

Dialect Corpora: An Interactive Digital Platform for Uzbek Dialects

MUYASSAR KHOLOVA

Termez State University, Termez, Uzbekistan

ABSTRACT

This research is about the stages of developing dialectal corpora for the Uzbek language, using the experience and models of existing dialectal corpora from around the world. The article focuses on analyzing the data and functionality of DialectCorpora, a resource designed for audiences interested in Uzbek dialects and researchers working in the field of Turkic languages. Accordingly, the dialectal corpus is dynamic in nature; it currently includes 338 recordings, over 61,000 tokens and 64 topics. At the same time, we sought to evaluate the corpus – considering its structure and diverse usage capabilities – as a unified ecosystem.

Keywords: Dialectal corpus, Uzbek dialects, digital platform, corpus linguistics, dialectology, audio material, geostatistics.

INTRODUCTION

The digitization of linguistic units along with the dialectal variations they contain has come to require automation in corpus linguistics. Thus, dialect corpora and related research have begun to play an integral role in the operation of smart devices today. Today, we can observe the operation of highly prominent dialectal corpora: FRED (Freiburg English Dialect Corpus) covers English dialects, and Szmrecsanyi & Hernández (2007) developed a user guide for this specific corpus [6]. The corpus contains material collected from 300 informants across various regions of England between 1970 and 1990, and Anderwald &

Wagner have commented on the productivity of this corpus [2]. Furthermore, the work done by J. Johannessen on the NorDic corpus [5], the perspectives of Dash & Ramamoorthy regarding corpus-based dialect studies[4], and the work carried out on corpora such as NECTE [3] and SCOTS [1] serve as noteworthy examples (cf. Table 1)

Table 1. *Comparative table of the world's dialectal corpora*

Corpus name	Language/Dialect	Size	Annotation type	Year created	Main purpose
FRED	English dialects	2.5 million words	Morphosyntactic	1990-2000	Dialect syntax and morphology
NECTE	Tyneside English	1 million words	Morphosyntactic, phonetic	1960-2000	Diachronic variation
SCOTS	Scots	4.6 million words	Morphosyntactic	2000-2007	Written and spoken Scots
Nordic Dialect Corpus	Scandinavian languages	2.5 million words	Morphological, syntactic, phonetic	2009-2012	Multilingual dialect research
CORDIAL-SIN	Portuguese dialects	700,000 words	Morphosyntactic	2000-2010	Dialect syntax
Czech DIALECT	Czech dialects	2 million words	Morphological	2010-2014	Diatopic and diachronic variation
CHARG	Guayaquil Spanish	100 hours of audio	Phonetic, morphological	2020-2023	Radio speech analysis
English Dialects App	English dialects	50,000+ samples	Phonetic	2015-2018	Crowdsourced dialect mapping
Mishar Tatar Corpus	Tatar dialect	50,000 words	Morphological	2015	Documenting Turkic dialects
Kazakh Dialectal Corpus	Kazakh dialects	200,000 words	Morphological	2022	Kazakh dialect analysis

A corpus is recognized as a linguistic corpus only if it meets the established requirements during its creation; accordingly, intralinguistic and extralinguistic parameters come to the forefront.

Representativeness is considered one of the key requirements in this regard; In analyzing the relationship between corpus and dialectal research, Dash & Ramamoorthy examine this approach

in detail [4]. They raise the following issues regarding representativeness:

- geographical coverage;
- social stratification;
- speech genres;
- various levels of language.

As we construct language corpora and their sub-corpora, we consider an approach that takes into account the specific characteristics of each language to be the most optimal.

THE MAIN FINDINGS AND RESULTS

Uzbek national sub-corpus

The development of the linguistic foundations for the Uzbek dialectal corpus can be pursued in two directions.

In the first stage, the material is collected within a dialect corpus based on a thematic approach that reflects the ethnic population's way of life, distinctive views, culture, and occupational activities, while taking into account extralinguistic parameters.

Tasks assigned to the direction

1. Developing a passport for the materials in the dialectal corpus.
2. Categorizing and entering materials into the dialectal corpus by topic.
3. Incorporating materials from various genres into the dialectal corpus.
4. Ability to conduct research utilizing materials from the dialectal corpus (Figure 1).

Text meta-markup

While the process of automatically classifying and sorting texts within the dialect corpus involves categorizing them into poetic, prose, and mixed (poetic-prose) forms based on their constituent

lexemes, words, and sentences, thematic meta-annotation entails assigning small dialectal text segments to specific categories based on the lifestyle, mentality, and worldviews of the ethnic population. Genre metathematics involves the process of dividing dialectal texts into genres in a corpus of dialects. Passport metadata in a dialect corpus represents comprehensive information regarding the dialectal texts specifically, the time and location of the recording, the speaker, and the person who prepared the material for the corpus.

The screenshot shows the 'DialectCorpora' website interface. On the left is a sidebar menu with options like 'Asosiy panel', 'Content', 'Ilimiy manba', 'Sahifalar', and 'Passport Meta'. The 'Passport Meta' section is expanded, showing a list of fields: Collector, Dialect, Equipment, Genre, Grade, and Location. The main content area displays the 'Meta' form, which is titled 'Passport meta data'. The form contains several dropdown menus and input fields: 'To'plovchi', 'Lahja', 'Genre', 'Source', 'Matn', 'Topic', 'Year start', 'Year end', and 'Amalda'. The 'Genre' and 'Matn' fields are highlighted with red boxes. At the bottom of the form are three buttons: 'Yaratish', 'Yaratish va yana boshqa yaratish', and 'Bekor qilish'.

Figure 1. Representation of extralinguistic parametr in DialectCorpora.uz

When compiling a unified corpus for the Kipchak, Karluk, and Oghuz dialects, we must while certainly accounting for the regional affiliation of the speaker representing each dialect include details such as the time, place, and region of the recording; the topic of the conversation; the genre of the oral or written source; and the textual format of the recorded material. In practical terms, a website named dialectcorpora.uz has been launched for this purpose, and it fully represents the corpus of categorized dialects.

In the field of corpus linguistics, dialect corpora represent a new area of research for our linguistics community; although the linguistic foundations for launching this site have been established, building the database requires a substantial team

comprising experts in both dialectology and information technology. Site interface components:

1. Home page.
2. Scientific resources related to the field.
3. Contact.
4. About the project.
5. A keyboard for transcription symbols.

Each page evokes a certain imagination within the context of the dialect. For instance, the Uzbek language is classified as a three-dialect language; consequently, one can gain a certain amount of information within the scope of these dialects:

Source for the Karluk dialect

The fact that most Turkic languages were formed on the basis of ancient Turkic ethnic groups and tribes is reflected in history. For example, while the languages of the Kazakhs, Kyrgyz, and the Turkic peoples of Siberia were formed on the basis of the language of the Kipchak, Azerbaijani, Turkmen, Turkish, and Gagauz, were formed on the basis of the Oghuz and Uyghur language took shape based on the language of the Karluk clans and tribes, the Uzbek language was formed through the merging of the languages of all three ethnic groups. The Uzbek language was formed on the basis of the languages of the Karluk groups and tribes. The Karluk dialect. This dialect group primarily covers urban-type dialects and sub-dialects. This dialect differs from other dialects in that it is divided into two groups according to the preservation of synharmonism. That is, in urban and city-type dialects, the law of synharmonism does not apply (only some of its manifestations exist), therefore, in urban and city-type dialects, the number of vowels does not exceed 6 or 7, while in the northern Uzbek and Andijan dialects, synharmonism is preserved and the number of vowels in them is 9.

This dialect is divided into 4 subdialect groups: 1) Tashkent group dialects, 2) Fergana group dialects, 3) Samarkand-Bukhara group dialects, 4) Turkistan-Shymkent-Sayram group dialects.

Source for the Oghuz dialect

Sh. Shoabdurahmonov links the dialectal diversity of the Uzbek language to the process of the Uzbek people's formation and emphasizes that the phonetic, lexical, and grammatical differences found in the language arose from the sequential migrations of Turkic tribes. According to him, the migration of the Karluks, followed by the Oghuz and later the Kipchaks, to Central Asia led to the formation of the dialectal composition of the Uzbeks. Oghuz dialect. In ancient times, the Oghuz people formed along the Ile River and in the 9th-10th centuries lived in the middle and lower reaches of the Syrdarya and the Aral Sea regions. Later, they migrated westward. Speakers of this dialect reside in the Khorezm Region and the Republic of Karakalpakstan of the Republic of Uzbekistan; in the Forish, Olot, and Qorako'l districts and certain villages of the Samarkand Region; in the Toshhovuz area of Turkmenistan; and in the settlements of Qamoq, Iqon, Qorabuloq, Qora Murt, and Mankent in Kazakhstan. The Oghuz dialect is predominantly widespread in the Khorezm oasis, though this does not mean the oasis consists solely of this dialect. As F. Abdullayev points out, dialects belonging to the Kipchak group are also observed in this region.

Source regarding the Kipchak dialect: Kipchak tribes also lived in the Altai region in the distant past. In the 7th and 8th centuries, they were part of the Turkic Khaganate. In the mid-13th century, they remained part of the kimak // imak (yimak) state. After this state disintegrated in the 11th century, the Kipchaks became politically active. In the 7th–8th centuries, the Kipchaks lived along the Irtysh River, but in the 9th–11th centuries, they migrated to the Volga region. Today, the descendants of the Kipchaks inhabit regions of Mongolia, the Altai, Central Asia, and the Volga area. Speakers of this dialect live in the Tashkent, Syrdarya, Samarkand, Navoi, Kashkadarya, Surkhandarya, and Jizzakh regions of the Republic of Uzbekistan, as well as in the Sozoq district of Kazakhstan and the villages of Suvnoq and Chomoq in the Turkistan region.

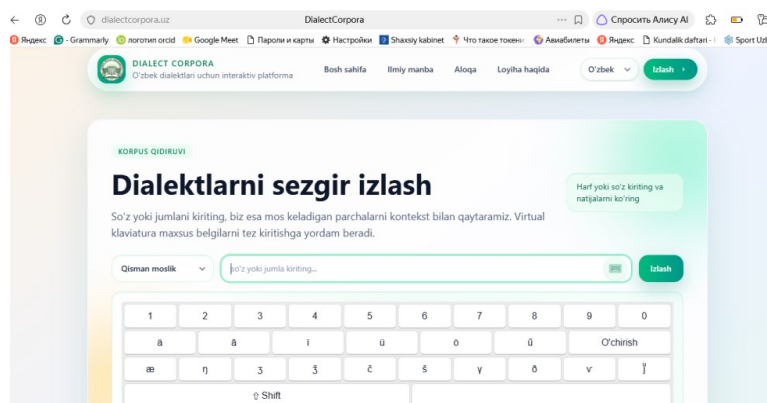


Figure 2. <https://dialectcorpora.uz/search> search bar

The second stage of work on the dialectal corpus and the tasks set for it:

1. Developing the phonetic variability of dialects within the dialectal corpus based on differentiation and integration, and selecting the most suitable variant for the sub-corpus.
2. Developing a phonetic meta-annotation system for dialects within the dialectal corpus and preventing errors in the implementation of automatic analysis processes.
3. Developing the morpholexicon of the dialects.
4. Establishing a morphological meta-annotation system for the dialectal corpus.

Here is the question: why is the lexical tagging of the dialectal corpus not developed? In dialectology, it is advisable to conduct the process of lexical meta-tagging based on a thematic corpus; a distinctive feature of dialects is that one can observe how lexical units sharing a common basis undergo gradual evolution and acquire diverse interpretations. For instance, when naming plants, the local population focused on the color of the plant itself or its produce *ko'kat*, *oqpaycha*, *siyodona*; in naming herbs known for their medicinal properties, associations with other objects were taken into account *shibit* (*ukrop*) *arpabodiyon*, *tuyato'pon*, *ituzum*; substances extracted from certain plants had

their own distinct names *ko'knor-chakida*; human influence on plants is reflected in specific phytonyms *chukuri-tashbasti-tasangi*; and plants were named according to their developmental stages *roba-tataron-kavrak* [7].

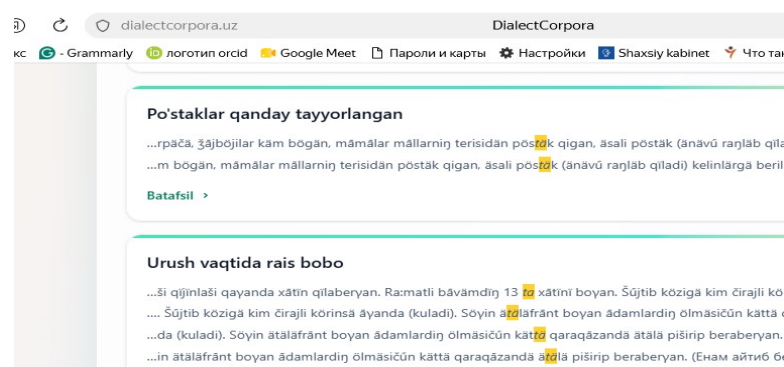


Figure 3. The bigram search result for the "tä" bigram at <<https://dialectcorpora.uz/search>>

3 modules of the metatexts regarding dialects should be structured as follows

Module 1: Text meta-markup containing information about the text

1. the time and place where this recording (audio, video, manuscript, voice recording) was made;
2. purpose of speech;
3. the fundamental essence of speech (explanation, memory, information, everyday life needs, emotional experiences);
4. persons mentioned in the speech process;
5. whether the event occurred before or after independence: collective farms, socio-economic changes (construction of a school or factory; history of the opening and closing of a cinema, cultural center, government institution, or medical facility);
6. list of speech texts.

Module 2: Autobiography of the informant from whom the text was recorded (year of birth, place, social background, education, current place of residence).

Module 3: Illustrative module (images or photographs that provide information, correspond to the text, and reveal the essence of the topic; e.g., regarding the yurt: the yurt itself, *keraga*, *beldav*, *uvuq*, *changaroq*, *jandari*, *qalam*).

A dialect corpus dataset not only offers a partial solution to pressing issues regarding dialectal units but also embodies a unique speech culture and a collection of spiritual perspectives that reflect the daily lives of the local population. Within the corpus, text genres and sub-genres (or sub-categories) function as integral components of the dialectal text's passport metadata.

Writing sample:

No. Questionnaire

- Location of recording (detailed information by MFY, district, city, and region facilitates search and sorting easily)
- Date of recording (year, day, month) _____. _____. ____.
- Informant (surname, first name, patronymic)
- Informant's gender (female, male)
- Informant's place and year of birth (district, region, republic)
- Informant's education (no formal education, higher education (Bachelor's, Master's, PhD), higher education (5-year program))
- Informant's profession
- Data collector (surname, first name, patronymic)
- Institution where work is being conducted
- Type of recording (video, audio, written) -
- Notebook or record No. ____
- Period of the event:
 - Pre-independence - year
 - Post-independence
- Topic: e.g., {iqtisodiy} <xalq xo'jaligi><chevarlik><chopon tikish>
- Text form: (oral, written)
- Genre: conversation

DISCUSSION

Analyzing and compiling statistics on a dialect within a dialectal corpus enables extensive research into the new lexical-morphological and stylistic units present among the ethnic population; simultaneously, it allows for the study of various themes and speech genres reflecting dialectal communicative interactions, as well as the cognitive characteristics of dialect-based communication and the worldview associated with it. O. Kryuchkova notes that the thematic tagging of a dialectal text corpus makes it possible to study identified themes as a unified text or to examine a specific concept across multiple domains [8].

Table 2. *Example of a dialectal metatext module*

No.	Questionnaire	Responses
1.	Recording location (district, region, republic)	Toqchi Mahalla, Boysun District, Surkhandarya Region, Republic of Uzbekistan
2.	Recording date (year, day, month)	May 11, 2023
3.	Informant (surname, first name, patronymic)	Satova Tursunoy Jabborovna
4.	Axborot beruvchining(jins: ayol, erkak)	Female
5.	Informant's place and year of birth (district, region, republic)	Boysun District, Olmazor
6.	Informant's education (secondary education, higher education (bachelor's, master's, PhD), higher education (5-year program))	Secondary education
7.	Informant's profession	Chopon maker
8.	Data collector (surname, first name, patronymic)	Muyassar Xolova (Termez State University)
9.	Affiliated institution	Termez State University
10.	Recording type (handwritten notes, video camera, audio recording, voice recorder, still camera)	Video camera, still camera, voice recorder
11.	Notebook or data record No.	-
12.	Period of the event: Pre-independence Post-independence	{iqtisodiy} <xalq xo'jaligi><chevarlik><chopon tikish>
13.	Topic	Chapan-making
14.	Text form: (oral, written)	Oral
15.	Genre	Conversation

Module 1. Conversation about *Chopon* making. Text meta-markup containing information about the text.

1. this recording (video camera, voice recorder, and still camera) was made on May 10, 2020, in the Toqchi neighborhood, featuring Tursunoy Jabborovna Sattorova, a practitioner of the traditional craft of chopon making;
2. purpose of the speech: to identify and explore dialect-specific words by providing secrets for acquiring professional skills and qualifications.
3. the fundamental essence of speech: emotional experiences, information, explanation, memory, and everyday life needs;
4. individuals mentioned during the conversation: spouse, father-in-law, an elderly female neighbor (whose name she could not recall), daughter Umida, daughter-in-law Jabborova Gulnora;
5. whether the event took place before or after independence: the narrative recounts that the events related to chopon making specifically, her father-in-law turning a *chopon* inside out occurred before independence, in the 1960s, after she had married;
6. list of speech topics:
 - a. General annual information regarding the profession;
 - b. The need for skills to master the profession;
 - c. Motivation to master the profession;
 - d. Transmission of the profession from generation to generation;
 - e. Unique aspects (secrets) of the profession.

Module 2. Autobiography of the informant (year and place of birth, social background, education, current place of residence).

Module 3. Illustrative: A photograph of the person providing the information, or an image or photo that corresponds to the text and reveals the essence of the topic.

In dialectological discourse, a respondent's autobiographical narratives unlike formal autobiographies lack a formal character and do not always follow a strict chronological sequence during oral communication. Nevertheless, just like the author of a

formal autobiography, the respondent retains the right to select the aspects of their life and past that they deem significant. Most such accounts are structured around the progression of events relevant to the listener meaning the narrator approaches the events from a subjective perspective [9].

Now, we would like to share some initial experiments that were developed using Surkhandarya dialects as a model but were designed for collecting materials across various regions of Uzbekistan. DialectCorpora (dialectcorpora.uz) is an interactive digital platform created for Uzbek dialects. By consolidating materials from the Kipchak, Oghuz, and Karluk dialect groups into a unified corpus, it enables researchers to search, compare, and analyze dialectal data. The platform's homepage is designed to provide users with an immediate, comprehensive overview of the corpus's structure and scope upon arrival. A navigation bar at the top directs users to key sections, including the homepage, scholarly resources, contact information, and project details. The interface supports multiple languages and features a quick-search function, making it easier for users to work with the corpus. The purpose of the corpus is stated concisely and clearly at the center of the page: it is presented as a digital collection covering three language groups the Kipchak, Oghuz, and Karluk dialects. Each dialect is featured in a separate column, providing a brief description along with information on its historical development and geographical distribution. This approach enables the user to grasp the differences between the dialects right from the start.

An information panel displaying live corpus statistics is located on the right side of the interface. This panel presents the corpus's key metrics: 338 dialectal recordings, 61,034 tokens, and three Uzbek dialects (note: the corpus is classified as a dynamic corpus). Additionally, the panel provides data that further enriches the corpus 33 data collectors, 26 genres, and 64 topics. These metrics provide an immediate overview of the corpus's size and diversity. The platform's functionality is organized around three main areas. The dialect search feature enables searches based on partial or exact matches and allows users to retrieve specific information from contextual excerpts. The scholarly resources section facilitates the study of dialectal layers

through filtering by genre, topic, and source. Meanwhile, the connections section is designed for researchers interested in academic collaboration and contributing to the corpus. Significantly, DialectCorpora is integrated with other related projects within a unified ecosystem. Among these, it provides access to and enables the use of platforms such as Dialekt.uz an open database for Uzbek dialects and Areal Dialect (areal.dialect.uz), which allows for the creation of surveys and the preparation of statistical and geostatistical reports.

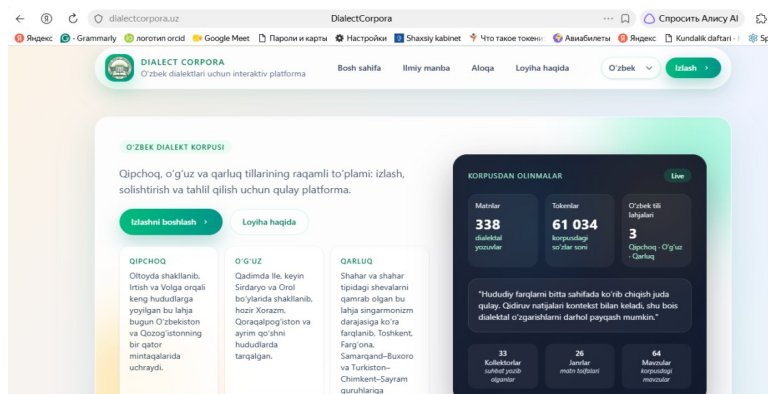


Figure 4. The interface of DialectCorpora (<http://dialectcorpora.uz/>)

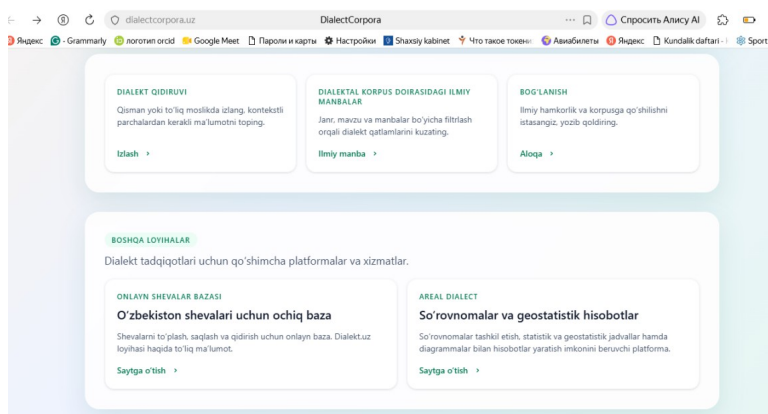


Figure 5. DialectCorpora interface (<http://dialectcorpora.uz/>)

This integration enables the user to carry out the entire research process from collecting corpus data to dialectometric mapping within a single environment.

One of the platform's significant new features is its capability to handle audio material. When a user searches for a word, the system directs them to the corresponding written text; there is also the option to listen to the material in full on that page. Thus, the written text and the audio recording are presented together.

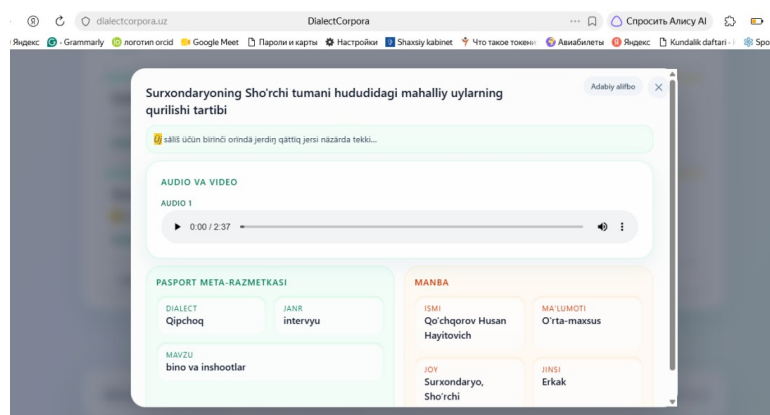


Figure 6. Redirecting from DialectCorpora's lexeme search to audio material

CONCLUSION

This aspect holds particular value for dialectology, as it enables the study of dialect-specific pronunciation, intonation, and phonetic features through direct listening rather than relying solely on text.

Overall, the platform features a simple, intuitive, and user-friendly interface. It creates an environment conducive not only to storing and searching for dialectal materials but also to observing and deeply analyzing regional variations. It serves as a medium for preserving and conveying dialectal speech encompassing everyday interactions, conversations between rural residents and individuals from other socio-cultural backgrounds,

information shared by dialect speakers, and unique interpretations of traditional customs [8].

A dialect corpus compiled in this manner offers a partial solution to pressing issues facing dialects; simultaneously, it serves as a repository of speech culture and the spiritual values that reflect the daily lives of the local population.

The passport metadata of the dialect corpus serves as a unified platform for research based on parameters such as the recording location, time, and the informant's details (surname, first name, patronymic, gender, place and year of birth, education, and profession), as well as information regarding the data collector (dialectologist, researcher, student, or independent collector), the affiliated institution, the timeframe of the recorded event, the topic, the genre, and the text format.

Studying a dialect within a dialectal corpus enables extensive research into new lexical-morphological and stylistic units present among the ethnic population; simultaneously, it allows for the analysis of diverse topics and speech genres reflecting dialectal communicative interactions, as well as the cognitive characteristics of dialect-based communication and the worldview inherent to the speakers. The thematic tagging of the dialectal text corpus facilitates the study of identified themes through both inductive (from specific to general) and deductive [11](from general to specific) approaches enabling, for instance, the analysis of a single concept across multiple domains or as a holistic entity.

REFERENCES

1. Anderson, J., Beavan, D. & Kay, C. 2007. *SCOTS: Scottish Corpus of Texts and Speech*. https://doi.org/10.1057/9780230223936_2.
2. Anderwald, L. & Wagner, S. 2007. *FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data*. https://doi.org/10.1057/9780230223936_3.
3. Beal, J., Corrigan, K., Rayson, P. & Smith, N. 2007. *Writing the Vernacular: Transcribing and Tagging the Newcastle Electronic Corpus of Tyneside English* – NECTE.

4. Dash, N. S. & Ramamoorthy, L. 2019. *Corpus and Dialect Study*. https://doi.org/10.1007/978-981-13-1801-6_9.
5. Johannessen, J. B., Priestley, J., Hagen, K., Nøklestad, A. & Lynum, A. 2012. *The Nordic Dialect Corpus*.
6. Szmrecsanyi, B. & Hernández, N. 2007. *Manual of Information to Accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. Freiburg.
7. Xolova, M. 2020. Brief Notes on Phytonyms in the Lexicon of the Boysun Dialect. *Scientific Bulletin of Namangan State University*, 3, 316.
8. Kryuchkova, O. Yu. & Goldin, V. E. 2008. A text-based dialect corpus as a model of traditional rural communication. In *Proceedings of the International Conference "Dialog 2008"* (pp. 268-273).
9. Voloshina, S. V. 2008. *The Speech Genre of Autobiographical Narrative in Dialect Communication* [Dissertation abstract in philology]. Tomsk.
10. Kryuchkova, O. Yu. *A Dialect Corpus as a Source for the Linguocultural Study of Russian Folk Dialects* (pp. 67-85).
11. Deduction. *Wikipedia*. <https://uz.wikipedia.org/wiki/Deduksiya>.

MUYASSAR KHOLOVA
TERMEZ STATE UNIVERSITY,
TERMEZ, UZBEKISTAN.