The algorithms for complex analysis of the corpuses of poetic texts in the Kazakh language

V B Barakhnin¹, ², O Yu Kozhemyakina¹, A M Bakiyeva² and M K Sodboev²

¹ Institute of Computational Technologies of SB RAS, Lavrentiev av. 6, Novosibirsk, 630090, Russian Federation
² Novosibirsk State University, Pirogov Str. 1, Novosibirsk, 630090, Russian Federation

bar@ict.nsc.ru, olgakozhemyakina@mail.ru, m_aigerim0707@mail.ru, matveysodboev@gmail.com

Abstract. This paper outlines the main approaches to the automation of the process of statistical analysis of the lower structural levels (meter, rhythm, phonetics, vocabulary, grammar) of poetic texts, as well as the algorithms for complex analysis of poetic texts in the Kazakh language are presented in order to automate the process of creation of metric dictionaries and concordances. The results of this analysis will significantly expand the possibilities of the philologists who study both these levels of poetry and their semantic and pragmatic characteristics, and also will rid the philologists of routine work, will expand the range of analyzed works, by reducing the dependence of the quality of comparative analysis on the personal erudition of the researcher.

1. Introduction

The multicomponent structure of a literary text implies the complexity of objective analysis, due to the emerging need to take into account the set of data belonging to different levels of poetics. The levels of the verse structure represent a certain hierarchy (see, for example [1]): meter, rhythm, phonetics, vocabulary, grammar, speech genre (compositional-speech whole), thematics, literary genre. At the same time, the process of verse analysis provides for the initial consideration of each level as an independent semantic unit with their subsequent interconnection.

There is a certain correlation between the levels of the structure of an arbitrary message and a verse: meter, rhythm and phonetics correspond to the syntactic level, vocabulary and grammar correspond to the semantic level. As for the thematics, in relation to the analysis of poems, it refers not only (and in many ways not so much) to the semantic, but also to the pragmatic level, since in the analysis of the lyrical poem, the analysis of the thematics often includes the study of the emotional impact on the reader. Finally, the analysis of genres (speech and literature) implies the study of messages within the poetic text: because, according to [1], the speech genre implies not only a certain type of speech subject, but also the same specific type of speech addressee, the interaction of the speech subject and the speech addressee creates the specificity of a literary genre.

The systematic study of the influence of the lower levels of the verse structure on the higher levels began, apparently, with the works of K. F. Taranovsky, who presented in 1963 at the Fifth Congress of Slavists the report "About the interaction of poetic rhythm and thematics", in which, based...
on the analysis of several dozen Russian poetic texts, the interaction of rhythmic features and genre application of the pentameter chorea was studied. It has been shown that in many verses written by this size (starting with "I go out alone on the road... by M. Yu. Lermontov), "the dynamic motive of the way is opposed to the static motive of life" (see [2]). In this work the technique of definition of the semantics of this or that poetic size consisting in the study not of its single usage, but of the tradition of its genre and thematic usage [1] that assumes the analysis of corpuses of poetic texts of large volume was offered.

The purpose of this work is to present the main approaches to the automation of the processing of poetic texts in Kazakh language, written with the use of syllabic-tonic poetry. It should be noted that Abay Kunanbayev was an innovator in the field of Kazakh syllabic style [3]. He expanded the theamics of Kazakh poetry, approved in the literature the new genre forms, perfected the expressive means of verse, the intonation and rhythmic variety of poetic sizes, introduced the new flexible poetic forms. His reforms are a whole complex of smooth overflows, increasing dynamic rhythmic and logical stresses and accents, i.e. an alloy of activity, plasticity and smoothness of the verse, where the rhyme can not absorb the meaning of the content and merges with it. In the language of Abay, for example, the verbal form of the мак/мек, which exists in many Turkic languages (ex. Uzbek – корма, Turkey – гормек, Tatar – курмэк, Uighur – курмак) [4] received a wide field of usage with a special stylistic colouring.

The monograph [5], which withstood two editions, examines the questions of great importance for the study of the art form of poetic works in the Kazakh language, their poetic language and style, the problems of art skill. The establishing of the rhythmic communication with the various aspects of poetic form is very important – with the intonation, the syntactic construction of speech, its sound structure, because the rhyme in the Kazakh verse is characterized by exceptionally high proximity of the sounds, the fullness of the harmony.

Thus, a comprehensive study of poetic texts in the Kazakh language and the creation of computer algorithms for the automation of this process, including the analysis of large volumes of poetic texts, is a very actual task of computer linguistics.

2. History of question

The appearance of the first publications in the field of automation of the analysis of meterrhythmic characteristics of poetic texts dates back to the middle of 1990s: M. Hayward conducted a computer study of the features of metrics in the poems of various poets [6]. But the mass interest in the research in this area appeared only in the late 2000s-early 2010s. Thus, D. Kaplan and D. Blay analyzed the styles of American poets on the basis of spelling, syntactic and phonetic characteristics of their poems [7]. A similar study in order to identify the features of the style of professional poets that distinguish them from the style of amateur poets, done by John.Kao and D.Jurafsky [8]. The system SPARSAR which was described in the work of R. Delmonte [9], includes the automated comprehensive analysis of poetic texts with the purpose of studying their style. Web application Metricalizer2 [10], developed by K. Bobbenhausen and B. Hammerich, allows to do the automatic analysis of the metric characteristics of the poems in German.

The series of publications of the authors of this article, V. B. Barakhnin and O. Yu. Kozhemyakina, is devoted to the study of automation of processes of complex analysis of poetic texts in the Russian language, for example, [11, 12].

Among the works of such thematics which are relating to the analysis of poetry, including in Turkic languages, we can mention the article by A. Kurt and M. Kara [13], where the algorithm of recognition and analysis of poems written in a special, characteristic for east (Arabic, Persian, Turkish) poetry, the system of poetry arud, is proposed.

It should be noted that the algorithms of analysis of the lower levels of poetic texts strongly depend on the characteristics of a particular language, as well as on the degree of development of computer technologies for this language. Thus, the peculiarities of phonetic analysis of texts in English are poor paradigm of word changes and the presence of a large number of network dictionaries of phonetic
analysis, in German – quite simple and strict rules of morphological changes of words and phonetic characteristics of word forms, in Russian – the absence of any general rules of morphological changes of words and their phonetic characteristics (primarily emphasis), in Turkish and other Turkic languages, including Kazakh, – almost deterministic rules of the formation of the word forms and of the change of their phonetic characteristics following from the law of synharmonism which is characteristic to Turkic languages.

3. **Algorithm for extraction of metrorhythmic characteristics of poetic texts in the Kazakh language**

According to its typology and morphological structure, the Kazakh language belongs to the group of agglutinative languages. This means that new words are formed by sequentially attaching of the word affixes (grammatical suffixes and endings) to the root or base.

Let’s note that for the Kazakh language the algorithm of phonetic analysis is much simpler than for the Russian language, since in the Kazakh language the unstressed vowel sounds are not reduced, and the process when the end consonants become voiced and voiceless is reflected in writing by direct replacement of the corresponding letters.

The emphasis in Kazakh, as in other Turkic languages, is fixed and is put on the last syllable of the word.

The exceptions, according to [14, 15] are, that is, do not take the emphasis:
- suffixes of adjectives дай/дай, тай/тей (adjectives of similarity);
- suffixes of verbs шы/ши;
- negative suffixes ба/бе, па/пе, ма/ме;
- personal endings мын/мн, сын/син, сындар/сіздер etc.
- the end of the ablative мен, пен, пенен, менен;
- words that refer to generalizing pronouns: барша, күллі, барлық, etc.
- words that are used in the dialogue in the form of confirmation, consent: мейлі, болды, мақұл, etc.

On the basis of this rule, the algorithm was developed for accentuation (placement of emphasis) and search for metrics and rhymes in Kazakh poems.

Let's note that Python 3 was chosen as the programming language with the use of the Pycharm development environment, since this development environment has a built-in debugger, which is not in the standard development environment from the official Python site.

Let’s consider the algorithm of splitting words into syllables. The algorithm is based on the fact that there are 3 types of syllables in the Kazakh language [16]:
- open syllable which consists of one vowel or begins with a consonant and ends in a vowel, such as а-же;
- a semi-closed syllable that begins with a vowel and ends with a consonant (one or more), such as ант;
- a closed syllable, in which a vowel is enclosed between consonants, e.g. сан, қант, сарт, спорт, спринт.

On the basis of these rules, a software module was developed, implemented as a function, in which a word is applied to the input, and an array of syllables is get in the output.

Thus, not the whole word, but a set of syllables of the word are given to the input of the program module, which carries out accentuation. The module checks whether the word is an exception from the general rules of emphasis. The return values are the position of the beginning of the stressed syllable, its length, the number of syllables after the stressed syllable.

For automatic determination of the metric structure of the poetic text we have used the algorithm described in [17] and implemented by us for the Russian language in [11]. 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 chorey.
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.

Figure 1 shows the result of the software application that determined the iambic dimeter.

![Figure 1](image1)

Figure 1. The work of the software application to determine the size.

For the determination of the type of rhyme in stanzas when the poetic text is splitting in quatrains, as the basic variants of checking the rhyme we distinguish:
- a ring rhyme: the first line rhymes with the fourth, the second with the third;
- a related rhyme: the first line rhymes with the second, third with the fourth;
- a cross rhyme: the first line rhymes with the third, the second with the fourth.

The rhyming lines are denoted by the same letters.

In the absence of these types of stanzas, the algorithm tries to check all possible variants of rhyming lines. Figure 2 shows the finding of the enough complex stanzas: the second line rhymes with fourth, the fifth with the sixth and the eighth.

![Figure 2](image2)

Figure 2. The definition of the schema of emphasis and rhyming of the lines.

In addition, the system determines the number of male (rhyming line ends with a stressed syllable – the situation is most typical for poems in the Kazakh language) and female (rhyming line ends with an unstressed syllable and the penultimate stressed syllable) rhymes.
The algorithm of the creation of the frequency dictionaries and concordances

An important task of the automation of the processing of poetic texts is to develop the algorithms for the creation of the dictionaries of the language of a poet, which include all the words used in the languages of this poet, without taking into account word forms, and the concordances, which take into account all word forms. The stemmer [18], which extracted the unique words, has been implemented in the generated software system with the aim of obtaining the dictionaries of the language of poets. The work of stemmer is based on Porter's algorithm [19], as well as algorithms of morphological analysis and synthesis of words of the Kazakh language from the work [20]. The task of stemming the words of the Kazakh language was divided into two subtasks: 1) the stemming of the verb forms of the word; 2) the stemming of the other parts of speech.

The verbs play a central role in a sentence because the basic information (time, negation, question, person, and number) is shown through this part of speech and its affixes. In agglutinative languages, the verbs play an important role, while in Kazakh they have a very strict position in the sentence: they stand at the end of the sentence or, if it is a complex sentence, before the comma. Thus, first the stemming of the verbs is realized and then of the nouns and adjectives (of course, the ambiguous situations are considered in the implemented algorithm, where the basis ends with a certain combination of letters that can act as an affix). An example of the stemming can be seen on figure 3.

Figure 3. An example of the work of stemmer.

The building in automated mode of the concordances is rather trivial. The main problem is the division of homonyms (homographs) and their assignment to the necessary sockets of lexemes. Currently, when we solve this problem, we do not see an alternative to the work of a linguist (in practice – a sufficiently competent native speaker) in manual mode using a convenient software interface.

4. Algorithm testing

The algorithm was tested on the body of Abay's poetic texts. About 40 poems were used for testing, some of which were the translations of works by Russian poets (for example, the novel in verse “Eugene Onegin” by A. S. Pushkin).

In the process of testing it was found out that the Kazakh syllabic-tonic versification is somewhat different from the European one, i.e. the metrics known to us, such as yamb, chorey, etc., are poorly tracked. On figure 4 we can notice that the first line rhymes with the second and third, the metric of the form:”+-+--+-+-+”(yamb-yamb-anapest yamb-yamb), however it is broken in the first line. It should also be noted that foreign names (Macedonia) retain their emphasis, but at this stage the system does not take this into account (because there is no dictionary of borrowed proper names).

Here we can notice that there are cross-rhyme lines, but there is no systematic metrics again.

It becomes clear in the analysis of male and female rhymes that in the Kazakh versification the male rhyme (line ends on a stressed syllable) dominates with a large margin, but in the translations of Pushkin we see other situation. The fact is that the stress in the Kazakh language almost always falls on the last syllable of the word, and this is a male rhyme.
Figure 4. Analysis of the poem “Ескендір” (“Alexander”).

Let’s consider another example. Figure 5 shows the translation of the letter “Tatyana to Onegin” (“Татьянаның Онегинге жазған хаты”) from the novel “Eugene Onegin”.

Figure 5. Analysis of the passage from “Татьянаның Онегинге жазған хаты”.

5. Conclusion
This paper outlines the main approaches to the automation of the process of statistical analysis of the lower structural levels (meter, rhythm, phonetics, vocabulary, grammar) of poetic texts, as well as the algorithms for complex analysis of poetic texts in the Kazakh language are presented in order to automate the process of creation of metric dictionaries and concordances. The results of this analysis will significantly expand the possibilities of the philologists who study both these levels of poetry and their semantic and pragmatic characteristics, and also will rid the philologists of routine work, will expand the range of analyzed works, by reducing the dependence of the quality of comparative analysis on the personal erudition of the researcher.

Acknowledgments
The work was partially supported by RFBR grant No. 18-07-01457, and by the topics of the state task No. AAA-A17-117120670141-7 (No. 0316-2018-0009).

References
[1] Magomedova D M 2004 Philological analysis of lyrical poems (Moscow: Publishing center “Academy”)
[2] Taranovsky K 2000 About the relationship between poetic rhythm and subject *About poetry and poetics* (Moscow: Languages of Russian culture) 372–403

[3] Alimkulov T 1971 Herald of truth. In honor of the 125th anniversary of Abai *Kazakhstanskaya Pravda* May 22

[4] Auezov M O 1960 About traditional and innovative in the Kazakh Soviet literature *Rep. the XXV Int. Cong. of Orientalists*

[5] Akhmetov Z A 1964 Kazakh versification: the problems of the development of the verse in pre-revolutionary and modern poetry (Alma-Ata: Nauka)

[6] Hayward M 1996 Analysis of a corpus of poetry by a connectionist model of poetic meter *Poetics* 24 (1) 1–11

[7] Kaplan D M and Blei D M 2007 A computational approach to style in american poetry *7th IEEE Int. Conf. on Data Mining* 553–58

[8] Kao, J and Jurafsky D 2012 A computational analysis of style, affect, and imagery in contemporary poetry *NAACL Workshop on Computational Linguistics for Literature*

[9] Delmonte R 2013 Computing poetry style *Proc. of 1st Int. Workshop ESSEM* (CEUR Workshop Proc. no 1096), eds C Battaglino, C Bosco, E Cambria, R Damiano, V Patti and P Rosso 148–55

[10] Bobenhausen K and Hammerich B 2015 Literary metric, linguistic metric and German algorithm metric settings *Languages* 199 67–87

[11] Barakhnin V, Kozhemyakina O and Zabaykin A 2016 Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts *2016 Int. Conf. “Education Environment for the Information Age” (SHS Web of Conf. vol 29)*

[12] Barakhnin V, Kozhemyakina O and Pastushkov I 2017 Automated determination of the type of genre and stylistic coloring of Russian texts *2017 Seminar on Systems Analysis (ITM Web of Conf. vol. 10)*

[13] Kurt A and Kara M 2012. An algorithm for the detection and analysis of arud meter in Diwan poetry *Turkish J. of Electrical Engineering & Computer Sciences* 20 (6) 948–63

[14] Oralbaeva N, Madina F and Abulkayev A 1993 *Kazakh language* (Almaty: Kazakh University)

[15] Kozhemyakina 2002 (Astana: Astana Polygraphy)

[16] Kozmin A V 2006 Automatic verse analysis in Starling system *Proc. Int. Conf. «Dialogue 2006»* (Moscow: Publishing center of the RSUH) 265–68

[17] Bakieva A M 2017 *The program of generation of word forms of the Kazakh language*

[18] Porter M F 1980 An algorithm for suffix stripping *Program* 1 (3) 130–37

[19] Barakhnin V B, Fedotov A M, Bakieva A M, Bakiyev M N, Tazhibayeva S Zh, Batura T V, Kozhemyakina O Yu, Tussupov D A, Sambetbaiyeva M A and Lukpanova L Kh 2017 The software system for the study the morphology of the Kazakh language *The Eur. Proc. of Social & Behavioural Sci. XXXIII* 18–27