The OSCAR PUBLISHING Services
American Journal of Philological Sciences

VOLUME Vol.05 Issue08 2025
PAGE NO. 96-97
DOI 10.37547/ajps/Volume05Issue08-25

Problems And Solutions in Creating The National Corpus of The Uzbek Language

Eshqulova Nilufar Uktamovna
Independent Researcher, Uzbekistan

Received: 30 June 2025; **Accepted:** 29 July 2025; **Published:** 31 August 2025

Abstract: This article analyzes the challenges in creating the national corpus of the Uzbek language and explores possible scientific and practical solutions. The importance of the national corpus, its integration with modern digital technologies, and comparisons with international experience are discussed. In addition, the prospects of enriching, standardizing, and applying the national corpus in the educational process are considered.

Keywords: Uzbek language, national corpus, linguistics, digital technologies, artificial intelligence, machine translation, electronic resources, text database, standardization, philology, lexicography, automation, computational linguistics, education, development.

Introduction: In today's era of globalization and digital technologies, new approaches are also required in the field of linguistics. National corpora serve as an important tool in studying a language on a scientific basis, promoting it internationally, and ensuring its broad application through modern technologies. In recent years, the issue of creating a national corpus of the Uzbek language has become urgent, bearing not only linguistic but also social and cultural significance. In the process of forming a national corpus, it is necessary to digitize texts, analyze them morphologically, syntactically, and semantically, standardize them, and take into account international experiences. This article provides a detailed overview of the problems encountered in this process and their scientific and practical solutions.

MAIN PART

In creating the national corpus of the Uzbek language, the first step is to form the source base. This includes literary works, scientific articles, textbooks, press materials, as well as samples of oral speech. Converting texts into digital format, marking them linguistically, and conducting morphological and syntactic analysis is a complex and time-consuming process. In addition, the dialectal richness of the Uzbek language, as well as orthographic and stylistic differences, create additional challenges in corpus development. To address these issues, it is advisable to draw on international experience, particularly from the national corpora of English, Russian, and other major languages. In addition, the use of artificial intelligence and machine learning algorithms makes it possible to improve the quality of the corpus and present it to users in a more convenient form. The development of the corpus creates broad opportunities in the education system, in scientific research, in automatic translation systems, and in the popularization of the language.

The Necessity of Creating a National Corpus of the Uzbek Language

The Uzbek language, as the state language, is used in all spheres of social life. However, for the language to be studied comprehensively from a linguistic perspective, applied effectively in education, and promoted internationally, the existence of a national corpus is essential. The corpus makes it possible to determine the frequency of word usage, their function in real speech, and their variants across different stylistic layers. In addition, a national corpus serves as a primary resource in developing modern technologies such as machine translation, artificial intelligence systems, and automatic text analysis. Moreover, corpus materials also play an important role in teaching the Uzbek language within the education system.

Existing Problems in Creating a National Corpus of the

Uzbek Language

In the process of forming the national corpus of the Uzbek language, a number of problems are observed. Firstly, there are technical problems: collecting texts created in different periods, converting them into digital form, encoding, and standardizing them requires significant effort. Secondly, there are linguistic problems: since Uzbek is an agglutinative language, the morphological and syntactic analysis of words is very complex. The third problem is related to theoretical approaches, namely which principles should be taken as the basis in corpus creation, what criteria should be applied for text selection, and how stylistic layers should be classified — all of which have not been sufficiently developed. Fourthly, there are practical problems: a shortage of trained specialists, limited financial resources, and the absence of a unified platform.

Solutions and Prospects

To address the existing problems in creating the national corpus of the Uzbek language, it is necessary to work in several directions. First of all, it is important to strengthen the national digitization policy and adopt the national corpus project as a separate program at the state level. Many technical problems can be solved through the extensive use of artificial intelligence and natural language processing (NLP) technologies, as well as by automating morphological and syntactic analysis. In addition, it is necessary to study international experiences, collaborate with foreign national corpora, and integrate the Uzbek language corpus into international scientific platforms. The use of the national corpus in the educational process, in scientific research, and in electronic dictionaries and translation systems will make a significant contribution to the development of the Uzbek language.

CONCLUSION

In conclusion, the creation of a national corpus of the Uzbek language has not only linguistic but also social and cultural significance. The national corpus serves as an important tool for the development of the language, its promotion on an international scale, and its application through modern technologies. To solve the problems encountered in this process, cooperation between the state, scientific institutions, and IT specialists is necessary. In the future, the national corpus of the Uzbek language is expected to be widely used in artificial intelligence, machine translation, electronic dictionaries, and educational resources.

REFERENCES

Sharipov A. O‘zbek tili milliy korpusi va uning istiqbollari. – Toshkent, 2022.

Karimov B. Lingvistika va raqamli texnologiyalar integratsiyasi. – Toshkent, 2021.

National Corpus of the Russian Language. – Moscow, 2020.

Biber D., Reppen R. Corpus Linguistics. – Cambridge University Press, 2015.

Crystal D. Language and the Internet. – Cambridge University Press, 2006.

McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge University Press, 2012.