AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

PHONETIC ANNOTATION OF DIALECT UNITS IN THE UZBEK LANGUAGE DIALECTAL CORPUS (BASED ON THE EXAMPLE OF THE DIALECT OF THE JARQORGAN DISTRICT)

https://doi.org/10.5281/zenodo.20713588

Kurbonazarova Marvarid Bozor kizi
Termez State University
First-year Master's student in the field of Computer Linguistics
marvaridqurbonazarova@gmail.com

Annotation

This study examines the issues of phonetic annotation of Jarkurgan district dialect units within the dialectal corpus of the Uzbek language. Scientific views on the study of the phonetic features of dialects in Uzbek dialectology are analyzed, and a linguistic description of the phonetic changes observed in the Jarkurgan dialect is given. The theoretical and practical aspects of phonetic annotation models used in world dialectal corpora are also highlighted. The main part of the research is devoted to the development of linguistic and software for the phonetic annotation of Jarkurgan dialect materials in the Uzbek dialectal corpus. The corpus's search engine, the system of extralinguistic parameters, phonetic tagging criteria, and annotation processes implemented in the Google Colab environment are described. The research results serve as a methodological basis for implementing corpus technologies in Uzbek dialectology, digitizing dialectal materials, and automating phonetic research.

Key words

dialectal corpus, phonetic annotation, Jarkurgan dialect, corpus linguistics, phonetic variation, UPOS, XPOS, phonetic tagging.

Dialects, which are considered regional manifestations of the language, are one of the important linguistic sources reflecting the historical development, ethno-cultural ties and language evolution of the people. Within the dialects of the Uzbek language, the dialect of the Jarkurgan district of the Surkhandarya oasis has its own phonetic, lexical, and grammatical features. The study of these features based on corpus technologies, along with traditional descriptive methods, is one of the important directions of modern dialectology.

Today, the creation of dialectal corpora and the implementation of multi-layered annotation systems in them are widely developing in world linguistics. Such corpora allow for the automatic analysis of the phonetic, morphological, and

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

syntactic features of dialects. In Uzbek dialectology, there is also a growing need to digitize dialect materials and present them in the form of a corpus. From this point of view, the development of a methodology for the phonetic annotation of Jarkurgan dialect materials is of scientific and practical importance.

CHAPTER 1. ISSUES OF STUDYING AND ANNOTATING THE PHONETIC NATURE OF DIALECT UNITS IN UZBEK AND WORLD DIALECTOLOGY

1.1. Works done in Uzbek dialectology to study the phonetic nature of language units

In the process of formation and development of Uzbek dialectology, the study of phonetic phenomena developed as a separate direction. In the initial studies, the phonetic composition, sound system, and pronunciation features of Uzbek dialects were studied based on the descriptive method. Later, methods of comparative-historical, linguogeographical, and experimental phonetics began to be widely used in dialectological research.

The works of scholars such as V. Reshetov, Sh. Shoabdurakhmonov, A. Gulyamov, F. Abdullaev, B. Juraev, and N. Rajabov play an important role in the study of Uzbek dialect phonetics. In these studies, the vocal and consonant systems of dialects, variants of vowel and consonant phonemes, phonetic differential features, and regional differences were studied.

In traditional dialectological research, phonetic phenomena were primarily analyzed based on the following methodological approaches:

First, based on a descriptive approach, the sound system in the dialect and its relationship with the literary language are determined.

Secondly, the origin and historical development of phonetic phenomena were investigated using the comparative-historical method.

Thirdly, based on the linguogeographic approach, the territorial distribution and isoglosses of phonetic phenomena have been determined.

Fourthly, the acoustic and articulatory properties of sounds were studied using experimental phonetics methods.

In recent years, as a result of the development of corpus linguistics, the trend of using electronic corpora in dialectological research has intensified. This made it possible not only to describe phonetic units but also to perform statistical analysis, determine the frequency of use of variants, and perform an automatic search. Thus, phonetic research in Uzbek dialectology is moving from the traditional descriptive stage to the digital and corpus-based research stage.

1.2. Phonetic changes of language units in Uzbek dialectology (on the example of the Jarkurgan dialect)

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

The dialect of the Jarkurgan district has a phonetically active variable system within the Surkhandarya dialects. In this dialect, phenomena of sound exchange, assimilation, reduction, syncopation, and phonetic variation are regularly encountered. Phonetic variation is considered one of the most active phenomena in the dialect. One language unit can be used in several phonetic forms in relation to the literary language.

For example:

kelayapti → kevotti;

borayapman → borvommam;

qilayapti → qivotti.

In these examples, new pronunciation variants emerged as a result of phonetic reduction and the interaction of sounds. Certain changes are also observed in the vowel system of the Jarkurgan dialect. In some cases, the vowel /a/ in the literary language approaches /o/ or /ə/. Such cases determine the phonetic individuality of the dialect. In the system of consonants, the phenomenon of assimilation occurs actively.

For example:

olib ber → obber;

olib kel → opke;

olib chiq → opchiq.

Here, the coordination of consonants occurs as a result of the articulatory convergence of neighboring sounds. The phenomenon of syncope is also widespread in the dialect.

For example:

qarindosh → qarndosh;

bo‘laman → bo‘man;

kelaman → keman.

In this case, some sounds within the word are dropped, and the pronunciation becomes economical. Another feature of the Jarkurgan dialect is that certain phonetic phenomena differ across villages. For example, in some regions, the form "kevotti" is actively used, while in other regions, the form "kelvotti" may prevail. This indicates the need to be defined in the corpus based on extralinguistic parameters. Thus, phonetic changes in the Jarkurgan dialect are related not only to linguistic but also to regional and social factors, and their annotation in a corpus environment is scientifically important.

1.3. Experience of phonetic tagging of dialectal units in world dialectal corpora

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

In the 21st century, many dialectal corpora have been created as a result of the integration of dialectological research with corpus technologies. In these corpora, phonetic annotation was formed as one of the main layers. In dialectal corpora of the English language, phonetic annotation is often performed based on the IPA (International Phonetic Alphabet). The standard form and the real pronunciation form of each dialectal unit are stored in separate layers. This allows the user to search and compare phonetic variants. In German dialect corpora, audio recordings, transcriptions, and phonetic tags are integrated into a single platform. As a result, the user can analyze pronunciation patterns along with the text. In Russian dialectal corpora, phonetic phenomena are encoded based on a specific tag system. For example, the phenomena of assimilation, reduction, metathesis, and syncopation are marked with separate tags. This allows for the automation of statistical analysis.

World experience shows that phonetic annotation is carried out in the following stages:

Collection of dialectal materials.

Transcribe audio recordings.

Identify the standard shape.

Describe the phonetic phenomenon.

Tag attachment.

Placement in the database.

Search and create statistical modules.

These experiments serve as a methodological basis for creating the Jarkurgan dialect corpus.

CHAPTER 2. LINGUISTIC AND SOFTWARE SUPPORT FOR PHONETIC ANNOTATION OF JARKURGAN DIALECTS IN THE UZBEK DIALECTAL CORPUS

2.1. Representation of Jarkurgan dialects in the Uzbek dialect corpus based on extralinguistic parameters and the operating principles of the search manager

The Jarkurgan dialect corpus is organized in the form of a multi-layered database. To increase the scientific value of the corpus, extralinguistic parameters are included alongside linguistic data.

The following main tables are formed in the corpus structure:

Table 1. Informant database

Informant ID, age, gender, education, occupation, region of residence.

Table 2. Text database

Text ID, recording date, genre, size, theme.

Table 3. Token database

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

Token ID, word form, lemma, position in text.

Table 4. Phonetic annotation database

Token ID, phonetic event type, annotation code.

Table 5. Statistical database

Frequency of phonetic phenomena, territorial distribution, and indicators by age groups. As a result, the corpus operates based on at least five interconnected schedules. Search Manager provides the user with the following capabilities:

search by phonetic tag;

search by lemma;

territorial search;

search by age group;

gender search;

search by phonetic phenomenon type.

This search system serves to scientifically determine the frequency of phonetic variants, study intergenerational differences, and create dialect maps.

2.2. Linguistic support for the phonetic annotation of Jarkurgan dialects in the Uzbek dialectal corpus

Annotation criteria include:

determination of the literary variant;

determination of the dialect variant;

determination of the type of phonetic change;

recording the change position;

Attach UPOS and XPOS tags;

transfer to the statistical database.

This system allows for the recording of phonetic phenomena in a standardized form.

2.3. Software for phonetic annotation of Jarkurgan dialects in the Uzbek dialectal corpus

The software of the Jarkurgan dialect corpus is developed on the Google Colab platform.

The software architecture consists of the following modules:

Data loading module

Audio recordings, transcripts, and metadata files are uploaded to the system.

Tokenization Module

Text is automatically divided into word units and assigned identifiers.

Morphological tagging module

The lemma, UPOS, and XPOS tags are automatically determined for each token.

AMERICAN JOURNAL OF MULTIDISCIPLINARY BULLETIN
ISSN: 2996-511X (online) | ResearchBib (IF) = 9.512 IMPACT FACTOR
Volume-4| Issue-6| 2025 Published: |30-06-2026|

Phonetic annotation module

Phonetic phenomena are automatically identified based on a pre-formed phonetic dictionary.

Database module

Annotated data is stored in SQLite or CSV formats.

Search Module

The user can search for information by phonetic tags, region, or lemma.

Statistical module

The system automatically:

number of phonetic phenomena;

interest rates;

distribution by region;

calculates differences by age groups and presents them graphically.

Visualization Module

The results are displayed in the form of charts, tables, and maps.

The main advantage of the proposed software is that it allows for the automatic identification, search, and statistical analysis of phonetic phenomena in the Jarkurgan dialect. As a result, a methodological and technological basis is formed for the creation of one of the first phonetically annotated dialect corpuses in Uzbek dialectology.

CONCLUSION

The dialect of the Jarkurgan district was formed within the southern dialectal zone of the Uzbek language and possesses unique phonetic characteristics. In the course of the study, the existing experience in the study and annotation of phonetic units in Uzbek and world dialectology was analyzed. Phonetic variation, assimilation, syncopation, reduction, and other phonetic phenomena found in the Jarkurgan dialect were identified, and criteria for recording them in the corpus environment were developed. Additionally, a system of extralinguistic parameters for the dialectal corpus, a phonetic tagging model, and an annotation scheme integrated with UPOS and XPOS tags were proposed. A software model based on Google Colab allows for the automatic annotation, search, and statistical analysis of phonetic phenomena. The results of this research serve to widely implement corpus technologies in Uzbek dialectology, digitize dialects, and form a new stage of linguistic research of dialectal materials.

REFERENCES:

1.Under the editorship of **Abdulla Qodiriy**. *Explanatory Dictionary of the Uzbek Language*. In 5 volumes. – Tashkent: Uzbekistan National Encyclopedia Publishing House, 2020.

2.*Uzbek Dialectology*. – Tashkent: O‘qituvchi Publishing House, 1977.

3.**Sh. Shoabdurahmonov, M. Mirzayev et al.** *Modern Uzbek Literary Language*. – Tashkent: O‘qituvchi Publishing House, 1980.

4.**A. Nurmonov.** *History of Uzbek Linguistics*. – Tashkent: National Society of Philosophers of Uzbekistan Publishing House, 2012.

5.**B. O‘rinboyev.** *Issues of the Uzbek Language and Dialectology*. – Tashkent: Fan Publishing House, 2004.

6.**Tony McEnery and Andrew Hardie.** *Corpus Linguistics: Method, Theory and Practice*. – Cambridge: Cambridge University Press, 2012.

7.**Geoffrey Leech.** *Introducing Corpus Annotation*. – London: Routledge, 2005.

8.**Douglas Biber, Susan Conrad, and Randi Reppen.** *Corpus Linguistics: Investigating Language Structure and Use*. – Cambridge: Cambridge University Press, 1998.

9.**Steven Bird, Ewan Klein, and Edward Loper.** *Natural Language Processing with Python*. – Sebastopol, CA: O’Reilly Media, 2009.

10.**Academy of Sciences of the Republic of Uzbekistan.** *Collection of Scientific Articles on Uzbek Dialectology and Dialectography*. – Tashkent, various years.