6 Conclusion

In this article we present the algorithms of statistical analysis of lower structural levels (meter, rhythm, phonetics, vocabulary) of Russian poetic texts, also we describe the software tool for the processing of the poetic texts which is developed on the base of these algorithms and can be an important instrument of philologists' learning of the complex analysis of poems. The results of the such of analysis will allow to expand essentially the possibilities of the philologists who study as the listed levels of verses, as their semantic and pragmatic characteristics, and also to free the philologists from routine work, to expand the range of analyzed works by reducing the dependence of the quality of the comparative analysis on the personal knowledge of the researcher, and also to apply the different methods of intellectual analysis of data.

References

1. V.M. Zhirmunsky, *Introduction to literary studies: Lecture* (St. Petersburg University Press, St. Petersburg, 1996) (in Russian)
2. D.M. Magomedova, *Philological analysis of lyric poems* (Publishing center «Academy», Moscow, 2004) (in Russian)
3. The analysis of the poetic texts online - <http://poem.ict.nsc.ru>
4. K. Chukovsky, *Alexander Blok as man and poet* (A.F.Marx, Petrograd, 1924) (in Russian)
5. K. Taranovsky, *About the relationship between poetic rhythm and topic*, (Languages of Russian culture, Moscow, 2000) (in Russian)
6. M.L. Gasparov, *Semantic aureole of the meter: the semantics of Russian iambic trimeter*, In: *Linguistics and poetics* (Nauka, Moscow, 1979) (in Russian)
7. D. Samoilov, *Book about Russian rhyme* (Khudozhestvennaya Literatura, Moscow, 1982) (in Russian)
8. A.V. Kozmin, *Automatic analysis of verse into the Starling system*, In: *Computational linguistics and intellectual technologies: Proceedings of the international conference «Dialogue 2006»* (Publishing center of the RSUH, Moscow, 2006) (in Russian)
9. *The Tower of Babel. An etymological database project. Russian dictionaries and morphology* – <http://starling.rinet.ru/indexru.htm> (in Russian)
10. *The morphological analyzer* <http://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win> (in Russian)
11. V.B. Barakhnin, O.Yu. Kozhemyakina, *About the automation of the complex analysis of Russian poetical text*, *CEUR Workshop Proceedings*, 934 (2012) (in Russian)
12. N.V. Lapshina, I.K. Romanovich, B.I. Yarkho, *Metrical guide to the poems by A.S. Pushkin* (Moscow; Academia, Leningrad, 1934) (in Russian)
13. J.T. Shaw, *Pushkin: A Concordance to the Poetry: 1, 2* (Slavica, Columbus, 1984)
14. *Project AOT* - <http://nlpub.ru/AOT> (in Russian)
15. *Rules of Russian orthography and punctuation. Full academic Handbook* (Eksmo, Moscow, 2007) (in Russian)
16. *Dictionary of synonyms of Russian language: in 2 volumes* (Nauka. Leningrad department, Leningrad, 1970 - 1971) (in Russian)