Sarcouncil Journal of Arts and Literature

ISSN(Online): 2945-364X

Volume- 04| Issue- 04| 2025

Review Article **Received:** 10-07-2025 | **Accepted:** 30-07-2025 | **Published:** 26-08-2025

Concept of the Dialectical Corpus

Muyassar Kholova
PhD in Philology, Associate Professor, Termez State University

Abstract: This article provides a brief overview of the role and development stages of dialectology in global linguistics, along with information on electronic texts, online resources, and cloud-based databases. It also highlights research carried out within the framework of linguistic corpora, as well as upcoming studies aimed at addressing the challenges facing the subcorpus of Uzbek national dialects.

Keywords: Turkology, Devonu lug'otit turk, "globalizing" future generations, electronic text corpora, Brown University's language corpus (USA), the explanatory dictionary of the Uzbek language, literary language, dialectological corpus schools, phonetic and morphological distinctiveness of word forms.

INTRODUCTION

In global linguistics, dialectology has its own rich history and distinct stages of development. The study of dialects within Turkology traces its origins to the 11th century, beginning with Mahmud Qoshg'ariy's *Devonu lug'otit turk*, approached from the perspective of contrastive linguistics. Interest in the living speech of the people — and in regional language research more broadly — emerged in Spain, while dialectology as a recognized scientific discipline took shape in Germany toward the end of the 17th century.

Over the course of its historical development, the Uzbek people have experienced highly complex and diverse stages, all while maintaining social, political, and cultural ties with other nations. These interactions have enriched the Uzbek language at the phonetic, lexical, and grammatical levels. The vocabulary of Uzbek — shaped over many centuries — is a shared cultural resource, widely understood across the population, frequently used in socio-economic life, and capable of conveying precise meanings. It forms a solid foundation for word formation and grammatical expression, boasting a wealth of words, word forms, and phraseological units.

The re-examination of dialects after a certain period is a natural and necessary process. In our rapidly developing era, technological advancement and the growth of cultural and everyday life are gradually displacing dialects from the speech of "globalizing" future generations, replacing them with the literary language and foreign words entering through it. Dialectal forms are increasingly preserved only in the speech of the older generation. Considering this, research into dialects remains crucial. Although numerous books, monographs, and articles have been published on the phonetics, lexicon, and grammar of Uzbek dialects, the study of their dialectal vocabulary should still be at the forefront of linguistic research today.

In the modern world, it is difficult to imagine the development of any sphere of socio-economic life without computer technologies, and linguistics is no exception. The field is steadily moving toward solving its challenges through the use of electronic text corpora and databases built with the help of these technologies. Below, we will explore the structure and function of text databases and corpora.

An electronic text corpus is a collection of texts brought into a standardized format and organized into a structured system. From a linguistic perspective, such corpora are invaluable for comparing all variants of a language across social, regional, functional, and historical dimensions. They serve as an indispensable tool for information retrieval and linguistic research (Zakharov, V. P. 2005).

Online Database

An online database, unlike a database stored locally on a network or on the computer's own storage (e.g., a CD), is hosted on a website and accessed through the internet. These databases are available as software-as-a-service products, accessible via a web browser. Such databases may be free to use or require a subscription fee, such as a monthly payment. Some also offer advanced features, including collaborative editing and email notifications.

Cloud Database

A cloud database is one that operates via the internet rather than being stored locally. Instead of keeping the database in a single physical location, an organization can host it online so that all

***Corresponding Author:** Muyassar Kholova

DOI- https://doi.org/10.5281/zenodo.16953093

departments can access and update it in real time. Most database service providers offer web-based consoles (graphical user interfaces) that allow end users to manage, configure, and customize database instances (2).

Dialectal Database

A dialectal database is an online, categorized collection of lexical units from the speech of smaller regional or ethnic communities within a nation that otherwise shares a common literary language. Such a database provides opportunities for online access, search, and filtering. Unlike literary language corpora, dialectal databases present entries in phonetic transcription.

Thematic Dialectal Corpus

A thematic dialectal corpus is a specialized collection of dialectal speech material from small regional or ethnic communities within a national language, reflecting their lifestyle and unique cultural mentality. The corpus may include poetic, prose, or mixed poetic-prose texts, organized according to thematic, genre, and metadata-based (passportized) classifications. (Look up Tomsk corpus)(3).

Annotated Dialectal Corpus

This refers to a corpus of phonetic, lexical, and grammatical units belonging to the dialect (language) of a small regional or ethnic community within a national language, enriched with meta-annotations and designed for automatic analysis using artificial intelligence tools.

The first research into linguistic corpora began in the 1960s. In 1963, at Brown University in the United States, the first large-scale text corpus — the *Brown Corpus* — was created and stored on a mainframe computer. Compiled by W. Francis and H. Kučera, it consisted of 500 prose texts of 2,000 words each, representing 50 genres in American English.

In global linguistics, corpus creation initiatives have emerged in Austria, China (Mandarin dialect corpora), Germany, Portugal, the Czech Republic, Finland, Scandinavia, Poland, Lithuania, Georgia (e.g., *Cor-Dial-Sin*, *Helsinki Corpus of English Dialects*, *Norwegian Dialect Corpus*, *Archiv für gesprochenes Deutsch*, *Die Bayerische Dialektadatenbank*, *lexdialgram*, and others). Significant dialect corpus projects have also been developed in central cities of Russia such as Moscow, Slavyansk-on-Kuban, Vologda, Saratov, and Kazan, including Диалектный подкорпус в составе Национального корпуса русского языка (Dialect Subcorpus within the National Corpus of the Russian Language), Саратовский диалектологический корпус (Saratov Dialectological Corpus), Электронная библиотека русских народных говоров (Electronic Library of Russian Folk Dialects), and Электронный корпус диалектной культуры Кубани **(**Electronic Corpus of the Dialect Culture of Kuban — Tregubova E.N., Sfin'ko O.S., Balatsenko N.S., Litus E.V.).

In Russian linguistics, the development of the national corpus — Национальный корпус русского языка (National Corpus of the Russian Language) — along with research carried out at institutions in Saint Petersburg and Saratov, as well as the Tomsk dialectological corpus schools, further illustrates the breadth of these efforts.

In Uzbek linguistics, research has been conducted in the fields of computational linguistics, lexicographic text processing, and linguostatistical analysis by scholars such as A. Po'latov, M. Ayimbetov, S. Muhammedova, S. Karimov, A. Babanarov, D. O'rinboyeva, N. Abdurahmonova, and A. Norov. Among these, Hamroyeva (Eshmuminov, A. 2019) defended a PhD dissertation on *"The Linguistic Foundations of Constructing an Uzbek Authorial Corpus"*, A. Eshmuminov(Begmatova, G. Kh. 2019) worked on *"The Synonym Database of the Uzbek National Corpus"* (PhD dissertation), and G. Begmatova researched an idiom database for the Uzbek National Corpus (TerSU) (Karimov, R., Mengliev, B. 2020). Alongside this, investigations have also been carried out into the development of the Uzbek-English parallel corpus toolkit (Abdurakhmanova, N. 2022), computer models of the Uzbek electronic corpus(Akhmedova, D. 2020), the linguistic foundations and models of lexico-semantic tagging for Uzbek corpus units(Khamroeva, Sh. M. 2021), the linguistic provision of an Uzbek morphological analyzer(Kholiyorov, O. 2021), the linguistic foundations of building an educational corpus of Uzbek(Gulyamova, Sh. 2022), the linguistic foundations of an Uzbek semantic analyzer(Gulyamova, N. 2023), the creation of an Alisher Navoi authorial corpus and its semantic tag base(Boysarieva, S. 2024), as well as the linguistic foundations for compiling an electronic morpheme dictionary(Kholova, M. A. 2020). These studies can be regarded as bold steps in the practical development of corpus linguistics in Uzbekistan.

Publisher: SARC Publisher

Although efforts to establish Uzbek as a language of information technology have been somewhat limited during the years of independence, raising Uzbek to the status of an Internet language remains essential for ensuring its long-term vitality. Between 2018 and 2022, research has been carried out on expanding the database of the Uzbek National Corpus. During this period, the process of unifying dialectal varieties under a single corpus has entered a stage of active development. This process has now become one of the most urgent and pressing challenges facing Uzbek linguistics. At present, the following can be identified as tasks still awaiting resolution:

- To study, analyze, and investigate the principles of constructing existing dialectal databases within the framework of world corpus linguistics.
- To examine international standards for building dialect databases and analyze the linguistic foundations of Uzbek dialects (25).
- To apply the positive experiences gained from studying global dialectal corpora, taking into account the agglutinative nature of the Uzbek language.
- To explore the description and interpretation of dialects presented in explanatory dictionaries of the Uzbek language during the process of developing a dialect database (26).
- To ensure comprehensive coverage of the dialectal variants found in explanatory dictionaries of the Uzbek language within the dialect database and corpus, and to analyze their relation to the literary language(Abdulhakimovna, K. M. 2020).
- To investigate the process of creating electronic texts and files by digitizing dialectal units, words, sentences, and manuscripts.
- To develop unified transcriptional symbols with unique Unicode standards for use in the dialect database and corpus(Kholova, M. 2020).
- To design a system that enables the use of transliteration (a Latin-based representation of dialect pronunciation) within the dialect database and corpus (27).
- To establish principles for the passportization (metadata documentation) of dialect corpora (Kholova, M. A. 2023).
- To develop thematic metadata for the dialect corpus that reflects the lifestyle of ethnic communities.
- To create a system of dialectological atlases within the dialect corpus that function on the basis of variable linguistic data (28).
- To provide theoretical foundations for the development of unified symbolic markers for phonetic (23), morphological, and lexical-grammatical tagging of texts in Uzbek dialects.
- To classify, describe, compare, analyze, and study national words and word forms that share features with literary works, while ensuring compliance with orthographic norms.

CONCLUSION

The creation of a system capable of demonstrating the literary form of nationwide, dialect-specific, and historical-dialectal words and word forms—as well as the commonalities and differences within dialectal units—serves as one of the key factors ensuring the completeness of a dialectal corpus (24).

As a result, it becomes possible to study existing dialects not only from a linguistic perspective but also from a practical one, thereby expanding the opportunities to preserve dialectal units and pass them on from generation to generation. The importance of studying general principles for the linguistic identification of dialectal units within national language corpora, developing meta-annotation of dialectal corpora by genre, and improving linguistic support systems for dialectal transcription is steadily increasing. At present, issues remain urgent regarding the formation of two main types of dialectal corpora based on the transcription (recording) of dialectal units: thematic corpora and tagged text corpora.

REFERENCE

1. Zakharov, V. P. “Corpus linguistics: Educational method. Allowance”. – St. Petersburg, (2005). ¬ P. 48; C.4
2. https://uz.wikidea.ru/wiki/Online_database
3. https://losl.tsu.ru/?q=corpus
4. Lai, H. L. "The nccu corpus of spoken chinese: Mandarin, hakka, and southern min." *Taiwan Journal of Linguistics* 6 (2008): 119-144.
5. Hall, K. P. “Purely practical revolutionaries: A history of Stalinist theoretical physics”. *Harvard University*, (1999)
6. http://rusling.narod.ru/qqq_corp_nonslav_other.htm//
7. Tregubova, E. N., Sfinko, O. S., Balatsenko, N. S., Litus, E. V. “Dialectal culture of Kuban in light of ethnolinguistic analysis (according to the electronic corpus of the dialect culture

Publisher: SARC Publisher

of Kuban), monograph," (2017). http://www.dialog-21.ru/media/4276/zemicheva.pdf
8. Khamroeva, S. h. "Linguistic foundations for creating an author's corpus of the Uzbek language." *Doctor of Philosophy (PhD) ... diss. Author - Bukhara*.
9. Eshmuminov, A. "Base of synonymous words of the national corpus of the Uzbek language." *Doctor of Philosophy (PhD) ... diss. Author - Karshi*. (2019).
10. Begmatova, G. Kh. "Creation of a base of idioms in the national corpus of the Uzbek language" *Doctor of Philosophy (PhD) ... diss. Author - Termez*, (2019).
11. Karimov, R., Mengliev, B. "Theoretical foundations for creating an Uzbek-English parallel corpus / Journal of critical reviews". 7. 17. (2020): 907-911;
12. Abdurakhmanova, N. "Computer models of the electronic corpus of the Uzbek language." *Phil. Sci. Doctor of Philosophy… Diss. Author – Fergana*, (2022).
13. Akhmedova, D. "Linguistic foundations and models of lexical-semantic tagging of units of the Uzbek language corpora": *Phil. Sci. Doctor of Philosophy… Diss. Author – Bukhara*, (2020).
14. Khamroeva, Sh. M. "Linguistic support for the morphological analyzer of the Uzbek language." *Phil. Sci. Doctor of Philosophy… Diss. Author – Tashkent*, (2021).
15. Kholiyorov, O. "Linguistic foundations for the creation of an educational corpus of the Uzbek language." *Phil. Sci. Doctor of Philosophy (PhD)… Diss. Author. – Termez*, (2021).
16. Gulyamova, S. h. "Linguistic foundations of the semantic analyzer of the Uzbek language". *Phil. sciences. Doctor of Philosophy (Doc.)… dissertation. Author. – Fergana*, (2022).
17. Gulyamova, N. "Creation of a database of the author's corpus of Alisher Navoi and its semantic tags (cabinet "Badoe ul-vasat")." *Phil. sciences. Doctor of Philosophy (PhD)… dissertation. Author. – Tashkent*, (2023).
18. Boysarieva, S. "Linguistic foundations of creating an electronic morphemic dictionary" *Phil. sciences. Doctor of Philosophy (PhD)… diss. Author. – Tashkent*, (2024).
19. Kholova, M. A. "The relationship of Uzbek dialects to the standard language and their place in the language system (on the example of the "j"-speaking dialects of the Boysun district) //International Journal of the Art of the Word". – (2020). – T. 3. – No. 5.
20. Abdulhakimovna, K. M. "Transcription in the corpus of the uzbek national boysun dialect (on the example of Baysun district" j" dialects)." *Journal of Critical Reviews* 7.5 (2020): 844-847.
21. Kholova, M. "Peech genres of texsts and passport mete-metrics in the dialect corpus (Baysun district on the example of the sound "J" dialects)." *Journal of Advanced Research in Dynamical and Control Systems* 12.6 (2020): 1103-1107.
22. Kholova, M. A. "Glance at World Dialectal Corpora // "Uzbek National Educational Buildings: Theoretical and Practical Issues of Creation" *International Scientific-Practical Conference*. 2. 2. (2023).
23. Kholova, M. "A Study of the Corpus of Uzbek National Dialects // Catalogue of Monographs.". 1. 1. (2024) 3–124.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Kholova, M. " Concept of the Dialectical Corpus." *Sarcouncil Journal of Arts and Literature* 4.4 (2025): pp 53-56.

Publisher: SARC Publisher