pISSN:2349-5707 Volume:13,Issue 2, February -2026 eISSN:2349-5715

ISSUES OF CREATING LINGUISTIC SUPPORT FOR THE UZBEKISTAN ELECTRONIC CORPS OF THE DIALECTS OF THE FERGANA REGION

4Teacher of Fergana State University
azizbekvosiljonov@gmail.com

Abstract

This article discusses the issues of creating a dialectal linguistic resource for the electronic corpus of the Uzbek language based on the phonetic, lexical and morphological characteristics of the dialects of the Fergana region. The study analyzes the theory of corpus linguistics, the principles of constructing a dialectal corpus, linguogeographic classification, areal localization, and the stages of entering written and audio texts into the corpus database. The principles of classifying dialectal units, forming a metadata database, and creating a search interface are substantiated using the example of the dialects of the Bogdod and Buvayda districts. The results of the study will serve to enrich the national corpus of the Uzbek language, develop research in the field of dialectology and computer linguistics, and create an important linguistic resource for speech processing systems based on artificial intelligence.

Keywords

corpus linguistics, dialectal corpus, Fergana dialects, linguistic support, linguogeography, areal localization, metadata, transcription, national corpus of the Uzbek language.

Introduction. In world linguistics, corpus linguistics has emerged as an independent scientific direction in recent decades, strengthening the empirical basis for language research. The issues of creating national corpora, improving existing corpora, and creating special (subject, dialectal) corpora have become one of the urgent tasks.

In Uzbek linguistics, the need to create electronic resources, systematize language units in a digital environment, and expand the capabilities of automatic analysis and search has also increased. In particular, the insufficient reflection of the dialectal layer in the national corpus of the Uzbek language requires special research in this area.

The dialects of the Fergana region occupy a special place in Uzbek dialectology. The phonetic and lexical features formed as a result of the historical, ethnic, and social factors of the region are important empirical material for the corpus. Therefore, this article will cover the theoretical and practical aspects of creating the linguistic base of the dialects of the Fergana region.

Corpus linguistics and theoretical foundations of dialect corpus creation

Corpus linguistics is a field that deals with the creation of a language corpus, its annotation, placement in a search engine, and linguistic analysis. A language corpus is a set of texts selected on the basis of certain principles, stored in electronic form, and processed using special software tools.

A dialect corpus is a component of a national corpus, which reflects regional variants, oral speech samples, and phonetic and lexical units specific to the dialect. The following principles are important in creating a dialect corpus:

1. Representativeness - sufficient coverage of regional variants.
2. Annotation - indication of phonetic, morphological, and lexical features.
3. Metadata system - inclusion of indicators such as the informant's age, gender, region of residence, and type of speech.
4. Searchability - creation of a user-friendly interface.

In world experience, dialectal corpora serve as an important source for preserving the historical and territorial layer of the language and improving automatic speech recognition systems.

Linguogeographic classification of dialects of the Fergana region

pISSN:2349-5707 Volume:13,Issue 2, February -2026 eISSN:2349-5715

The territory of the Fergana region constitutes a complex linguogeographic area. Dialects belonging to the Karluk-Chigil-Uyghur dialect predominate in the region, but as a result of contact with neighboring regions, some phonetic and lexical peculiarities have emerged.

The dialects of the Baghdad and Buvayda districts provide interesting material in terms of areal localization. According to the results of the study:

• Lengthening and shortening of some vowels;
• Consonant alternation at the beginning of words;
• Active use of local lexical units;
• Variants of morphological forms are observed.

As a result of the linguogeographic analysis, common and distinctive features were identified, which were displayed in the corpus through a separate tagging system.

Stages of entering dialectal texts into the corpus

The creation of a dialectal linguistic database was carried out in several consecutive stages:

1. Material collection

• Oral speech samples (audio recordings);
• Written sources (local texts, folklore samples).

2. Transcription

Audio texts were transcribed based on the phonetic principle. Sound changes specific to the dialect were indicated by special symbols.

3. Annotation

Texts were marked at the following levels:

• Phonetic;
• Morphological;
• Lexical;
• Dialectal tags (region name, informant information).

4. Formation of a metadata database

Each text was assigned indicators such as the informant's age, gender, profession, and place of residence.

5. Creation of a search interface

A filter search system was developed to quickly identify dialectal units. This system allows you to search for a specific phonetic or lexical unit within a region.

Linguistic and practical significance

The linguostatistical analysis conducted on the basis of the dialect database made it possible to determine the frequency and functional load of dialect-specific units. The results showed that local lexical units have high activity in certain semantic areas.

This database:

• Enriches the national corpus of the Uzbek language;
• Creates an empirical basis for dialectological research;
• Serves as a resource for automatic speech recognition and synthesis systems;
• Can be used as practical material in the educational process.

Conclusion

Creating the linguistic support of the Fergana regional dialects allows enriching the national corpus of the Uzbek language with a territorial layer. The results of the study served to develop a methodology for systematically collecting, annotating and placing dialectal units in a search system.

Areal analysis conducted on the example of the Baghdad and Buvayda dialects made it possible to identify regional differences and reflect them in the corpus. The dialectal linguistic base is important not only theoretically, but also practically, and serves as a necessary resource for artificial intelligence technologies.

In the future, press The gradual inclusion of these regional dialects in the corpus will lead to the creation of a complete dialectal map of the Uzbek language.

pISSN:2349-5707 Volume:13,Issue 2, February -2026 eISSN:2349-5715

References:

1. Abdurahmonova N.Z. O‘zbek kompyuter lingvistikasi asoslari. – Toshkent, 2020.
2. Po‘latov A. Kompyuter lingvistikasi va uning istiqbollari. – Toshkent, 2019.
3. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge University Press, 2012.
4. Sinclair J. Corpus, Concordance, Collocation. – Oxford University Press, 1991.
5. Kennedy G. An Introduction to Corpus Linguistics. – London, 1998.
6. Захаров В.П. Корпусная лингвистика. – Москва, 2005.
7. Plungian V.A. Vvedenie v korpusnuyu lingvistiku. – Moskva, 2009.
8. O‘zbek tili milliy korpusi: www.uzbekcorpus.uz