UDC 81’33
SRSTI 16.31.02
DOI10.48371/PHILS.2022.66.3.006

TECHNOLOGY FOR PREPARING A DIALECTOLOGICAL INTERNAL NATIONAL CORPUS OF THE KAZAKH LANGUAGE

Zhanabekova A.A.[1], Kozhakhmetova A.K.[2], *Tlegenova G.B.[3]
[1]Head of the Department of Applied Linguistics at the A. Baitursynov Institute of Linguistics, Doctor of Philology, Almaty, Kazakhstan,
[2]PhD student, Junior Researcher of Department of Applied Linguistics at the A.Baitursynov Institute of Linguistics, Almaty, Kazakhstan,
[3]PhD student, Researcher of Applied Linguistics at the A. Baitursynov Institute of Linguistics, Almaty, Kazakhstan,
[1]e-mail: aiman_miras@mail.ru, [2]e-mail: akony_8484@mail.ru,
[*3]e-mail: gulden_20.88@list.ru
[1]https://orcid.org/0000-0002-6199-7444
[2]https://orcid.org/0000-0003-3783-3199
[3] https://orcid.org/0000-0001-9842-3324

Abstract. The article deals with the world experience of creating a dialectological corpus. The introduction of the dialect corpus into the national corpus of the Russian language on the materials of oral speech, the problems of phonetic transcription and their spelling, the development of prosodic notation and methods of automatic morphological analysis of dialect usage.

The main purpose of the article is to describe the development of the dialectological corpus of the Kazakh language on the basis of the world's experience in creating a dialect corpus and ways to improve it.

Dialectological corpus is necessary for researchers of the language to simplify and speed up research required for scientific articles, monographs and dissertations.

The article uses methods of review, description, narrative, analytical analysis, algorithmic programming in developing the dialectological division of the National Corpus of the Kazakh language, studying applied works in world linguistics in this area.

In conclusion, it should be noted that in the corpus both standardized words of the literary language and words of dialect character should be divided primarily into roots and affixes. In the dialectal corpus, a dictionary of dialect keywords is created in the same way, which is placed in the corpus database. The second stage of morphological analysis, the second part of the word forms, consists in dividing the suffixes into morphemes and labeling them according to their grammatical nature. At the same time, the dialect word forms in the dialect corpus must also be subjected to morphological analysis.

In the future, it is planned to record oral language materials from the regions into a dialect corpus and include them in the corpus. The article gives recommendations for organizing the work in accordance with this goal. This increases the value of the work.

Keywords: dialect, dialect corpus, contextual editor, analyzer, internal corpus, metamarkup, Regional Dictionary, metadata.

This article BR11765619 was published as part of the program-targeted research work «Development of the national corpus of the Kazakh language as an information-innovation base of the state language: a scientific-methodological Internet resource».

Basic provisions

Traditionally, the study of folk dialects was conducted mainly in the linguogeographical aspect. According to this approach, dialectologists first of all studied the structural features of dialectal speech that distinguish one dialect from another: the sound structure of dialects, their grammatical features, vocabulary. However, in the second half of the XX century – the beginning of the XXI century, communicative, cognitive, linguocultural approaches to the study of dialectal speech in dialectology gradually began to strengthen (V.E. Goldin, T.A. Demeshkina, E.L.Berezovich, A.A. Kamalova and L. As a result of the work of A. Savelova, T.V.Makhracheva and others). Linguocultural phenomena, such as oral literature, rituals and systems of customs, preserve and convey long-established worldview and worldview elements in a symbolic generalized and aesthetically processed form, representing a cultural tradition rather than its diachronic section, and then relevant ethnocultural information through modern dialectical speech. elements of other semiotic systems that coexist in a certain society and in the minds of its representatives can also be objectified in the form of actual knowledge. Here the study of folk dialects from the linguistic and cultural point of view is an urgent task of modern dialectology.

Introduction

A dialectological corpus should serve the study of local language features in a linguistic-structural and speech-discursive way. At the same time, dialectological corpora in other countries usually record conversations (audio, video) in the local language and place the audio recordings and their transcripts (text) in the corpus database. Because local differences are observed not only at the level of vocabulary (words), but also at the level of local discourse. While lexical features represent the vocabulary of which local discourse consists, grammatical variation and contextual use, prosodic pronunciation and the pragmatic meaning of this lexical resource in communication differ in local colloquialism. Therefore, local differences in the national language are reflected in the dynamics of linguistic units. At the same time, the local linguistic features of the natural language are manifested not only in the registered words, i.e., dialects given in dictionaries, but also in conversations in everyday communication. At the same time, dialectological corpuses have been recorded in languages from local discourse and compiled into a corpus database. Such local conversations not only reflect the peculiarities of speech in the vernacular, but also reflect culture, worldview, and mentality. After all, language is the only tool that reflects the worldview of a nation, a local country.

Material and research methods

The article uses methods of review, description, narrative, analytical analysis in the study of applied work in the field of world linguistics in the development of dialectological division of the national corpus of the Kazakh language. We rely on the method of algorithmic programming in the technology of dialectological corpus development.

The research material is taken from the register dictionary "Regional Dictionary" of the National Corpus of the Kazakh language, illustrative materials of the "Explanatory Dictionary of the Kazakh Literary Language".

Literary review

In any language, local features are a way of enriching the vocabulary of that language as a branch of a literary language. Because ethnic mixing and territorial boundaries do not affect the national language. Because such relations are established in different regions of the language of one people, the presence of regional peculiarities has become a natural phenomenon. It is not to say that even poets and writers do not use regional features at all.

The peculiarities of the local language in the Kazakh language belong to one of the well-studied areas. S. Amanzholov [1], O. Nakysbekov [2], Sh. Sarybaev [3], etc. are studied in the works of scientists. The peculiarities of the local language, which became a subject of research since the 1930s, were also considered at the level of candidate and doctoral dissertations. Special regional expeditions were organized, and the peculiarities of the western, eastern, southern, and northern regions of Kazakhstan were determined. In addition, the subject of research was the language of the diaspora Kazakh in other countries, Karakalpak, Chinese, etc. The linguistic features of the Kazakhs have been studied. The oral discourse recorded during the expeditions was recorded, and these texts were described by lexical, phonetic, and grammatical levels. Thus, former Soviet dialectological features were derived from both spoken language and regional writing.

Recordings of local language conversations are taken from different speech situations, different environments, and different types of speech. All types of linguistic communication provide information about the characteristics of the language, worldview and culture of the local country. The study of dialect speech in Russian linguistics began to be actively pursued in the late twentieth and early twenty-first centuries. In this case, Kachinskaya I.B. [4, p.57], [5, p. 245], Letuchiy A.B. [6, p. 215], [7, p. 114], Sichinava D.V. [8, p. 58] and others. The works of scientists can be named. Scientists write about the importance of studying dialect speech: "Dialect speech (stories of dialect speakers, everyday communication in dialect, dialogues of villagers with representatives of other socio-cultural environment) is the most important translator of traditional folk culture. If such linguacultural phenomena as oral folklore, rituals and ceremonies, represent cultural tradition rather in its diachronic section, preserving and transmitting in a symbolically generalized and aesthetically processed form the elements of worldview and world understanding that have developed over time, the actual ethno-cultural information (the modern picture of the world of traditional folk culture speakers) is represented primarily by modern dialect speech, in which, as actualized knowledge, there can also be the In this regard, the linguacultural study of folk accents is an urgent task of modern dialectology [9].

Thus, when local features of the national language are recorded in spoken language, one can observe the pragmatic and worldview and mentality of the interlocutors. From this point of view it is necessary not only to fix local features in the language (i.e. to mark the vocabulary in dictionaries), but also to record them on an audio tape and collect them in a database. This allows for a deeper understanding of the peculiarities of local speech in the natural language and the character of the people of the region. The dialect-representative corpus is a source of structural, communicative, cognitive, linguacultural research of the local spoken language. For example, the Saratov dialectological corpus, created by the Saratov State University named after N.G. Chernyshevsky, is specifically created as a source of the above research. This corpus contains records of rural conversations in the local language. This dialectological corpus consists of three subdivisions. The first is the dialect of the village Belogornoe in the Saratov region of the Volga District, the second is the dialect of the village Zemlyanoy Khutor in the Atchara district, and the third is the dialect of the village Megra in the Vytegorsky district of the Vologda Region. This dialectological corpus contains extralinguistic data, as well as conversations from local audio recordings. For example, they include video illustrations, maps, drawings, photographs, geographical information, historical information about the region, etc. They are continuously placed in the corpus in the form of an information chain.

In the Saratov dialectological corpus, audio recordings of speeches are stored in three forms: 1 – audio recording, 2 – symbolic record/text of the audio recording, 3 – as a dedicated text module.

According to the developers of the corpus, the text of the recorded audio recordings is written close to the orthographic norm, but the lexical and grammatical features of the local language are preserved. For example:

Вот это у нас образа/ оне называются старообрядческие/ поморские/ вот наша вера какая/ мы не монашки/ мы не какие поповцы/ и не это… вот оне эти вот/ монашки-ти/ оне… как сказать… оне… тоже беспоповцы/ но… у них брак/ и считает-ся за грех/ оне называются безбрачные// оне вот доживают/ до шестьдесят лет/ женщина/ если с мужчиной живет/ это было раньше/ и у них закон такой/ доживают до шестьдесят лет/ она сносит кануны/ и говорит/ вот/ Ваня там или Вася/ всё/ я больше прекращаю с тобой жизню жить/

The genre and title of the audio recording are listed on the corpus tag. For example, audio recordings of stories, fairy tales, proverbs, songs, etc. The theme is family, religion, nature, history, etc. The developers of the corpus have numbered their types (#25@1) to identify a given genre and thematic group. From the materials given, it provides metatextual information in the form of a series of linguistic (morphological) definitions. For example:

#25@1ну{ну(ну)=PART} <…> когда{когда(когда)=ADV/ CONJ} стало{стать(стать)=V,сов=изъяв,прош,ед,сред} мне-{я(я)=S-PRO,ед,од=дат} годов{год(год)=S,муж,неод=мн,ро д=*} шестнадцать{шестнадцать(шестнадцать)=NUM=им} или{или(или)=CONJ} пятнадцать{пятнадцать(пятнадцать)=NUM=им}/ это{это(это)=PART} уже{уже(уже)=ADV} по шла{пойти(пойти)=V,сов=изъяв,прош,ед,жен} мода{мода(м ода)=S,жен,неод=ед,им}

панталоны{панталоны(панталоны)=S,жен,неод,мн=вин} белые{белый(белый)=A=мн,вин,нео д} с{с(с)=PR} кружевами{кружева=S,сред,неод,мн=твор}// <…> и{и(и)=CONJ} нижние{нижний(нижний)=A=мн,вин,н еод} юбки{юбка(юбка)=S,жен,неод=мн,вин} белые{белый(б елый)=A=мн,вин,неод}// <…> чтобы{чтобы(чтобы)=CONJ} немного{немного(немного)=ADV}/ из-под{из-под(из-под)=PR} п латьев{платье(платье)=S,сред,неод=мн,род} <…> видать{видать(видно)=PRAEDIC=*} было{быть(быть)=V,несов=изъяв ,прош,ед,сред}// и{и(и)=CONJ} начали{начать(начать)=V,сов =изъяв,прош,мн} уже{уже(уже)=ADV} юбки{юбка(юбка)=S, жен,неод=мн,вин}–то{то(то)=PART} короткие{короткий(кор откий)=A=мн,вин,неод} носить{носить(носить)=V,несов=ин ф}/ а+то{а то(а то)=CONJ} раньше{раньше(раньше)=ADV}– то{то(то)=PART} не{не(не)=PART} были{быть(быть)=V,несов= изъяв,прош,мн=*}//@1#25

The provision of subject, genre codes in the audio text allows the user to search by this designation.

Each text contains 3 metatext modules. These are: 1) textual metaphor (information about informants, time of the text, place of the text, situation of communication, recipients of the conversation, topic of conversation (pre-Soviet, revolution, Civil War, collectivization, Great Patriotic War, postwar Soviet era, post-Soviet period; 2) informant biography; 3) illustrative module (photo informant, photo illustrations, associated with the received text).

In the Saratov dialectological corpus, these metadata and lexica-morphological notations are entered into the search system and work as a single software service. The texts obtained in the corpus text base reflect the everyday life, history, mentality, and worldview of the local inhabitants of the recorded area [10].

The audio recordings recorded in the Saratov corpus were thematically large macro groups on the topics of household, home, everyday life, and religion. On the materials of the Saratov corpus, researchers of dialectology, discourse, oral speech, stylistics, linguistics, cognition, and linguacultural studies use language as a source according to their goals. Scholars conduct various studies, considering dialectological corpora of other countries from different angles. For example, on the basis of the dialectological corpus, researchers were able to identify psychological and attitudinal differences between urban and rural life, rural life skills, understanding, life, consciousness, speech, etc. revealed cognitive, cultural, religious, linguistic features. He created a model of rural-urban confrontation.

One of the most pressing issues is the creation of the dialect corpus of the Tomsk dialectological school, which studies the Siberian prose. The importance of this project is not only for the internal use of the school, but also for the scientific community. The introduction of the achievements of machine linguistics in the study of the Tomsk State University it began in the 80's and 90's of the last century under the leadership of Rakov. In 2010, the first steps of the Tomsk Dialectic Corpus (TDC) were made on the idea of E.A. Yurina. At first, field records were scanned and the graphical principles of oral communication were taken into account. However, it remained in the form of

a sketch and the work was not carried out. Later, it was reopened on the initiative of young philologists.

The Tomsk dialect corpus is considered lexically and contextually. The textual Tomsk dialect corpus required a new type of notation, reflecting the peculiarities of text organization in dialect discourse. It is conventionally referred to as text type notation.

Since the peculiarities of dialect discourse are not yet fully understood, the Tomsk dialect corpus at this stage is carried out on the basis of regular markers in oral speech. These various markers include:

a) Text types according to the level of inexperience of pronunciation of speech phenomena.

b) Genres of speech.

The first definition was considered on the basis of the relationship of dialect users in textual form to the non-coercive use of dialects, which is inherent in the form of speech. That is, purposeful discussions by dialectologists on a particular topic, verbal gestures during a conversation (Oh, my grandson woke up), complex words when informants change the topic during a conversation (What do I tell you? Try, touch). Unlike the texts of the previous type, it is oriented on a general theme ("Names of plants", "Traditions", etc.). A certain genre was indicated for each situation (the informative group included different types of messages, predictions, explanations, statements, warnings; the imperative – a request, order, instruction, order, gratitude, invitation, wish; evaluation – praise, blame, self-assessment, evaluation). Biographical narratives also include stories, descriptions, interviews, tales, etc. [11, p. 58].

However, since it is impossible to cover all these genres, it is necessary to specify the most common genres. That is, these are autobiographical stories – own stories about the informant's informal life, memoirs – a genre of speech in the narrative monologue about a person's past life, stories about other people; stories about the life of the informant himself or his close relative. Fairy tales, proverbs, etc., which occupy a special place in the culture of speech. Attention is also paid to genres.

The compilers of the Tomsk dialect corpus are working to digitize the rich vocabulary of users of Middle Pribean dialects, and dialect dictionaries have been published. The Tomsk dialect corpus plans, as a first step, to provide an explanation of the meanings of non-literal units in lexical notation. In the future, extending the corpus-dictionary connection involves identifying descriptions of lexemes (old, new, infantile, coarse, etc.) using other dictionary sources. This information is an effective tool used in the study of dialectal vocabulary and the compilation of new lexicographic works using the new corpus. The creation of the Tomsk dialect corpus, in turn, will still be the basis of dialect materials and dictionaries.

The National Corpus of the Russian Language has been building a Corpus dialectics on its website since 2005. They also recorded the dialect of the language on audiotape and made a transcription of the spelling. According to

the scientists, metadata (time, place, addressees, title, etc.) of the originally written texts were given, but the words were not linguistically labeled. Linguistic processing of each token was then initiated in 2009-2011. This dialectological corpus of the Russian language included audio recordings from various parts of the Russian region. Studies are even being conducted on these materials [12, p. 14].

The dialectological part of the Russian language also includes the original audio and phonetic transcriptions of the texts recorded on audio cassette. All dialectal texts have linguistic and metatextual (genre, thematic, phonetic features (common)) designations.

The search for metatextual notation in the dialectal division of the Russian language includes the following parameters:

1) Written geography (region: region, district)
2) Topic
3) Genre
4) Time of recording
5) Historical period, period of the event being described
6) Certain phonetic features
7) Information about informants (gender, age, education, religion)

Linguistic markup:

1) Grammar (including dialectal features)
2) Semantics (intermediate level)
3) Token search.

The dialectal division of the Russian language is said to have 2 different ways of recording grammar. The first is to analyze the peculiarities of the grammatical applications of Russian dialects by finding an invariant in accordance with the orthographic norms of the literary language and using the morphological analyzer in the main corpus. In this case, the orthographic editor (subtitle) brings the transcription to the literary norm. After the orthographic editor adjusts the transcription to the orthographic norm, it passes through the morphological analyzer used in the main corpus. If the analyzer cannot find a grammatical entry, it shows it as a dialectal grammatical feature.

In addition, words missing from the main case dictionary are highlighted as words not found in dialectal texts.

The second is to pass the morphological analyzer by putting the grammatical variation vocabulary at the base of the case. In Russian there are such grammatical features in the form of preposition, species, plural, singular only, etc. It turns out that, for example, in the imperative form *поставь (поставь), удари (ударь), посоль (посоли)*. Such a form is called dialectal. And if the local difference is at the token/base level, it is called a dialect. The dialectal word-formation is also characteristic of the Russian dialects.

According to the developers of the Russian dialectological corpus, this corpus includes not only oral texts from the regions. District newspapers, district letters, diaries, poems, memoirs, notebooks.

The texts collected in the dialectological corpus of the Russian language are divided as follows:

1) Non-folklore oral texts (dialogues, stories, narratives, answers to questions, discussions, etc.);

2) Non-folklore written texts: autobiographies, memoirs/memoirs, personal letters, congratulations, personal diaries, notebooks, official letters, statements, references, descriptions, artistic texts written by residents of the region, poems;

3) Folklore texts [13, p. 593].

The dialect corpus of the Russian language does not reflect the lexical meanings of words, grammatical designations are given. In the Russian corpus, even in the main corpus, the lexical meanings of words are not given. The developers of the dialectal corpus say that the lexical meanings of dialethism's in the dialectal corpus can be created by providing references to dialectological dictionaries, but there is no such database of dictionaries on the Internet.

The creators of the Russian dialectological corpus developed its concept in several ways. They:

1) The selection of dialect material and the definition of the criteria for the representation of the dialect corpus;

2) Principles of classification of the continuum of speech in the corpus;

3) Options for the transmission of text fragments;

4) Forms of transmission of dialect texts by case;

5) Metabolic parameters of dialect texts;

6) Rules and types of annotation of the case text base;

7) The transmission of extra-linguistic information in a dialect case;

8) Effective ways to create a dialectological research capability according to the needs of the user [14].

According to the parameters specified by O. Kryuchkova, the criterion of representativeness of the dialect corpus is determined by the maximum coverage of the territorial distribution of the national language. At the same time, dialect texts are much more difficult to insert into the corpus than normalized texts. According to the scholar, the dialect corpus is conducted with the participation of dialectologists, i.e., with the participation of scholars who study certain dialects, and many difficulties arise here, such as sound recording, writing on paper, and handwritten notes. In order to preserve the principle of representation in the dialect corpus, it is necessary to cover all genres and topics of linguistic communication in local regions, i.e. to tie important dialect types in the national language and balanced materials by regional subdivisions.

There are also problems with the placement of text fragments in the dialect corpus. Does the audio record cover the entire dialectal fragment or short fragments? For example, in the national corpus of the Russian language it consists of 3-4 sentences. Scholars believe that the smaller unit of the dialectal text corpus is a paragraph, and the larger unit is an entire audio text. There is also a big problem with the orthography or orthoepy (transcription) of audio transcripts in the dialect corpus. Orthography allows scholars to analyze texts

with a conventional morphological analyzer, but, on the other hand, prevents recognition of phonetic features of speech. A transcript is a symbolic representation of an audio or video text. According to Russian scholars, in the dialectological division of the National Corpus of the Russian language, the symbolic transformation of tape recordings was carried out differently in different dialectological centers. And in the Saratov dialect corpus, it is recorded on paper according to a special instruction. It combines oral (audio) and symbolic recording, the possibility of transition from oral text to written text and vice versa from written text to oral text (audio or video).

Linguistic designations in the Russian and Saratov dialect corpus are mostly made at the morphological level, as in the regular main corpus. The thematic and genre designations in the Saratov corpus will allow future researchers to study dialect speech in specific genres and thematic groups.

Results and discussion

Meta-textual designations in the Russian dialect corpus:

- Term of entry;
- Place of entry;
- Operator;
- General topic;
- Text size.

According to Kryuchkova, these metatextual designations in the Russian dialect corpus are not enough; it is also necessary to include a lot of additional information as extralinguistic materials, such as photographs, geographical maps, historical, ethnographic materials, etc.

Once the textual base of the dialect corpus is established, it is important to 1) allow the user to find, search, and 2) consider the possibility of researching other sources as well. Only then will the National Corpus, like the Dialectic Corpus, become an unprecedented source of research.

It is important to create a similar dialectological corpus for the National Corpus of the Kazakh language.

The National Corpus of the Kazakh language currently has the Main Corpus, which includes 5 styles of literary language. Its textual base is currently 21 million words. It contains metatextual designations specific to each style.

The dialectological division of the National Corpus of the Kazakh language is based on the materials of the «Regional Dictionary of the Kazakh Language» [15]. To do so:

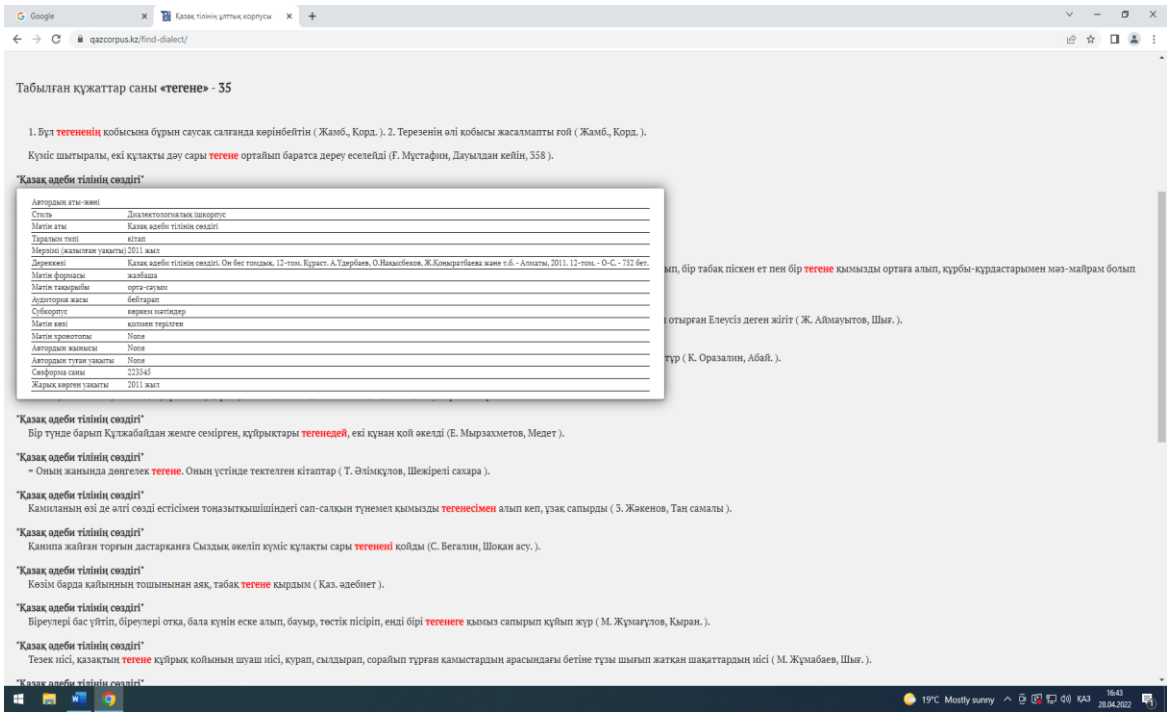
- First, the register dictionary of the "Regional Dictionary" was included in the case as a database of key words of the dialectological corpus;

- Second, examples substantiating the meaning of the word for each dialethism are included in the case as a textual base of the dialectological unit, i.e., the volume of text is made up of examples given in the "Regional Dictionary." In addition, illustrative materials from the onesome "Explanatory Dictionary of the Kazakh Literary Language" are added. The total volume is 3,294,049 words.

- Third, each word in the "Regional Dictionary" is written according to the region, which allowed to "Search by Region" in the Finder;

- Fourth, each word in the register is labeled with a word. In the linguistic window the words searched for from the dialectological division, above all, if they are in a modified face, are brought to the base (main and derived root), i.e. the lemma is separated and a lexical sign is put on this base.

It is important for local historians to know the genre of the discourse texts recorded in the corpus, the age, gender, and social characteristics (education, occupation, etc.) of the interlocutors. In this case, when collecting local dialects in the dialectological corpus, it is necessary to introduce metatextual designations in each of the dialect fragments and create a passport to these texts. That is, the dialectology unit of the National Corpus, like the Main Corpus, must be ticked. However, in the current situation, since the dialectology division is based on the materials of the Regional Dictionary, the text base in the Dictionary consists of 1 or 2-3 examples for each word. Most of the examples are collected from the country during expeditions to the regions. There are not many examples from literature. Therefore, for the time being we decided to present the illustrative (exemplary) texts of the Regional Dictionary and the illustrative text base from the "Explanatory Dictionary of the Kazakh Literary Language" only in 2 different metaphors.



As can be seen from the table, the Dictionary of the Kazakh Literary Language has been cited as the source of metatext notation. Similarly, the source of metatext notation for the examples from the regional dictionary is the regional dictionary itself. The reason we turn to these two dictionaries is that supporting markup for individual examples is unproductive. And in the main

corpus, texts are not inserted into the text base one at a time, but as a whole text (from a story to a novel).

The meta-definition includes the following parameters.

1) The lexicon of the dialect word;

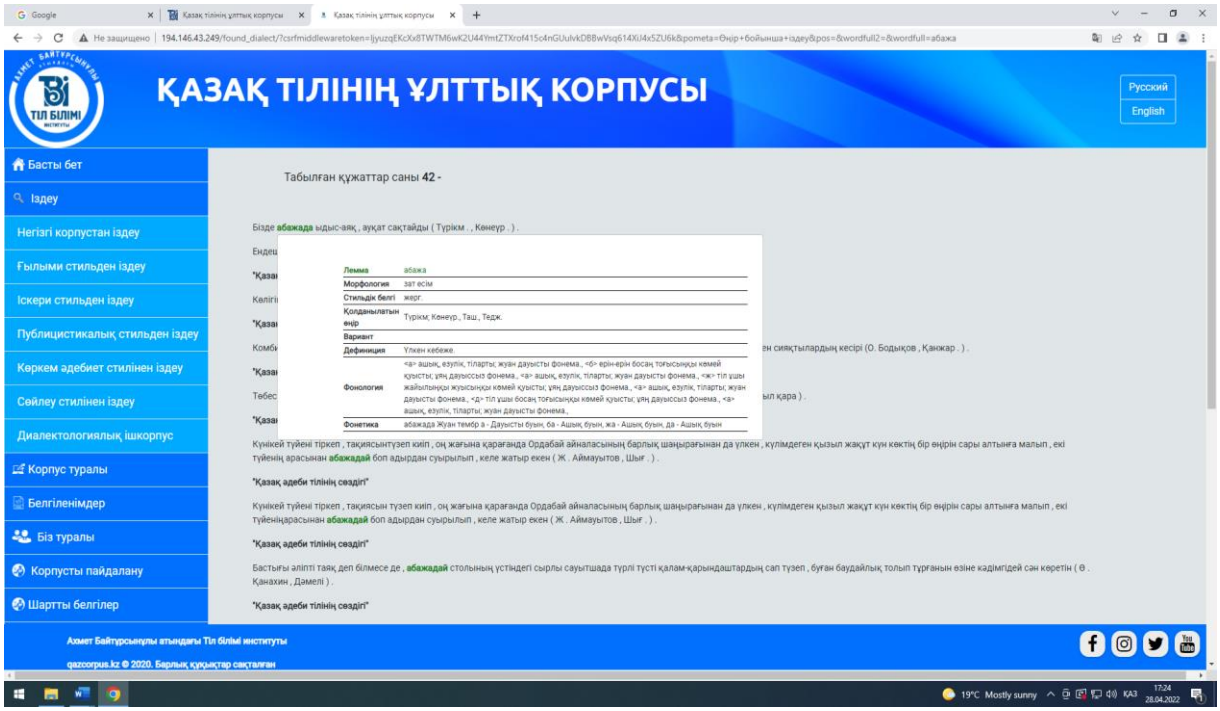
2) The geography of the dialect word (region, district, village, etc.), i.e., area;

3) Stylistic feature, i.e., it is a local word (local);

4) Variants, if the dialect is a variant word;

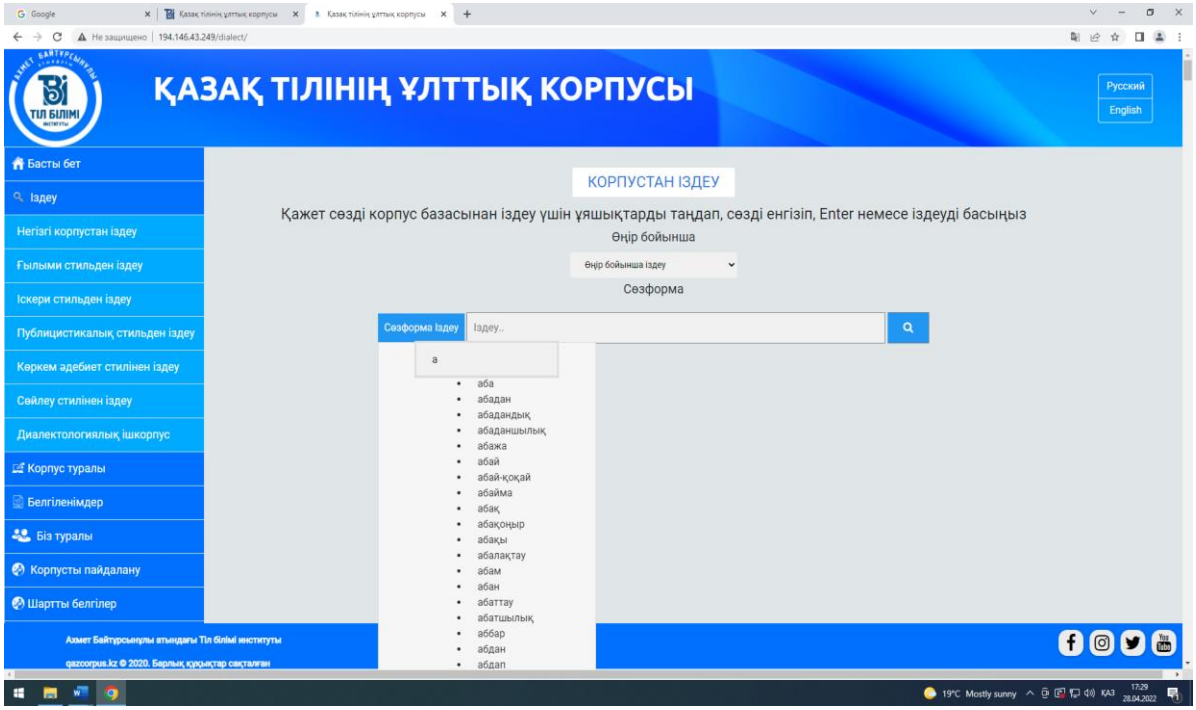
5) The definition of a dialect word;

Thus, in the word-field in the dialectal division metabolic and linguistic characters are given in combination.

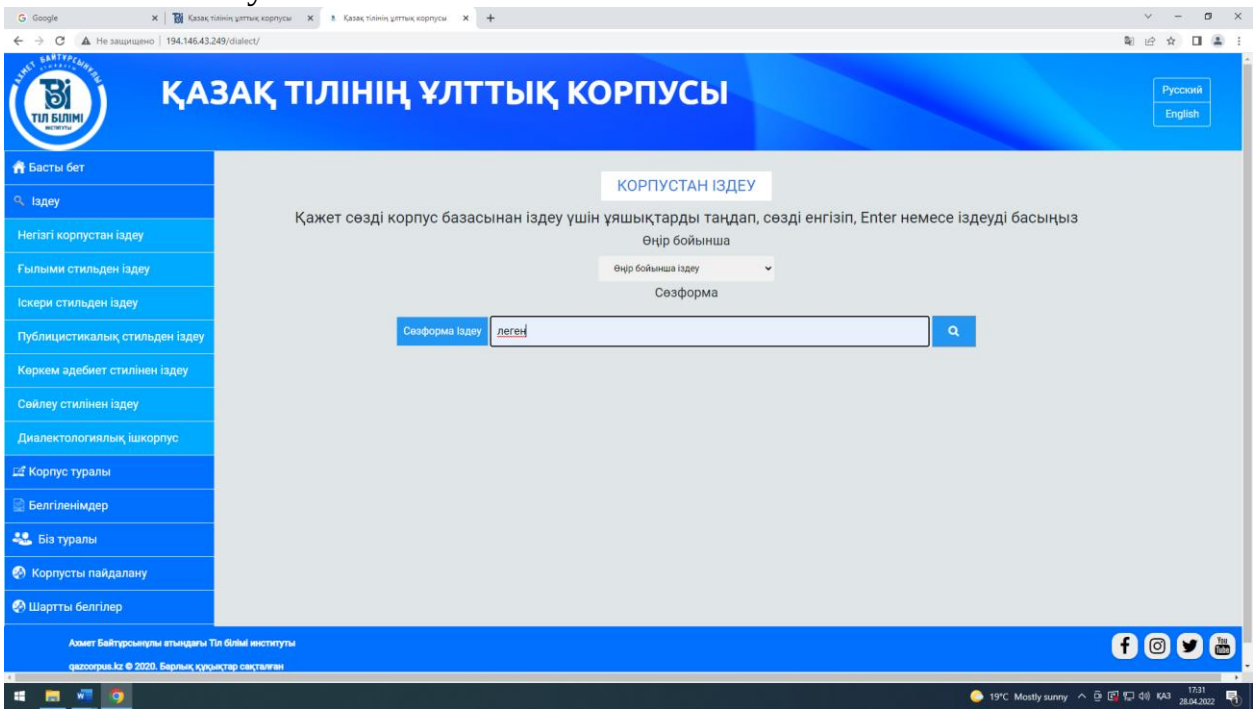


As for the dialect words, the information on the obtained texts, as mentioned above, is taken only from 2 sources ("Regional History Dictionary" and "Dictionary of the Kazakh Literary Language") and for these 2 is made a notation window metatext sources.

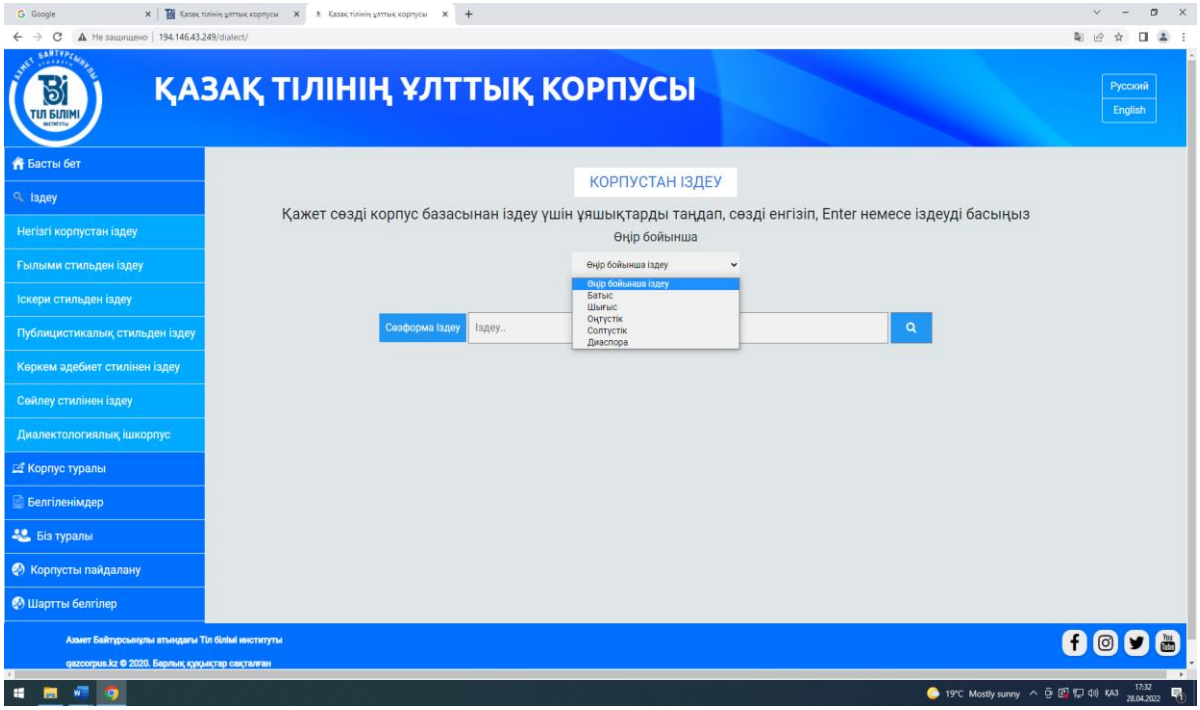
The search system of the dialectological corpus works with 3 parameters. 1 – Search in word-form:



2 – Search by word:



3 – Search by region:



This desktop is the original version of the dialectological corpus of the Kazakh National Corpus. The literature review at the beginning of the article analyzes the dialectological corpus of other languages. Thus, we have yet to study the world experience of creating a dialectological corpus and to improve the development of the Kazakh dialectological corpus. At the same time, in the future, we need to organize the work in several directions. They are:

- It is necessary to record the textual base of the dialectological corpus in the oral language of the regions, for which we need to organize dialectological expeditions;

- It is necessary to register the recordings by genre and subject, i.e., to conduct a philological expertise, which will ensure the representativeness of the case;

- When recording audio tapes, it is necessary to plan what parameters to record, i.e., what the audio recorder should pay attention to, what to record;

- It is necessary to conduct audio recording of the planned parameters, for which we must first determine the orthography or orthoepy (phonetics) of the spoken word and develop guidelines for recording transcripts, and now there are programs that automatically translate the spoken word into written form (speech analysis and synthesis), it is necessary to control the possibility of translation of these programs from oral to written, written to oral and use in the creation of the Oral Corpus, Dialectological Corpus;

- Metadata of recorded oral texts requires adding several new parameters to the metaphor window (text passport), such as information about the interlocutors (who, age, gender, education, etc.), when and where they were written and the topic of conversation, genre (Media, everyday speech, business conversation, etc.).

Conclusion

The dialectological corpus of the Kazakh language can be used by dialectologists-journalists-scientists, other scientists-linguists, philologists, historians, culturologists, ethnographers, and other non-linguistic users for various purposes. There is no doubt that dialects are part of culture. Given the importance of local dialects, the creation of a dialect corpus is one of the most pressing issues. Comprehensive sources of information are digitized every year for scientific research.

Users of dialectological subtexts and the various extra-linguistic and linguistic concepts contained in them can use ethnographic, ethno-cultural traditions, features of the Kazakh mentality, the Kazakh local language as a reference material, a source of scientific and pedagogical works. The basis of dialectological research can serve as a source only when the dialectal corpus is designed in accordance with the standard.

REFERENCES

[1] Amanzholov S. To the question of dialectology and history of the Kazakh language / S.Amanzholov. 2nd edition, vol. Almaty, Sanat, 1997. 2001. 608 p.

[2] Nakysbekov O. Variative dialect of the Kazakh language. Almaty, 1972. 176 p.

[3] Sarybayev Sh. Regional lexicon of the Kazakh language. Almaty, Nauka, 1989. 192 p.

[4] Kachinskaya I.B. Corpus of Dialect Texts in the National Corpus of the Russian Language: Status and Prospects // Lexical Atlas of Russian Folk Dialects (Materials and Research). 2009. St. Petersburg. PP. 57-68.

[5] Kachinskaya I.B. Dialect Subcorpus of the NCRY. The new delivery standard. New workplace // In: Russian oral speech. Materials of the international scientific conference “Barannikov Readings. Oral speech: Russian dialect and colloquial culture of communication. Interuniversity conference "Problems of creation and use of dialect corpora". Saratov, SSU, November 15-17, 2010, ed. O.Yu. Kryuchkova, A.I. Buranova, V.E. Goldin, L.V. Balashova. PP. 245-255.

[6] Letuchii A. B. Corpus of Dialect Texts: Tasks and Problems // National Corpus of the Russian Language: 2003-2005. M., Indrik. PP. 215-232.

[7] Letuchii A. B. Dialect corpus: composition and features of markup // National Corpus of the Russian language: 2006-2008. New results and perspectives. St. Petersburg, Nestor-History. Pp. 114-128.

[8] Russian oral speech. Materials of the international scientific conference “Barannikov Readings. Oral speech: Russian dialect and colloquial culture of communication. Interuniversity conference "Problems of creation and use of dialect corpora". Saratov, SSU, November 15-17, 2010, ed. O. Yu. Kryuchkova, A. I. Buranova, V. E. Goldin, L. V. Balashova. Yurina E. A. (2011), Tomsk dialect corpus: at the beginning of the journey // Bulletin of the Tomsk State University. university Philology. No. 2.PP. 58-63

[9] Kryuchkova O. Yu. Dialectological corpus as a source of linguoculturological study of Russian folk dialects. – Saratov https://cyberleninka.ru/article/n/dialektologicheskiy-korpus-kak-istochnik-lingvokulturologicheskogo-izucheniya-russkih-narodnyh-govorov

[10] Sveshnikova N.V. The history of the village in the folk word (the village of Belogornoye, Volsky district of the Saratov region) // Culturological and linguistic aspects of communication. Saratov, 2008. Issue. eight.

[11] Yurina E.A. Tomsk dialect corpus: at the beginning of the journey //Vestn. Volume. state university Philology. 2011. No. 2 (14). PP. 58-63.

[12] Puritskaya E.V. The Dialect Subcorpus of the National Corpus of the Russian Language as a Source for Studying the Lexical Dynamics of a Dialect // Northern Russian Dialects. Issue. 12: intercollegiate. Sat. / resp. ed. A.S. Gerd. St. Petersburg, Publishing House of St. Petersburg. un-ta, 2012. PP. 14-22.

[13] Sichinava D.V., Kachinskaya I.B. Corpus of Dialect Tests in the National Corpus of the Russian Language: Current State and Prospects // Computational Linguistics and Intelligent Technologies: Based on the materials of the annual International Conference "Dialogue" (Bekasovo, June 4-8, 2014). Issue. 13 (20). M.,Publishing House of the Russian State Humanitarian University, 2014. PP. 593-600.

[14] Kryuchkova O.Yu., Goldin V.E. Corpus of Russian Dialect Speech: Concept and Evaluation Parameters. – Saratov, https: // docplayer.com/57503155-Korpus-russkoy-dialektnoy-rechi-koncepciya-i-parametry-ocenki-1.html

[15] Regional Dictionary of the Kazakh language / Collection. G.Kaliev, O.Nakysbekov, Sh.Saribayev, A.Uderbayev and others. Almaty, Arys Publishing House, 2005. 824 p.

ҚАЗАҚ ТІЛІ ҰЛТТЫҚ КОРПУСЫНЫҢ ДИАЛЕКТОЛОГИЯЛЫҚ ІШКОРПУСЫН ӘЗІРЛЕУ ТЕХНОЛОГИЯСЫ

Жаңабекова А.Ә.[1], Қожахметова А.Қ.[2], *Тлегенова Г.Б.[3]

[1]филология ғылымдарының докторы, А. Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің меңгерушісі, Алматы, Қазақстан,

[2]докторант, А. Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің кіші ғылыми қызметкері, Алматы, Қазақстан,

[3]докторант, А. Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінің ғылыми қызметкері, Алматы, Қазақстан,

[1]e-mail: aiman_miras@mail.ru, [2]e-mail: akony_8484@mail.ru,

*[3]e-mail: gulden_20.88@list.ru

[1]https://orcid.org/0000-0002-6199-7444

[2]https://orcid.org/0000-0003-3783-3199

*[3] https://orcid.org/0000-0001-9842-3324

Аңдатпа. Мақалада диалектологиялық корпус жасаудың әлемдік тәжірибесіне шолу жасалады. Орыс тілінің ұлттық корпусында диалектілік ішкорпус ауызша тіл материалдары бойынша енгізілгендігі, ондағы фонетикалық транскрипция мәселелері мен оларды орфографиялау, просодикалық белгіленімдер әзірлемесі және диалектілік қолданыстарға автоматты морфологиялық талдаулар жасаудың жолдары көрсетіледі.

Мақала жазудағы басты мақсат диалектілік корпус жасаудың әлемдік тәжірибесі негізінде қазақ тілінің диалектологиялық ішкорпусының әзірлемесін сипаттап, оны жетілдірудің жолдары қарастырылады.

Диалектологиялық ішкорпусы тіл зерттеушілерге ғылыми мақалалар, монография, диссертациялық жұмыстарына қажетті зерттеулерді оңайлату және тездету үшін қажетті.

Мақалада Қазақ тілінің Ұлттық корпусының диалектологиялық ішкорпусын әзірлеуде, осы бағыттағы әлемдік тіл біліміндегі қолданбалы бағыттағы жұмыстарды зерттеуде шолу, сипаттау, баяндау, аналитикалық талдау әдістері, алгоритмдік программалау әдісі қолданылады.

Қорыта келгенде, корпуста әдеби тілдегі нормаланған сөздер болсын, диалектілік сипаттағы сөздер болсын ең алдымен түбір мен қосымшаға бөлінуі керек. Диалектілік корпуста да осылайша диалектизмдердің негіз сөздер сөздігі жасалып, ол корпус базасына салынады. Морфологиялық талдау жасаудың екінші қадамы – сөзформалардың екінші бөлігі – түрленім қосымшаларды морфемаларға бөлшектеп, олардың грамматикалық сипатына қарай шартты белгісін қою. Осы ретте диалектілік ішкорпуста да диалектизм сөзформалар морфологиялық талдаудан өтуі қажет.

Келешекте диалектілік ішкорпуста өңірлерден ауызша тіл материалдарын аудиотаспаға жазып, корпусқа енгізу жоспарланған. Мақалада осы мақсатқа сәйкес жұмысты ұйымдастыру бойынша ұсыныстар берілді. Бұл жұмыстың құндылығын арттыра түседі.

Тірек сөздер: диалект, диалектілік корпус, мәнмәтіндік редактор, талдауыш, ішкорпус, метабелгіленім, аймақтық сөздік, метаақпарат.

ТЕХНОЛОГИЯ РАЗРАБОТКИ ДИАЛЕКТОЛОГИЧЕСКОГО ПОДКОРПУСА В КОРПУСЕ КАЗАХСКОГО ЯЗЫКА

Жанабекова А.А.[1], Кожахметова А.К.[2], *Тлегенова Г.Б.[3]
[1]доктор филологических наук, Институт языкознания имени А.Байтурсынулы, заведующий отделом Прикладной лингвистики, Алматы, Казахстан,
[2]докторант, Институт языкознания имени А.Байтурсынулы, младший научный сотрудник отдела Прикладной лингвистики, Алматы, Казахстан,
[3]докторант, Институт языкознания имени А.Байтурсынулы, научный сотрудник отдела Прикладной лингвистики, Алматы, Казахстан,
[1]e-mail: aiman_miras@mail.ru, [2]e-mail: akony_8484@mail.ru,
*[3]e-mail: gulden_20.88@list.ru
[1]https://orcid.org/0000-0002-6199-7444
[2]https://orcid.org/0000-0003-3783-3199
[3] https://orcid.org/0000-0001-9842-3324

Аннотация. В статье рассматривается мировой опыт создания диалектологического корпуса. Изучены введение диалектного корпуса в национальный корпус русского языка на материалах устной речи, проблемы фонетической транскрипции и правописания, разработка просодических обозначений и способов автоматического морфологического анализа диалектных употреблений.

Основная цель статьи – описать развитие диалектологического корпуса казахского языка на основе мирового опыта создания диалектного корпуса и путей его совершенствования.

Развитие диалектологического подразделения Национального корпуса казахского языка необходимо исследователям языка для упрощения и ускорения исследований, необходимых для научных статей, монографий, диссертаций.

В статье используются методы обзора, описания, повествования, аналитического анализа, алгоритмического программирования при разработке диалектологического подразделения Национального корпуса казахского языка, изучении прикладных работ в мировой лингвистике в этой области.

В заключение следует отметить, что в корпусе как стандартизированные слова литературного языка, так и слова диалектного характера должны быть разделены в первую очередь на корни и аффиксы. В диалектном корпусе таким же образом

создается словарь ключевых слов диалектов, который размещается в базе данных корпуса. Второй этап морфологического анализа заключается в разбиении суффиксов второй части словоформ на морфемы и маркировке их по грамматической природе. При этом диалектные словоформы в диалектном корпусе также должны подвергаться морфологическому анализу.

В дальнейшем планируется записывать устные языковые материалы из регионов в диалектный корпус и включать их в состав корпуса. В статье даны рекомендации по организации работы в соответствии с этой целью. Это увеличивает ценность работы.

Ключевые слова: диалект, диалектологических корпус, текстовой редактор, анализатор, субкорпус, метаразметка, региональный словарь, метаданные.

Статья поступила 18.05.2022