Journal of Siberian Federal University. Humanities & Social Sciences
2023 16(8): 1256–1270

EDN: IVPVAN
УДК 81’342

# The National Corpus of Kazakh Language: Development of Phonetic and Prosodic Markers

**Zeinep M. Bazarbayeva[a], Sholpan K. Zharkynbekova*[b], Aisaule Zh. Amanbayeva[a], Zhanar T. Zhumabayeva[a] and Ainur A. Karshygayeva[a]**
*[a]The Institute of Linguistics named after A. Baitursynuly
Almaty, Republic of Kazakhstan*
*[b]L.N. Gumilyov Eurasian National University
Astana, Republic of Kazakhstan*

Received 22.03.2023, received in revised form 29.05.2023, accepted 05.06.2023

**Abstract**. The article focuses on the issue of corpus linguistics in Kazakh language studies. Nowadays, the sphere of corpus linguistics is being researched, and the base for the National corpus of the Kazakh language is in the process of preparation. The article discusses the ways of entering phonetic and prosodic markers to the oral text subcorpus of the Kazakh language. The analysis of vowels, consonants, and the three types of syllables is done. The system of linguistic knowledge is taken into consideration for entering the texts to the corpus base automatically. In particular, the article describes the rules of putting phonetic and prosodic markers, the studies of word melody, the analysis of automatically dividing a word into syllables, and distinctive features of phonemes. All the sounds of the Kazakh language are described in the article, and the differences and definitions of phonemes and phonetics are given. Also, to demonstrate the intonational features of words, the markers signifying tone, loudness, pauses, and intensity are given. The models of intonemes of sentences are created. For instance, the following model is described, and the ways of its automatization are given: in the beginning of the sentence, there is rising tone, and the sign of incomplete intoneme (↑) is given; at the end of the sentence the tone falls, and the sign of complete intoneme (↓) is given, while the tone stabilizes in the middle of the sentence (→). The results of the study will be useful for preparing the prosodic marking of the oral subcorpus, for the research in the fields of phonetics and phonology, and for writing Master’s and Doctor’s scientific papers.

**Keywords:** corpus, prosody, markup, phonetics, phonology, vocalism, consonantism, syllable, letter.

Research area: corpus linguistics and cultural studies.

* Corresponding author E-mail address: zharkyn.sh.k@gmail.com

The article is published in the scope of program-oriented funding titled “Development of the national corpus of the Kazakh language as an informational-innovative base of the state language: scientific-research and study internet source” (BR 11765619).

Citation: Bazarbayeva Z. M., Zharkynbekova S. K., Amanbayeva A. Zh., Zhumabayeva Z. T., Karshygayeva A. A. The national corpus of Kazakh language: development of phonetic and prosodic markers. In: *J. Sib. Fed. Univ. Humanit. soc. sci*., 2023, 16(8), 1256–1270. EDN: IVPVAN

# Национальный корпус казахского языка: фонетические особенности и просодические маркеры

**З. М. Базарбаева[а], Ш. К. Жаркынбекова[б], А. Ж. Аманбаева[а], Ж. Т. Жумабаева[а], А. А. Каршыгаева[а]**
[а]*Институт языкознания им. А. Байтурсынова*
*Республика Казахстан, Алматы*
[б]*Евразийский национальный университет им. Л. Н. Гумилева*
*Республика Казахстан, Астана*

**Аннотация.** В статье изучается проблема корпусной лингвистики в казахском языкознании. В настоящее время проводятся исследования, связанные с корпусной лингвистикой и подготавливается база национального корпуса казахского языка. В статье рассматриваются пути расстановки фонетической и просодической разметки в устном текстовом подкорпусе казахского языка, анализируются гласные и согласные звуки, а также открытые, полузакрытые и закрытые слоги. Для автоматического введения текстов в базу корпуса взята за основу система лингвистических знаний, в том числе описаны правила расстановки фонетической и просодической разметки, исследования, проведенные с целью определения мелодики слова, анализ автоматического деления слова на слоги, а также дифференциальные признаки каждой фонемы. В статье описываются все звуки, характерные для казахского языка, даются описания и различия фонем и фонетики. Также с целью демонстрации интонационных особенностей отдельных слов даны условные знаки, обозначающие тон, громкость, паузы и темп. Созданы модели интонем предложений. К примеру, в начале предложения присутствует повышение тона (интонема), то есть ставится знак интонемы незавершенности, а в конце предложения наблюдается понижение тона, и ставится знак интонемы завершенности (↓), тогда как в середине предложения тон стабилен (→). Приведенный пример описан в виде модели, даны пути ее автоматизации. Результаты исследования будут полезны в создании просодической разметки базы устного подкорпуса, в проведении фонетико-фонологических исследований, а также в написании научных работ магистрантов и докторантов.

**Ключевые слова:** корпус, просодика, разметка, фонетика, фонология, вокализм, консонантизм, слог, буква.

Статья опубликована в рамках программно-целевого финансирования «Разработка Национального корпуса казахского языка как информационно-инновационной базы государственного языка: научно-исследовательский и обучающий интернет-ресурс» (BR 11765619).

Научная специальность: корпусная лингвистика и лингвокультурология.

Цитирование: Базарбаева З.М., Жаркынбекова Ш.К., Аманбаева А.Ж., Жумабаева Ж.Т., Каршыгаева А.А. Национальный корпус казахского языка: фонетические особенности и просодические маркеры. *Журн. Сиб. федер. ун-та. Гуманитарные науки*, 2023, 16(8), 1256–1270. EDN: IVPVAN

**Introduction**

Corpus linguistics refers to the concept of language informatization through computer technology. In today's world languages, the problem of automating the language by means of a computer program is developing on a large scale along with the development of technology. There are scientists who have studied the "corpus" issue in general. They are: McEnery T., Wilson, (1); Svartvik J., (2); Meyer Ch. F. (3), K.P. Chilingarian (4), A. Zhubanov (5) etc.

Considering the studies related to corpus linguistics, the term "corpus" is defined as follows in the work of V.P. Zakharov "Corpus linguistics": «The term linguistic, or linguistic, corpus of texts is understood as a large, electronically presented, unified, structured, marked up, philologically competent array of linguistic data, designed to solve specific linguistic problems» (Zhakharov, 2005). Similar definitions can be found in the works of E. Finegan and V.V. Rykov.

In general, "Corpus Linguistics" itself consists of several sub-corpora. Oral subcorpus can be mentioned as one of them. In modern linguistics, scientists have dealt with the problem of oral subcorpus and determined its role in linguistics. However, if you look at any oral subcorpus, the level of creation and execution is different. For example, the national corpus of Russian linguistics includes oral, accentological and multimedia corpus types. The entries in the corpus database of the Russian language contain oral speech of young, middle-aged and elderly people. In the oral and accentological corpora, since the accent changes the meaning of a word in Russian, the accent is placed on the word and the words are written in the form of orthography. There are 13 million words in the oral corpus of the Russian language, 133 million in the accentological subcorpus, and 5.7 million in the multimedia subcorpus. And 10 percent of the British corpus base is based on the oral subcorpus. The oral subcorpus consists of two parts. The first demographic section contains transcriptions of spoken language texts of people from different social groups, while the second section contains transcriptions of speech from business people or government officials via radio shows and telephone conversations.

The general 10-million oral subcorpus is divided into two equal parts: 1) spontaneous, natural dialogues with their transcription which are included in the demographical part; 2) contextual-manageable part which carries out an important role. In this part, public speeches from significant meetings and events are presented.

And the database of the oral corpus of the Czech national corpus consists of 4 million words. Corpus of spoken German texts. The creation of corpora of spoken language texts began early in Germany. The main form of preservation of spoken language corpora was recording on magnetic tape. However, later the acoustic form was replaced by phonetic transcription. The full version of phonetic transcription has not been successful due to the difficulty of transcription and user acceptance. As a result of several experiments on the introduction of the spoken language into a graphic form, a partial version of the phonetic transcription was created, which preserved some features of speaking and represented the sequence of speech acts, i.e., their simultaneous occurrence. In the abbreviated version of the phonetic transcription, intonation, regional features of pronunciation, degree of clarity of speech, etc. imaging tools were used. That is, if we look at the study of the oral subcorpus of the languages of the world and the current appearance, it was found during the research that some of them provide orthographic transcription of audio recordings, and some provide phonetic transcription.

**Theoretical framework**

Creating an electronic corpus and developing its usage is one of the most advanced directions in contemporary linguistics. By a corpus we mean the data collected in a computer base with the help of a specific program. However, the corpus base not only collects and stores texts and data, but also upgrades the linguistic knowledge. There are certain methods which help to deal with corpora of texts.

The methods of corpus linguistic may be divided into following blocks: 1) philological methods; 2) theoretical-linguistic methods; 3) mathematical (statistic) methods; 4) informational technological methods. The first group method includes genre, text authors, contribution of genre texts, etc. methods related to problems. In the second group – methods related to general and individual linguistic problems of text selection and provision of linguistic information in the corpus. The third group consists of methods of mathematical statistics related to the number of sample texts, aspects of ensuring its representativeness (volume). The fourth group includes information methods that provide a computer representation of corpus data and its processing operations.

There are different types of corpus linguistics: written and spoken (audial) texts, parallel (texts of several languages), stylistic (publicistic, belletristic, official, scientific, colloquial), chronological (synchronic and diachronic) etc. Some languages include not only written texts, but also texts in oral form. The written texts and oral texts included in the corpus allow to define the distinguishing features between these two forms of speech. Written texts are written in accordance with traditional principles and linguistic norms, while oral texts show the dynamic nature of the language. That is, oral texts are a direct manifestation of natural living language. In this regard, texts in oral form are also included in the corpus database. Oral texts include official texts such as radio broadcasts, interviews, round table materials, and conversations with public figures. Creating prosodic markings for text is mainly a problem of oral corpus. When searching for the required word, the audio/video of each example from the oral corpus concordance is next to it. Because prosodic signs mainly reflect the phonetic process in oral speech. For example, syntax, pause, stress, intonation. That is, prosodic markings are placed on oral texts. In prosodic notation, the text is divided into syntagms, phrases, pauses are determined, and orthography is created. Since the problem of creating an automated system for prosodic level notation is very complicated at the moment, the main work of notation-coding is done "by hand". When asked why the oral corpus might be needed, the scientists provide the following opinion: «Obviously sounding corpus will be a good help for researchers and teachers of phonetics and orthoepy. Moreover, since the corpus is expected to be fairly balanced in terms of chronology, and, in addition, the years of birth will be indicated for a significant number of performers, it will be possible to set a kind of historical task on the corpus – to consider certain phonetic phenomena in the history from the 1930s to today» (Grishina &Savchuk, 2008).

Each nation has its own language intonation. Intonation is unique by its inner characteristics. Such features include the following prosodic tools: *melody, intensity, longitude, pause, intensity, rhythm, tone.* Prosodic methods transform, articulate and beautify the speech according to the tone and change of the voice, give light to the speech of the speaker and develop the expressed thought. Through this, the meaning of the sentences in the text becomes clear. Prosodic methods break down the sentences in the text, reveal the relationship between the individual segments, determine and explain the thought expressed. The most important and universal prosodic method of intonation in all languages is melody. The reason is that the unique melody of that language plays a big role in expressing its difference from other languages. In general, the function of the melody in the sentence is pleasant. Firstly, the melody and delay are what divides the flow of the spoken language into parts and joins it. Secondly, it plays a big role in determining the communicative basis of the sentence, that is, in the division of the sentence into the purpose of its pronunciation. Thirdly, together with the grammatical and lexical methods, it participates in defining the general meaning of the sentence and determining its features. First of all, it has a great value in conveying the logical and emotional content of the sentence, combined with the syntagmatic accentuation of the sequence of words. The melody of the voice can change depending on the individual's speech style and according to the features of speech, and be characterized by stylistic coloring. However, the main form of the melody is preserved according to the norm, because without it, the main content of the sentence may change, become unclear, and the idea might be altered.

In developing the prosodic specification of the oral corpus, the text was divided into syntagms and rhythmic groups. At the same time, among the prosodic methods, the outline of the melody (high, low, high-low, low-high, steady), types of delay (emotional, hesitating), expression of stress in discourse (thought stress, emotional stress) were guided. The place of delay is special in placing prosodic markings in the text. This is because when a person speaks, he does not say all the words one after the other, sometimes he pauses a little in terms of time, sometimes he pauses for a long time and continues his speech, that is, he makes a delay. The main task of the voice delay is to separate the text into parts and to determine the relationship between those parts, to explain the meaning, to read and understand the written text. The linguistic function of a pause is not only its division of a whole text into phrases, syntagms, rhythmic groups, words, but also, combined with other components of intonation, its participation in determining relations between the sentences which have different content (Bazarbayeva, 2015).

At the same time, one of the prosodic methods of intensity can be identified in the oral corpus. Together with other components, it has some content. Intensity is often combined with the main tone frequency of sounds and participates in the accentuation of words in a sentence. If you need to accent one segment of the word chain in the sentence, it will strengthen the intensity there. But the increase in speed often sounds louder to the ear. The loudness of the voice is related not only to the intensification, but also to the rise of the main tone of the voice. Of the two segments of the sentence with the same intensity, the one that sounds louder to the ear is the higher of its main tone. The absolute value of vowel intensity is also affected by their articulation features. It has been proven experimentally that open vowels are more intense than short vowels.

All the mentioned prosodic markings were implemented by hand. The reason is that if we attempt to put these markings in the computer system, we need to alter the audiotapes. For instance, in the Russian national corpus, the prosodic marking is also done manually. The authors explain it as follows: «Prosodic markings describe stress and intonation. This markup is accompanied by discourse markup, which serves to mark repetitions, reservations, and so on. Text markup is done using software, which reduces labor costs. For anaphoric and prosodic markups, creating such software tools is a complex task, so most of the work is done manually. The applied software requires manual post-editing (morphological homonymy and syntactic ambiguity), since the programs represent a number of variations of the solution, and the researcher himself chooses the right one. A full automated markup process is in the future» (Amiyeva et al., 2016).

In Russian, the basis of prosodic marking is the division into syntagmas and stress marks. Taking into account the phonetic system of the Kazakh language, the prosodic marking of Kazakh texts provides for the syntagmatic division of the transcribed sounding speech with the indication of the direction of movement of the tone, the marking of syntagmas into accent units (rhythmic groups) and pauses of hesitation. To be more exact, «of the three possible forms of the arrangement of oral speech on the website of the National Corpus (audio recording, transcription, standard spelling), the creators of the corpus, for reasons of both technical and ideological nature, chose the orthographic principle for presenting oral speech: the text is written in traditional Russian spelling. In this case, punctuation marks inside the sentence are generally removed (replaced by slashes), and only characters that are functionally equal to a period remain in the text. At the same time, of course, it should be remembered that slashes in transcripts marked up in this way do not have any semantic load. In this way, slashes in the corpus differ from the similar way of notation widely accepted in various phonetic traditions, where a single slash denotes a small pause, and a double slash denotes a large pause or its equivalents» (Contemporary Russian language. Theory. Analysis of linguistic units, 2002).

The prosodic markings for the texts developed for the oral internal corpus of the Kazakh language have their own characteristics. One of them is to remove the punctuation marks from the texts and replace them with short introduction (/), syntagma (//) and phrase (///) symbols. With this, the user of the corpus immediately understands the content of the spoken text. Since the prosodic notation is mainly done by hand and not by a special program, the range of texts in it is less than that of the written corpus.

Although the prosodic marking is being newly created, the phonetic-phonological marking is fully functional in the national corpus of the Kazakh language. In phonetic-phonological marking, syllables and sounds are given phonetic-phonological description. It can be said that syllable theory in the field of phonetics is one of the most complex and important issues. In this regard, the theoretical founder of the Kazakh language, prof. Kudaibergen Zhubanov says: "In order to understand the nature of the Kazakh language correctly, you need to understand the syllable system. If you don't know the syllables well, you don't know the spelling either: you can't understand the basis of morphology, such as word formation and word transformation, unless you know the syllables, you won't understand the secrets of language phenomena related to the sentence system, such as word stress (accent), word tone (intonation), You can't take classes like "impressive reading" and "melodic speech" properly. If you are not familiar with the place of syllables, you will not be able to properly familiarize yourself with the basis of the structure of the poem…" (Zhubanov, 1966).

A. Baitursynov overviews such terms as speaking and sentence, sentence and word, word and syllable, syllable and sound as interrelated and interconnected phenomena. While describing the ways of dividing words into syllables, he notes: "If the syllable-forming letter is surrounded by other letters, such a syllable should be called fully closed; if the end of the syllable is the syllable-forming letter, such a syllable should be called open; if the end of the syllable is a letter other than the syllable-forming one, such a syllable should be called closed", thus enumerating the mentioned types. Thus, the syllable-forming letters are vowels. The scholar also notes that "there are no syllables without the mentioned letters" (Baitursynov, 1992). So, the "syllable forming" sounds are vowels. The scholar notes that it is impossible to create a syllable without vowels.

According to the scientist Zh. Aralbayev, "one syllable in a word is pronounced in a more intense manner than other syllables, and this type of accent is called lexical stress". Stress and vowel harmony define the word limits in the Kazakh language, distinguish between words, create phonetic words. These features can be observed through examples. For instance: 1) *Су ал маған* (Bring me water) 2) *Су алмаған* (without water) 3) *Суалмаған* (unwatered) (cow) (Aralbayev,1988). Considering these examples, it can be seen that the intonation of the three examples is pronounced in three different ways depending on the stress, and even the meaning has changed. Oral speech is very rich in intonation variability. It gives our language a different melodic character. The concept of intonation includes the tempo of speech, the tone of speech, voice variability, and emphasis. A person's speech is closely related to many psychological processes (perception, will, feelings). That's why it is said that emotional coloring prevails in the spoken word rather than in the written word (Manasbayev et al., 1974).

Based on the studies of syllables, the model of its automatic recognition was created.

**Problem statement**

In order to introduce the prosodic marking into the corpus base, it is better to find and study the mechanisms through scientific research. In the prosodic notation of the corpus of the Russian language, only the sentences are divided into syntagms and accented. And in the prosodic notation of the Kazakh language, the orthography of each sentence is created, divided into syntagms, and intonemes (rising, falling, flat) are given. Rhythmic groups and syntagms reflected in speech (text) are the basis for distinguishing meaning from the phonetic and semantic-intonational point of view (Bazarbayeva, 2012). That is, each sentence is internally divided into syntagms, and the semantic-intonational features of the speech (text) are shown. Syntagma is a universal, spoken, semantic unit that appears during text segmentation, characteristic of all languages. When dividing the text into syntagms, it is important to correctly set the boundaries of the syntagm so that the meaning of the sentence is not violated. Demarcation of syntagms affects not only the transfer of semantic content, but also the transfer of intonation characteristics. For example: *Балалар жазда | демалысқа шығады (Children go on vacation during summer).* – this sentence consists of 2 syntagms, and while the first syntagma is a theme that shows the known information, the second is a rheme which shows the new information. These two help connect the sentence and make it meaningful. The first syntagma is characterized by rising intonation which depicts the unfinished character of the content, while the rheme has falling intonation which concludes the idea.

We rely on scientific data when putting the prosodic marking of the oral subcorpus, that is, when determining the intonation of each sentence. Z.M. Bazarbayeva says that intonation has five main functions in Kazakh linguistics. They are: determining the general communicative types of the sentence (differentiates the informative, interrogative, imperative types of the sentence); distinguishing individual communicative types (divides the sentence into topic and rheme); distinguishing differences of the content (according to the content, enumeration, opposition, combination, continuation, superposition, sentence intonation); separating sentences, syntagms, texts from each other (correctly expresses types of sentences, distinguish between their categories, pronounce words clearly); definition of emotion (not only the shade, but the change in meaning). The scientist says that a sentence *I know you very well* can have a variety of meanings with the help of intonation change (mocking, respect etc.) (Bazarbayeva, 2016). We take Z.M. Bazarbaeva's research as a basis when developing the national corpus of the Kazakh language. One of the intonation components in the prosodic notation was guided by the melody. Also, accent plays a special role in oral speech. Accent function is different in different languages. Among them, the mobility of the accent in the Russian language (at the beginning and at the end) makes it one of the inflected languages, and the fact that the accent in the Kazakh language often falls on the last syllable of the word indicates that the Kazakh language belongs to the group of agglutinative languages. That is, as I. Kenesbayev says, the connection between the accent and the typological differences of these languages (Kazakh being agglutinative and Russian being inflected) is supported by various scientific research (Kenesbayev, 1945). In this regard, A. Zhunisbek concludes: "As we have noticed, the words in the Kazakh language do not always have independent stress within a sentence" (Zhunisbek, 2018).

**Methods**

Prosodic marking of spoken text is done manually. However, when the speaker's speech is made according to orthography, it is semi-automated. The automaton creates the orthography of some sounds in the spoken text according to the algorithm. The algorithm was created according to the orthographic norm of the Kazakh language. The orthography of the Kazakh language is based on the law of harmony. In this regard, the laws of harmony were used as guidance. In the same way, its algorithm was created for giving the phonetic-phonological character of a syllable and each sound. The Praat program was implemented in giving the acoustic-articulatory character of some sounds.

The oral corpus consists of several methods. After the prosodic notation is applied to the spoken text, it is entered into the corpus database. The researcher can listen to the speech of the speaker and see how he speaks through the text. Then you can see the standard version of the spoken text. The transmission of prosodic marking by such a method indicates its distinctiveness from other corpora. The reliability of the corpus research method largely depends on the corpus used and its technical features. In turn, the degree of representativeness of the corpus data affects the possibility of coming to general conclusions, not only related to the corpus, but also to the language as a whole.

**Results and discussion**

At the initiative of professor A.K. Zhubanov, a corpus base was created and a national corpus was built in Kazakh language studies. The national corpus of the Kazakh language has been working since 2009. In the implementation of the corpus base, analyzes were first carried out in the field of morphology. The morphological subcorpus is the main one. Based on world experience, the scope of the national corpus of the Kazakh language is expanding. In connection with this, the oral subcorpus base was also launched. The works of Zh. Aralbaev, A. Zhunisbek, N. Vali and other scientists were taken as a basis when developing the oral subcorpus. In this regard, we base the development of the oral subcorpus on 3-step guide provided by A. Zhunisbek.

In the 3-step guide by A. Zhunisbek, the texts are described by three different characteristics. They are the melody of the word, automatic syllabification, and phonetic descriptions of sounds (letters). According to these characteristics, the following guide was prepared for the program.

The first step: Showing the melody of the word (or timbre). In the Kazakh language, the words are divided into two types of melody according to the thickness or thinness of the vowel. The program incorporated in the corpus base should automatically distinguish between "thick" and "thin" words. When affixes are being added to the roots of Kazakh words, they are attached according to the thickness or thinness of the vowel in the word. So, the word becomes either thick or thin. In order for the program to differentiate these types of words, the following rule must be entered there. If the vowels in the first syllable of the word (or in the word itself) are thick (this includes *а, ы, ұ, о*), then such a word is considered thick from the viewpoint of vowel harmony. If the vowels in the first syllable of the word (or in the word itself) are thin, then such a word is considered thin from the viewpoint of vowel harmony.

The second step: Dividing the words into syllables and giving descriptions of them. The syllables in the Kazakh language are differentiated according to the number of vowels in a word. The number of vowels in a word equals to the number of syllables. For example:

Әліппе (Primer). There are 3 vowels here, hence 3 syllables: Ә-ліп-пе

Ана (Mother). There are 2 vowels here, hence 2 syllables: А-на

Бала (Child). There are 2 vowels here, hence 2 syllables: Ба-ла

Жаңбыр (Rain). There are 2 vowels here, hence 2 syllables: Жаң-быр

Құттықтады (Congratulated). There are 4 vowels here, hence 4 syllables: Құт-тық-та-ды

To describe the structure of a vowel, we take the globally recognized signs for vowels and consonant, which are V for a vowel and C for a consonant. In order to implement phonetic marking, the program needs to assign the mentioned V and C signs to the vowels and consonants. For example: А- V; Ә – V; Ы – V and so on. Also, the consonants: П – С; Б – С; М–Х and so on.

K. Zhubanov was the first to present his opinion on the issue of distinguishing syllables. In his article "How can we distinguish between syllables?", the scholar shows the following shortened markings of vowels and consonants: Ды – vowel; Дз – consonant; сДз – prolonged consonant; кДЗ – voiceless consonant.

Professor K. Zhubanov presented six types of syllables according to their sound content.

1) Ды (V) syllable, or a fully open syllable (is made of a single vowel).

2) Дз-Ды (CV) syllable, or an open syllable.

3) Ды-Дз (VC) syllable, or a light closed syllable.

4) Дз-Ды-Дз (CVC) syllable, or a light fully closed syllable.

5) Ды-кДз-сДз (VCC) syllable, or a heavy closed syllable.

6) Дз-Ды-кДз-сДз (CVCC) syllable, or a heavy fully closed syllable.

The first four types of syllables, as K. Zhubanov believes, are the main types that can be met in the Kazakh language.

K. Zhubanov notes that “the syllables can be distinguished easily, mechanically” through the mentioned models (Aralbayev, 1988).

Taking into consideration the conceptions of the Kazakh linguists related to the types and structure of syllables, we enter the following descriptions to the program that implements phonetic marking:

Vowel syllable – V (а-на, о-тан, а-та, ә-ліп-пе, е-ді, ө-мір, ы-дыс, і-ні т.б.);

Open syllable – C+V (ба-ла, қа-ла, бөл-ме, ә-ке, са-бақ т.б.);

Closed syllable – V+C (ақ-ша, ақ-та-ды, ар-нау, өр-ле-ді, ор-на-лас-қан);

Closed syllable – V+C+C (ант, айт-ты);

Fully closed syllable – C+V+C (бай-лық, қой-ды, бер-ді, ә-кел-ді, та-быс, ең-бек, жаз-ды т.б.);

Fully closed syllable – C+V+C+C (бұлт, қант, жалт, жырт-ты, құрт-ты т.б.);

Fully closed syllable – C+C+V +C (Өз-бек-стан, Қа-зақ-стан т.б.).

In order for the corpus program to divide words automatically, it is necessary to give instructions about the positions of the given syllables. For example, V the vowel syllable is only met in the beginning of the word (а-на, о-тан, а-та); C+V the open syllable can be met in any position within a word (са-бақ, бер-ме-ді, бөл-ме); V+C the closed syllable is only met in the beginning of the word (ақ-ша, ақ-та-ды, ар-нау); V+C+C the closed syllable is only met in the beginning of the word (ант, айт-ты); C+V+C the fully closed syllable can be met anywhere in a word (қой-ды, бер-ді, ә-кел-ді, та-быс), and the same is true about the fully closed syllable C+V+C+C (бұлт-ты, жаң-ғырт-ты, жаз-дырт); and the fully closed syllable C+C+V+C can be met anywhere but the beginning of the word (Өз-бек-стан, Қа-зақ-стан-да).

The corpus program divides words automatically following the given algorithm. However, it cannot divide some words in the Kazakh language that have the phonemes *и* and *у*. For example: *баруы (going), келуі (coming), балуан (wrestler), алуан (various).* In these words, the vowels preceding the given phonemes are omitted. If these words were written as *барұуы, келүуі, балыуан, алыуан,* the program would be able to divide them as the instructions prescribe.

The next phonetic description in phonetic marking of the corpus is the analysis of sounds. For this purpose, a list of sound characteristics involving the role of tongue, jaw, and lips is prepared for the program. For example: А – open, back, unrounded; Ә – open, central, unrounded; Ы – closed, back, unrounded; etc. We are taking these descriptions from the work of A. Zhunisbek.

A. Zhunisbek gives the following descriptions to the phonetic analysis of the word structure in the corpus:

ӘЛІППЕ – thin (soft), ә-ліп-пе: ә – vowel syllable, ліп – fully closed syllable, пе – open syllable; ә – open, central, unrounded; л – alveolar, approximant, sonorant; і – closed, central, unrounded; п – bilabial, plosive, voiceless; е – diphthong, central, unrounded.

At the bottom of the window, words are described by the three-step system of A. Zhunisbek. The first step: the melody (timbre) of a word is shown depending on whether it is pronounced hardly or softly. The second step: the word is automatically divided into syllables, and the types of syllables are described. The third step: each sound in the word is analyzed from the phonetic viewpoint. This phonetic description included in the oral subcorpus ensures that a program that can automatically segment future syllables will work on the Internet. On the other hand, it will help to implement the functions of the text editor related to checking the transfer of words, as well as to solve many applied problems of linguistics in the future. In addition, each phoneme was described. Clicking on any word in a cell will give you a phonological description for each phoneme in that word. Different sounds are pronounced in the flow of speech, there are many types of their pronunciation, there is no limit, but they combine to form one type of sound, which we call a phoneme.

Z.M. Bazarbayeva gives the following definition regarding the phoneme: “The smallest functional unit of the language, the type of sound that separates the meaning of the word from the body of the word and is included in it.” Phonemes are paired and contrasted with each other in order to define their meaning, and then their main properties and features are determined. The number of phonemes in one word can be small or large, for one word to be different from another, it is not necessary for all phonemes to be different. One word differs from other words in terms of number, quality and order of phonemes. For instance, the phoneme <A>. It can be met in any position within a word. It can be paired with any vowel. When it comes after the phonemes **[Ш]** and **[й], [ж]** and **[й],** and within a soft syllable with the phoneme **[й],** its sound representation becomes **[ә],** which means the phoneme is palatalized. Now we list the words which have the **[ә]** variant of **<а>** phoneme. *Ж<а>й (simple) – ж[ә]й, ш<а>й (tea) – ш[ә]й, м<а>йсөк (fat) – м[ә]йсөк, бид<а>й (wheat) – бид[ә]й, ауж<а>й (crotch) – [ә]уж[ә]й, д<а>укес (arguer) – д[ә]укес, д<а>укестік (argument) – д[ә]укестік, ж<а>йбарақат (carelessly) – ж[ә]йбарақат, ж<а>йбарақаттану (being careless) – ж[ә]йбарақаттану, ж<а>йбасар (slow) – ж[ә]йбасар, ж<а>йғастыру (to settle) – ж[ә]йғастырұу, ж<а>йғасу (to settle) – ж[ә]йғасұу, ж<а>йғату – ж[ә]йғатұу, ж<а>йғызу (to make spread) – ж[ә]йғызұу, ж<а>йдақ (without a saddle) – ж[ә]йдақ, ж<а>йдарлы (jubilant) –ж[ә]йдарлы, ж<а>йдарман (smiley) – ж[ә]йдарман, ж<а>йдары (happy) – ж[ә]йдары, ж<а>й-күй (health and wealth) – ж[ә]й-гүй, ж<а>йқалу (to sway) – ж[ә]йқалұу, ж<а>йлы (comfortable) – ж[ә]йлы, ж<а>йма (dough) – ж[ә]йма, ж<а>йма-шуақ (bright sunlight) – ж[ә]йма-шұуақ, ж<а>йсаң (a comfortable place) – ж[ә]йсаң, ж<а>йша (just) – ж[ә]йша, ж<а>йшылық (nothing special) – ж[ә]йшылық, ж<а>йсыз (inconvenient) – ж[ә]йсыз, ж<а>йнау (to glitter) – ж[ә]йнау, ж<а>йт (situation) – ж[ә]йт.* Also, after the consonant **[к]** the sound representation of the phoneme **<а>** is **[ә]**, which refers to palatalisation: *қ<а>дір – қ[ә]дір, қ<а>дірлеп-қастерлеу – қ[ә]дірлеп-қ[ә]стерлеу, қ<а>дірлеу –қ[ә]дірлеу, қ<а>дірлі – қ[ә]дірлі, қ<а>жет – қ[ә]жет, қ<а>жетсіз – қ[ә]жетсіз, қ<а>стерлеу – қ[ә]стерлеу, қ<а>стерлі – қ[ә]стерлі.* … **[к]** is usually a voiceless hard consonant, and when it softens, we may call this sound harmony. Also, in the word *қауесет (gossip)* the phoneme **[ә]** is pronounced softly under the influence of **[е]** in the last syllable. However, the sound that we hear is between **[а]** and the regular **[ә].** (Orthographic dictionary). When the phoneme **<а>** comes between the consonants **[н]** and **[с],** its sound representation is **[ә]**: *н<а>сихат – н[ә]сійхат, н<а>сихатшы – н[ә]сійхатшы, н<а>сихаттау – н[ә]сійхаттау.*

In the softly pronounced borrowed words, the sound representation of **<а>** when it comes in the II syllable is **[ә]**: *ділд<а> – ділд[ә], тілм<а> ш – тілм[ә]ш, тілм<а> ш – тілм[ә]штық, дінд<а> р – дінд[ә]р* [5].

In the III syllable of softly pronounced borrowed words, the sound representation of **<а>** is **[ә]**: дүб<*а*>р<*а*> – дүб*[ә]*р*[ә]* [7].

The description of phonemes was prepared by the following model.

The nature of the phonemes presented in this sample was clearly shown in the part of the phonetics-phonological designation of the national corpus. Any word can be used, and its phonetic-phonological character is fully revealed.

If we say that one of the tasks of the national corpus is to show prosodic notation, what tasks does it include? In this regard, the first algorithm of vowels and consonants was created to automate the orthoepy of the text. The algorithm was guided by the results of scientific works on orthoepy, which have been studied for several years, and the model of vowels and consonants.

The model is as follows:

**А –** open, back, unrounded; vOt, vZy, vNg

**Ә–** open, central, unrounded; vOt, vSy, vNg

**Ы –** closed, back, unrounded; vZk, vZy, vNg

**І –** closed, central, unrounded; vZk, vSy, vNg

**Ұ –** closed, back, rounded; vZk, vZy, vGb

**Ү –** closed, central, rounded; vZk, vSy, vGb

**П –** bilabial, occlusive, voiceless; cGg, cSm, cGl

Table. Characteristic of phonemes in the Kazakh language

| № | phoneme | Description |
|---|---|---|
| 1 | **<а>** | open, unrounded, back; hard vowel phoneme. |
| 2 | **<ә>** | open, unrounded, front; soft vowel phoneme. |
| 3 | **<ы>** | closed, unrounded, back; hard vowel phoneme. |
| 4 | **<і>** | closed, unrounded, front; soft vowel phoneme. |
| 5 | **<ұ>** | closed, rounded, back; hard vowel phoneme. |
| 6 | **<ү>** | closed, rounded, front; soft vowel phoneme. |
| 7 | **<е>** | partly open, unrounded, front; soft vowel phoneme. |
| 8 | **<о>** | open, rounded, back; hard vowel phoneme. |
| 9 | **<ө>** | open, rounded, front; soft vowel phoneme. |
| 10 | **<п>** | bilabial, plosive, voiceless consonant phoneme. |
| 11 | **<б>** | bilabial, plosive, voiced consonant phoneme. |
| 12 | **<м>** | bilabial, nasal, sonorant consonant phoneme. |
| 13 | **<т>** | alveolar, plosive, voiceless consonant phoneme. |
| 14 | **<д>** | alveolar, plosive; voiced consonant phoneme. |
| 15 | **<н>** | alveolar, nasal, sonorant consonant phoneme. |
| 16 | **<қ>** | uvular, plosive, voiceless consonant phoneme. |
| 17 | **<к>** | velar, plosive, voiceless consonant phoneme. |
| 18 | **<ғ>** | uvular, fricative, voiced consonant phoneme. |
| 19 | **<г>** | velar, plosive, voiced consonant phoneme. |
| 20 | **<ң>** | uvular, nasal, sonorant consonant phoneme. |
| 21 | **<с>** | alveolar, fricative, voiceless consonant phoneme. |
| 22 | **<з>** | alveolar, fricative, voiceless consonant phoneme. |
| 23 | **<р>** | alveolar, trill, sonorant consonant phoneme. |
| 24 | **<ш>** | retroflex, fricative; voiceless consonant phoneme. |
| 25 | **<ж>** | retroflex, fricative, or post alveolar, fricative; voiced consonant phoneme. |
| 26 | **<л>** | alveolar, lateral approximant; sonorant consonant phoneme. |
| 27 | **<й>** | palatal, fricative; sonorant consonant phoneme. |
| 28 | **<у>** | labial; sonorant consonant phoneme. |
| 29 | **<ф>** | labio-dental, fricative; voiceless consonant alien phoneme. |
| 30 | **<ц>** | alveolar, fricative; voiceless consonant alien phoneme. |
| 31 | **<ч>** | alveolar, fricative; voiceless consonant alien phoneme. |
| 32 | **<h>** | pharyngeal, fricative; voiceless consonant phoneme. |
| 33 | **<х>** | velar, fricative; voiceless consonant phoneme. |
| 34 | **<в>** | labio-dental, fricative; voiced consonant alien phoneme. |
| 35 | **/я/** | refers to the diphthongs **[й+а]**; **[й+ә]** |
| 36 | **/ю/** | refers to the triphthongs **[й+ұ+у]**; **[й+ү+у]** |
| 37 | **<э>** | open, unrounded, back; hard vowel alien phoneme. |
| 38 | **/и/** | refers to the diphthongs [ый], [ій] |
| 39 | **/щ/** | In borrowed words: palatal, fricative, voiceless consonant phoneme. In native words refers to double шш sound. |

**Б –** bilabial, occlusive, voiced; cGg, cSm, cZv

**М –** bilabial, occlusive, sonorant; cGg, cSm, cSn

**Т –** alveolar, occlusive, voiceless; cAl, cSm, cGl

**Д –** alveolar, occlusive, voiced; cAl, cSm, cZv

**Н –** alveolar, occlusive, sonorant; cAl, cSm, cSn

Along with the segmental side of phonetics, supersegmental phonetics, that is, prosodic side, is also considered in the research, and descriptive, auditory and formal-descriptive methods are used. The formal-characteristic method is directly related to the graphic concept of intonation. The variety of intonation transcriptions allows to form a speech model as close as possible to the original.

Prosodic methods transform, express and beautify the speech according to the sound and change of the voice, give light to the speech and enliven the expressed thought, the meanings and relationships of the sentences in the text become clear and defined. Prosodic methods break down the sentences in the text, reveal the relationship between the individual segments, determine and explain the thought expressed. In this case, in order to give a prosodic designation to the corpus base, first of all, the orthoepy of each sentence is divided into syntagms from a prosodic point of view, and intonemes of each sentence are added. If possible, the tempo and pauses are also set.

In the scope of research, descriptive, auditory and formal-descriptive methods are used. The formal-descriptive method is closely connected to the graphic representation of intonation. The variability of intonational transcriptions allows to create a model of speaking which is as close to the original as possible. N. D. Svetozarova notes in her work: “The aim of intonational transcription is to convey the main signs of the intonational depiction of a single speech segment in the scope of intonation description theory” (Svetozarova, 1982). That is, it is better to include phonetic and intonation transcription in the transcription of the general text. An extract from A. Nurshaikov’s work “Truth and legend” from the style of fiction is taken into the study, and a phonetic and intonation transcription is made for it. Here, phonetic transcription involves the acoustic-articulatory change of sounds, the assimilation of words (progressive, regressive, interlaced), their transmission in accordance with the law of harmony, and in the intonation transcription, dividing the sentence into syntagms, rhythmic groups, and segment groups, placing 8 intonemes characteristic of the Kazakh language and taking into consideration the voice intensification, pauses and tempo.

>Сонұмен→ |– 1933 → жылдың айағында, |<↑– қақаған⁀ ғыста, <↑ | –шійнелімнің→> | ек’ ⁀етегі делеңдеп, –<↑ | шілемімнің→> | ек’ ⁀ғұлағы→ салпаңдап, –<↑ | Бұурныйға тоқтаған пойыздан→ >| дік етіп↑ ⌊– жерге түстүм. > ↓ || Пойызда геле жатқанда→> –| Тайғаның →| алдаң ⁀ғашқан> →| ақ түлкүсүнің құйрұғұндай↑< | жеб-жеңіл,> →| жұб-жұмсақ сыйақтанып, –<↑ | дала бетінде ––→>| бұлаңдап тұрған боранның↑< | жерге түскөнде→ –| екпіні→ <| ер жігітті алып соғардай< ↑–| қатты екен. ↓|| Маңдайында↑< –| қызыл жұлдұзы жарқыраған →| шошақ шілемнің→< | ек’ ⁀құлағын→ | ійегіме қамзау етіп↑< –| байладым да, >↑ | алдымен→ <| ауұл жаққа ғарадым. >↓ | Ауұл⁀гөрүмбейді. >↓ | Мынау сансыз→ >| – ақ түлкүнің құйрұғындай↑< | бұлаңдаған↓> | ақ түтөктердің< ↑| ар жағында,– → | төрт-бес шақырым→ | жерде ғана тұр. > ↓ || «Жүр-жүрлөп!» <↑– | жүрөкті <→| сағыныш жетелейді. >↓ || «Сабыр, сабыр!» деп↑< –| санам тартпақтайды. >↓ | Жас адамға→< –| жүрөк – бій. > ↓| Сана бійлейтін →<| шар тартқан→ –| кез емес →| еді ғой ол. >↓|| Сақылдаған айаз, →> | соғып тұрған→> | бораңға қарамастан: <↑ «Қайдасың айаулұм, –↑ | ауұлым? <↑ Қайдасың< –↑| ата-ана, → бауұрұм?!» <↑ – деп, <→| Мыңбұлақты бетке алып,– <→| жүрдүм де геттім. >↓

Translation of the passage: “So, at the end of 1933, during cold summer, the hem of my overcoat swaying, two ears of my helmet flapping, I jumped from the train which had a stop at Burnuy. The snow storm that was soft and playful as a white taiga fox while I was on the train – that storm now become so intense that it could kick a grown man down. I fastened the two sides of my helmet, with a red star shining on its front, under my chin and first looked at the village. The village cannot be seen. It is five or six kilometers over these countless white fox tails which sway and flap here. The despair leads me forward, saying "Let's go!" The reason pulls me back, saying "Keep calm!" A young man's heart is his ruler. Those days where not under the governance of reason. Despite the frost and the snow storm, I said "Where are you, my village? Where are you, my parents and relatives" – and went forward facing Mynbulak".

In the study, special marks were placed on the corpus base to convey intonation marks.

The fact that each intoneme is pronounced with a rising tone at the beginning of the sentence, and with a falling tone at the end, was given with special signs. Also, special conventional signs were placed to give the intonation of the text. That is, the transcription of 8 intonemes showing the intonation features of the Kazakh language, i.e. melody, voice intensification, pauses, tempo indicators were showed with special symbols (Bazarbayeva, 2002).

**Signs of intonational transcription:** The vertical line distinguishes syntagmas in speech (|), double vertical line (| | distinguishes phrases in speech.

1. The pause is depicted by a horizontal line (–).
2. The tone line in segments is depicted graphically. The movement of the frequency of the main tone is shown in three main directions:
   - Rising ↑
   - Falling ↓
   - Flat →
3. Defining the tempo of segments.

Depiction of the tempo:

- Slow - - - - - - - - - - - - - - - - - - - -
- Middle __ __ __ __
- Fast ……………….

4. Defining the voice intensification.
   - Increasing <
   - Decreasing >

During the research, the intonation transcription of the text was carried out with the help of auditory analysis. Also, the analysis of melody, voice intensification, tempo and pausation was made. Examples were given in order to signify that every component of intonation has a unique role in conveying the sentence meaning and its idea. For example:

<*Сонұмен* | *1933 жылдың айағында,* | *қақағаң ғыста,* | *шійнелімнің* | *ек' етегі делеңдеп,* | *ш*$^{i}$*лемімнің* | *ек' ғұлағы салпаңдап,* < | *Бұурныйға тоқтаған пойыздан*| *дік етіп* | – *жерге түстүм.* >– this sentence consists of ten syntagms.

**Melody and intensity.** The sentence that started with a flat intoneme in the first and second syntagms then acquires rising or unfinished intoneme in the third and fourth syntagms, rises further in the fifth and sixth syntagms and has a falling intonation at the end. So, the idea is completed. Each intoneme has its own role in conveying the thoughts.

**Tempo.** The beginning is slow, but the third and fourth syntagms are marked with intense tempo along with the placement of logical stress there, and then the tempo slows down at the end. In general, the tempo is necessary for identifying important and unimportant parts of the utterance and solidifying the idea.

**Pauses.** There are five pauses in the sentence. Each pause signifies the completion of an idea and has a role in distinguishing between the separate syntagms. Pauses, on the one hand, are physiological, but on the other hand they show the tidiness of the thought from the logical viewpoint.

<*Маңдайында* –| *қызыл жұлдұзы жарқыраған* | *шошақ шілемнің* | *ек'͡құлағын* | *ійегіме қамзау етіп* –| *байладым да,* > | *алдымен*| *ауұл жаққа ғарадым* > – this sentence consists of eight syntagms.

**Melody and intensity.** The first syntagma starts with a rising intoneme, and the in-between of the second and fourth syntagms is pronounced flatly. The fifts and the sixth syntagms are unfinished, thus they have a rising tone. Then the intoneme is flat again, and there is a fall at the end. Through the intonemes, a complete idea is conveyed here.

**Tempo.** The sentence starts slowly and intensifies in the middle only to slow down at the end. As the experimental research indicates, slow tempo refers to the significant parts of the sentence, while the increased tempo shows that this part of the sentence is unimportant.

**Pauses.** There are seven pauses in this sentence. These pauses, while being physiological, are also utilized to convey the meaning of the sentence and the complete idea.

To summarize, the orthography of words and phrases was given, the sentence was divided into syntagms, and intonemas were placed in the prosodic notation. Prosodic marking cannot be automated immediately. In the oral subcorpus, an algorithm is provided for the automation of the text, but it is carried out manually because it cannot cover the entire oral text. Later, as the scope of the database expands and the algorithm is improved, the text will be fully automated.

## Conclusion

To conclude, the first national corpus of the Kazakh language showed lighter and simpler notations, and included the least amount of text (10 million). In this regard, now the scope of the corpus, breadth of information is increasing, the quality of marking and all parameters and sub-corpus resources are being supplemented. One of them is the prosodic marking contained in the national corpus of the Kazakh language. Prosodic marking mainly consists of oral texts. The audio recordings of classical writers, modern poets and writers, public figures were included. In the oral texts, the speaker's orthography, prosodic features, and intonation decoration were clearly shown. In order for the user of the national corpus to distinguish between the standard orthography of the Kazakh language and the orthography of the speaker, as well as to determine the level of the correct pronunciation of the word, a prosodic marking was given. As for the question of how the user compares the normalized spelling and the speaker's orthography, clicking on a word in the spoken text with prosodic marking will show its normalized spelling, intonation features, and its audio.

The oral subcorpus of the Kazakh language is supplemented and improved with texts over the years. To automate it, the Praat program and the algorithm work together and expand the scope of the database. The algorithm has been developed on the basis of researches in the field of segmental and supersegmental phonetics of Kazakh linguistics and serves as the basis for text automation.

## References

Amiyeva, A.M., Filimonov, V.V., Sergeyev, A.P. (2016). [and others]. 2nd International conference of students, Masters, and young scientists "Informational technologies, telecommunications and systems of management": a collection of reports. Ekaterinburg: [URFU], 9, 251–260.

Aralbayev, Zh. (1988). *Kazakh fonetikasy boiynsha etudter* [Studies on the Kazakh phonetics]. Almaty: Science, 13, 144. (In Kazakh).

Baitursynov, A. (1992). *Til tagylymy* [Study of the language]. Almaty: Ana tili, 12, 448.

Bazarbayeva, Z.M. (2015). Universal Properties of Intonation Components. *Review of European Studies*, 8, 226–230. Available at: https://www.google.com/url (accessed 22 February 2022)

Bazarbayeva, Z.M. (2012). *Kazakh phonologiyasynyn negizderi* [Basics of the Kazakh phonology]. Almaty, 15, 120. (In Kazakh)

Bazarbayeva, Z.M. (2016). *Korpus kazakhskogo yazika: prosodicheskaya razmetka* [Corpus of the Kazakh language: prosodic marking]. In *Mezhdunarodnyj zhurnal prikladnyh i fundamental'nyh issledovanij [International Journal of Applied and Fundamental Studies],* available at: https://applied-research.ru/ru/article/view, 16, 334–337 (In Russ.) (accessed 15 March 2022)

Bazarbayeva, Z.M. (2002). *Kazirgi Kazakh tili intonatsiyasynyn negizderi* [Basics of contemporary Kazakh intonation]. Almaty: Complex, 20, 202. (In Kazakh)

Grishina, E.A., Savchuk, S.O. (2008). *Korpus zvuchashey russkoy rechi v sostave natsionalnogo korpusa russkogo yazika* [The corpus of Russian oral speech in the framework of the national corpus of the Russian language]. Proceedings of the International Conference "Dialogue 2008". 7, 125–136.

Kenesbayev, I.K., Musabayev, G. (1975). *Qazirgi qazaq tili: Leksika, fonetika*, [Modern Kazakh: Lexis, Phonetics], 17, 156. (In Kazakh)

Manasbayev, B., Balakayev, M., Tomanov, M., Zhanpeyisov, E. (1974). *Kazakh tilinin stilistikasy* [Stylistics of the Kazakh language]. Almaty, 14, 190. (In Kazakh)

Meyer, Ch. F. (2002). English Corpus Linguistics: An Introduction. Cambridge: Cambridge University Press, 3, 168.

McEnery, T., Wilson, A. (2001). Corpus Linguistics. Edinburgh: Edinburgh University Press, 1, 235.

Quirk, R. Svartvik, J., Quirk, R. (1980). A corpus of English Conversation. – Lund: Gleerup, 2, 893.

Sovremenniy russkiy yazik. Teoria. Analiz yazikovyh edinitz. [E.I. Dibrova, L.L. Kasatkin, N.A. Nikolina, I.I. Scheboleva]. Edited by E.I. Dibrova. (2006). [Contemporary Russian language. Theory. Analysis of linguistic units]. Moscow: «Academia" publishing house, 480. (In Russ.)

Svetozarova, N.D. (1982). *Intonatsionnaya sistema russkogo yazika* [Intonational system of the Russian language]. Leningrad: the publishing house of Leningrad university, 19, 173. (In Russ.).

Zakharov, V.P. (2005). *Korpusnaya lingvistika: uchebno-metod. posobiye.* [Corpus linguistics: an educational-methodological toolkit]. (In Russ.), Saint Petersburg, 6, 48.

Zhubanov, A.K., Zhanabekova, A.A. (2016). *Korpustyk lingvistika* [Corpus linguistics]. Almaty, «The Kazakh language» publishing house, 5, 336. (In Kazakh)

Zhubanov, K. (1966). *Kazakh tili boiynsha zertteuler* [Studies of the Kazakh language]. Almaty: Science, 11, 362. (In Kazakh)

Zhunisbek, A. (2018). *Kazakh til biliminin maseleleri* [Issues of Kazakh linguistics]. Almaty: Abzalay, 18, 368. (In Kazakh)

Chilingaryan, K.P. (2021). *Korpusnaya lingvistika: teoria vs metodologia* [Corpus linguistics: theory vs methodology]. In *RUDN Journal of Language Studies, Semiotics and Semantics*, 4, 196–218. (In Russ.)