ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23
**American Academic publishers, volume 05, issue 11,2025**

Journal: https://www.academicpublishers.org/journals/index.php/ijai

# UZBEK CORPUS LINGUISTICS: AN ANALYSIS OF PRACTICAL RESULTS AND SCIENTIFIC ACHIEVEMENTS

**Usmonova Mokhinur Ulugbek kizi**
Bukhara State Pedagogical Institute
Lecturer at the Department of Uzbek Language and Literature,
Independent Researcher, Department of Uzbek Linguistics and Journalism,
Bukhara State University
mohinurusmonova872@gmail.com

**Annotation:** The article discusses the formation and development of Uzbek corpus linguistics, which began with theoretical research in the 2010s and progressed with the creation of the educational corpus of the Uzbek language in 2021. It presents reflections on the establishment and advancement of this field. The article also outlines the main concepts and principles of Uzbek corpus linguistics, the types and typology of corpora, and provides information about the formation of the Uzbek National Corpus.

**Keywords:** Corpus linguistics, computational linguistics, corpus, national corpus, authorship corpus, text processing.

**Аннотация:** В статье рассматривается становление и развитие узбекской корпусной лингвистики, начиная с теоретических исследований 2010-х годов и продвигаясь к созданию образовательного корпуса узбекского языка в 2021 году. Обсуждаются основные понятия и принципы корпусной лингвистики, типы и типология корпусов, а также приводится информация, связанная с формированием Узбекского национального корпуса.

**Ключевые слова:** Корпусная лингвистика, компьютерная лингвистика, корпус, национальный корпус, авторский корпус, обработка текстов.

## Introduction

In recent years, theoretical research in Uzbek computer linguistics has focused on automatic translation, modeling of lexical units, studying issues related to corpus linguistics, and theoretically substantiating principles for creating linguistic tools. As a result of these studies, various applications have been developed, including natural language processing tools, text editing and analysis programs for Uzbek texts, speech synthesizers, educational corpora of the Uzbek language, and several types of modern electronic dictionaries.

Today, creating morphological analyzers, parsers, and semantic analyzers for Uzbek language corpora, as well as developing specialized sub-corpora (subsets) within the National Corpus of Uzbek, represent critical tasks for corpus linguists. Furthermore, preserving the purity of the state language, enriching it, raising the speech culture of the population, and actively integrating the state language into modern information technologies and communication systems remain priorities. A corpus is a collection of texts structured according to a search program that stores natural language in electronic form, either written or spoken, organized in a computer-based searchable system, which can operate online or offline. Language corpora are indispensable tools for linguistic research and practical applications. Unlike a typical electronic library, which aims to cover literary and journalistic works reflecting social, political, spiritual, and economic life,

ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23

**American Academic publishers, volume 05, issue 11,2025**

Journal: https://www.academicpublishers.org/journals/index.php/ijai

electronic libraries do not process texts from a linguistic perspective, limiting their usefulness for research. Electronic libraries are intended primarily to preserve cultural heritage rather than serve as scientific research databases.

**Methods**

In contrast, a language corpus aims to gather texts that are necessary, useful, and interesting for studying and researching the language. What distinguishes a corpus from an electronic library is that texts in a corpus are enriched with additional linguistic information and annotations, which form a unique part of the corpus — the corpus metadata. While a regular text editor can find any word a user needs, working with a system that "understands" the meaning, content, and structure of language phenomena in the text is much more effective and convenient. Searching for language units with such software, i.e., corpus tools, can greatly assist researchers and users. Tasks that once took years—like finding examples for research and compiling them manually—can now be completed in minutes with the help of global language corpora. Special search systems composed of several programs provide statistical reports and present search results in a user-friendly format. To get an accurate picture of language processes, it is important to expand the corpus coverage not only to written texts but also to spoken language materials. Such corpora allow researchers to draw precise conclusions about language development and expected changes.

**Results**

In the April 26, 2018 issue of the "Marifat" newspaper, Professors Bakhtiyor Mengliyev and Shahlo Hamroyeva, Doctors of Philology, published a concept paper on the requirements, needs, capabilities, and challenges of the Uzbek National Corpus ("The National Corpus of the Uzbek Language"). This article launched a series of monographic studies and practical developments in Uzbek corpus linguistics. The National Corpus is regarded as a national linguistic treasury. It is widely used by linguists, lexicographers, computational linguists, programmers, editors, translators, journalists, publishers, scientists, teachers, learners, and specialists across various fields. The fact that the Uzbek language, with its significant number of speakers worldwide, needs to have its own content on the internet and that these materials should be used for linguistic processing demonstrates the urgent need to develop Uzbek corpus linguistics. The Uzbek National Corpus, as a language resource, opens doors to modern information technologies, enabling examination of Uzbek natural language from multiple perspectives and integrating the Uzbek language with contemporary information technologies. This has created a need for theoretical studies on the development of Uzbek language corpora.

Recent dissertations on topics such as "Linguistic Foundations of Creating an Authorship Corpus of Uzbek" (Sh. Hamroyeva, 2018), "Synonym Word Database of the Uzbek National Corpus" (A. Eshmuminov, 2019), "Linguistic Foundations and Models for Lexical-Semantic Tagging of Attributive Units in Uzbek Corpora" (D. Akhmedova, 2020), and several others reflect the academic depth and progress in this field. More than ten doctoral dissertations have been defended on corpus linguistics topics, further advancing Uzbek corpus studies.

**Discussion**

The formation of the CorpLing center in Uzbekistan began theoretical research in the 2010s and expanded with the creation of the Uzbek educational corpus in 2021. Theoretical and applied corpus linguistics research in Uzbek linguistics has intensified, and its practical applications have been made available to users. Uzbek computer linguistics is a young but rapidly developing field rooted in corpus linguistics. Prominent scholars such as A. Polatov, A. Rahimov, B. Mengliyev, Sh. Hamroyeva, M. Abjalova, A. Eshmuminov, D. Akhmedova, O. Kholiyorov, N.

ISSN: 2692-5206, Impact Factor: 12,23
**American Academic publishers, volume 05, issue 11,2025**
Journal: https://www.academicpublishers.org/journals/index.php/ijai

Abdurahmonova, G. Toirova, and others have studied Uzbek corpora and their types extensively. Their research covers linguistic foundations for creating authorship corpora, natural language processing, synonym databases, and more.

Currently, master's theses and doctoral dissertations focus on developing national and other types of language corpora, contributing significantly to the advancement of Uzbek computational and corpus linguistics.

In general, Corpus Linguistics is considered a science of the 21st century. Today, it is recognized worldwide as the "body" of linguistics, while computational linguistics is regarded as its "engine."

**References:**

1. Khayriev, U., Usmonova, M., & Turdieva, G. (2024, March). An optimal quadrature formula in the space W~ 2 (2, 1) of periodic complex-valued functions. In AIP Conference Proceedings (Vol. 3004, No. 1). AIP Publishing.
2. Abdurahmonova, N. Corpus Linguistics (textbook). Tashkent: GlobeEdit, 2023. – 357 p.
3. Hamroeva, Sh. Linguistic Foundations for Creating the Uzbek Language Authorship Corpus. Monograph. Germany: GlobeEdit, 2020. – 250 p.
4. The National Corpus of the Uzbek Language – An Important Cultural Phenomenon (roundtable discussion) / "Ma'rifat", 11.08.2021. No. 32