pISSN:2349-5707 Volume:13,Issue 4, April -2026 eISSN:2349-5715

CHALLENGES IN THE DIGITALIZATION OF THE UZBEK LANGUAGE: ACHIEVEMENTS, PROBLEMS, AND PROSPECTS

Ibragimov Javlon Yormatovich
Associate professor, doctor of philology University of information technologies and management

Abstract

In recent years, linguistics worldwide has become closely integrated with digital technologies, giving rise to the interdisciplinary field of digital linguistics. The Uzbek language must actively participate in this transformation. Digitization is not limited to converting texts into electronic format; it involves adapting the language to artificial intelligence, machine translation, speech technologies, and corpus-based analysis. While notable achievements have been made in Uzbek – including spell-checkers, online dictionaries, corpus projects, and digital text databases – these efforts remain fragmented, lack standardization, and face significant technical and methodological challenges. The core problems include the absence of a unified national digital platform, inconsistent coding standards, and insufficient integration between linguistic resources. Additional barriers involve gaps in digital competence among linguists and limited linguistic knowledge among developers. This article examines the current state of Uzbek language digitalization, identifies key achievements and obstacles, and proposes pathways for developing a comprehensive national digital linguistic infrastructure.

Keywords: digitalization, corpus, Unicode, ASCII, artificial intelligence (AI), TTS (Text-to-Speech), STT (Speech-to-Text), OCR (Optical Character Recognition), tokenization, standardization, digital linguistics.

Introduction. The rapid advancement of digital technologies has fundamentally transformed linguistic research and language use. What was once viewed primarily as a technical process has evolved into a strategic, cultural, and scientific priority. Digitizing a language means creating the necessary infrastructure for artificial intelligence systems, machine translation, speech recognition and synthesis, and large-scale corpus analysis.

For the Uzbek language, this task has become particularly urgent. A well-developed digital infrastructure is not merely a technical tool but a means of preserving and developing national linguistic identity in the digital environment. However, the current efforts in Uzbekistan remain largely uncoordinated, leading to duplication of work, incompatibility of resources, and slow progress. This article analyzes the achievements and persistent challenges in the digitalization of the Uzbek language and outlines the requirements for building a sustainable national digital linguistic ecosystem.

Methods. The study employs a descriptive-analytical and comparative methodology. It is based on a review of existing digital projects in Uzbek linguistics, including spell-checking systems, online dictionaries, corpus initiatives, and speech technology developments. Institutional documents, project reports, and scholarly literature on digital linguistics and corpus studies were analyzed. Particular attention was given to issues of standardization, resource integration, and human capital development. The analysis also draws on international experience in building national corpora and digital language infrastructures.

Results. Significant progress has been achieved in several areas:

- Development of spell-checking tools and online dictionaries;
- Creation of initial digital text databases and corpus projects;
- Implementation of basic OCR (Optical Character Recognition) systems;
- Ongoing work on speech technologies (TTS and STT).

pISSN:2349-5707 Volume:13,Issue 4, April -2026 eISSN:2349-5715

These initiatives demonstrate growing awareness of the importance of digital linguistics in Uzbekistan. However, most projects operate independently, without a shared technical or methodological framework.

The main obstacles to effective digitalization can be grouped into several categories:

1. Multiple institutions are involved in digitization efforts, including the Institute of Uzbek Language, Literature and Folklore of the Academy of Sciences, universities, IT centers, and private startups. Each operates according to its own concept, resulting in fragmented outcomes. Linguistic annotation standards, for example, vary significantly across projects, making data interoperability difficult.

2. A critical problem is the absence of unified standards for digital encoding and orthography. Variations in the digital representation of characters such as “o‘” and “g‘” (different Unicode code points) create inconsistencies that affect search engines, machine translation, and speech synthesis systems. The parallel use of Cyrillic and Latin scripts further complicates the creation of unified digital resources.

3. Existing digital dictionaries, analyzers, and corpora are not interconnected. Different projects use incompatible data formats (XML, JSON, CSV), requiring additional conversion work. Unlike many international projects that follow open-source and interoperable standards, Uzbek initiatives are often closed and project-based, limiting reusability and long-term sustainability.

4. There is a significant disconnect between linguists and software developers. Most linguists lack sufficient digital competencies, while developers often do not possess deep knowledge of Uzbek linguistic structures. The profession of “language engineer” or computational linguist is not yet formally recognized in Uzbekistan, creating a serious bottleneck in developing sophisticated NLP tools.

5. Most projects are funded through short-term grants. Once funding ends, maintenance, updating, and expansion of resources often cease. In contrast, successful international models (such as the British National Corpus or Turkish digital language projects) benefit from stable, long-term governmental support.

6. The agglutinative structure of Uzbek, complex suffix chains, and specific phonetic features require specialized algorithms. Current machine translation and speech technologies frequently fail to account for these characteristics, resulting in grammatical errors and unnatural synthesized speech.

Discussion. The challenges outlined above are interconnected and reflect deeper systemic issues. The lack of a unified national strategy leads to duplication of effort and inefficient use of resources. While individual projects demonstrate innovation, the absence of coordination prevents the formation of a coherent digital linguistic ecosystem.

From a broader perspective, successful digitalization requires more than technical solutions. It demands a comprehensive national policy that combines linguistic expertise, technological capacity, sustainable funding, and educational reform. The development of open standards, interoperable data formats, and interdisciplinary training programs is essential.

International experience shows that languages with well-developed digital infrastructures benefit from open-access corpora and collaborative platforms. Adopting similar principles in Uzbekistan could significantly accelerate progress and enhance the global visibility of the Uzbek language in digital environments.

Conclusion. The digitalization of the Uzbek language has achieved initial successes but continues to face substantial structural, technical, and organizational challenges. Overcoming these obstacles requires moving beyond isolated projects toward a coordinated national strategy. Key priorities include:

- Establishing unified standards for encoding, annotation, and data exchange;
- Creating an integrated national digital platform;
- Developing interdisciplinary educational programs;

pISSN:2349-5707 Volume:13,Issue 4, April -2026 eISSN:2349-5715

- Ensuring long-term financial and institutional support;
- Promoting open-access principles for linguistic resources.

Only through such a comprehensive approach can the Uzbek language acquire a modern, functional digital infrastructure capable of supporting artificial intelligence, machine translation, and other advanced technologies while preserving its linguistic and cultural identity.

References:

1. Burada M., Tatu O., Sinu R. Language and Communication in the Digital Age. – Cambridge: Cambridge Scholars Publishing, 2023.
2. Fuertes-Olivera P., Bergenholtz H. e-Lexicography. – London: A&C Black, 2011.
3. Go E., Sundar S.S. Humanizing chatbots... // Computers in Human Behavior. 2019.
4. McEnery T. Corpora and Minority Languages. – London: Routledge, 2015.
5. Rehm G. Language Technology for All. – Berlin: Springer, 2022.
6. Stanciu N., Jager S.D. Digital language // A Multilingual Language in Its Multiple Dimensions. 2024.
7. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Санкт-Петербург, 2020.
8. Маник С.А. Цифровая лексикография... // Язык в эпоху цифровых трансформаций. 2024.
9. Шестакова Л.Л., Кулева А.С. Авторская лексикография в электронно-цифровую эпоху // Terra Linguistica. 2023.