



CREATION OF A DIGITAL CORPUS OF UZBEKISTAN DIALECTS AND PHONETIC DESCRIPTION

Kurbanazarova Marvarid Bazar kizi

Termez State University

1st year master's student, Department of Computational Linguistics

marvaridqurbonazarova@gmail.com

Abstract: This article analyzes the theoretical and practical issues of creating a digital corpus of Uzbek dialects and the principles of phonetic characterization of dialect materials. The processes of collecting, digitizing, transcribing and annotating dialect data are highlighted, and the importance of the phonetic corpus in linguistic research is substantiated. The possibilities of identifying and describing phonetic phenomena specific to Uzbek dialects based on corpus technologies are also shown. The results of the research can serve the development of Uzbek dialectology, corpus linguistics and computer linguistics.

Keywords: dialect, dialect, corpus linguistics, phonetic annotation, digital corpus, transcription, Uzbek language, computer linguistics.

Introduction

The rapid development of information technologies has created the basis for the formation of new research areas in linguistics. In particular, corpus linguistics is currently developing as an important scientific field that allows studying language units based on large volumes of electronic texts. The practice of studying national languages and their dialects based on electronic corpora is widely used in world linguistics. Such corpora allow for the systematic study of the phonetic, lexical, morphological and syntactic features of the language.

The Uzbek language has a rich dialectal composition, and its Karluk, Kipchak and Oghuz dialects, as well as many dialects, reflect the historical development, culture and ethnographic features of the people. However, most of the materials on dialects are preserved in the form of printed sources or audio recordings, and their electronic database is not sufficiently formed. This limits the possibilities of systematic study of dialect units and analysis based on modern technologies. Therefore, the creation of a digital corpus and phonetic description of Uzbek dialects is one of the urgent issues of today's linguistics. The digital corpus serves not only as an important resource for the development of dialectological research, but also as a basis for the creation of speech technologies, automatic text analysis, and artificial intelligence systems.

Purpose and objectives of the study

The purpose of the study is to develop the principles of creating a digital corpus of Uzbek dialects and to characterize dialect materials phonetically.

To achieve this goal, the following tasks were set:

- collection of linguistic materials on Uzbek dialects;
- digitization of audio and text data;
- development of a phonetic transcription system;
- phonetic annotation of dialect units;
- description of phonetic phenomena based on corpus data;
- determination of the scientific and practical significance of the dialectal corpus.

Research methodology

The study used descriptive, comparative-historical, corpus analysis, statistical analysis, and phonetic observation methods. Dialect materials were collected on the basis of audio recordings and dialectological sources and transferred to electronic format. Elements of the International Phonetic Alphabet (IPA) and transcription methods used in Uzbek linguistics were used in the phonetic description.

Principles of creating a digital corpus of Uzbek dialects

Creating a dialect corpus consists of several stages:

1. Data collection;
2. Digitization of audio recordings;
3. Transcription of texts;
4. Linguistic annotation;
5. Formation of a corpus database;
6. Creation of a search and analysis system.

The corpus includes dialect samples collected from different regions, materials of folk oral culture, conversations and dialectal texts. This allows for a comparative study of the regional characteristics of dialects.

Phonetic description of Uzbek dialects

Significant differences are observed in Uzbek dialects. In particular, in some dialects, the vowel “o” in the literary language can be pronounced as “u”, and the vowel “a” can be pronounced as long. Also, the occurrence of consonant shifts, assimilation and dissimilation is widespread.

For example:

- the word “bola” in the literary language is pronounced as bala in some dialects;
- the form “kevotti” is pronounced as kevotti;
- the form “boradi” is pronounced as baradi.

The identification of such phonetic variants in the corpus using special tags allows for an in-depth analysis of the phonetic system of dialects.

Results and discussion

As a result of the study, the necessary linguistic and technological foundations for creating an electronic corpus of Uzbek dialects were determined. The phonetic annotation system made it possible to identify sound changes, phonetic variants and regional differences in dialects. This will serve to bring Uzbek dialectology to a new level.

Conclusion

The creation of a digital corpus of Uzbek dialects and their phonetic description are one of the modern directions of linguistics. Dialectal corpora serve as an important tool for preserving, researching, and transmitting dialects to future generations. Phonetic annotation allows for the systematic study of the specific features of dialects. In the future, the integration of artificial intelligence and automatic speech processing technologies into this corpus will make a significant contribution to the development of Uzbek computer linguistics. The results of the study showed that the creation of a digital corpus of Uzbek dialects will not only bring dialectological research to a new level, but also create an important scientific basis for the preservation and systematic study of regional variants of the national language. Describing the phonetic features of dialects based on an electronic corpus allows for the analysis of sound changes, phonetic variants, and pronunciation differences in different regions based on accurate and reliable data.

As a result of the creation of a digital corpus, the processes of collecting, storing, searching and processing data on Uzbek dialects will be significantly improved. In particular, the phonetic annotation system will allow for the automated analysis of the specific phonological features of dialects. This, in turn, will serve the development of corpus linguistics, computer linguistics and speech technologies. During the study, it was found that the digitization and phonetic tagging of dialectal materials create a wide range of opportunities for statistical and comparative study of

similarities and differences between dialects. At the same time, it is possible to deeply study the historical development, territorial differentiation and modern development trends of the Uzbek language based on the corpus data. In the future, it is advisable to integrate the phonetic corpus of Uzbek dialects with artificial intelligence technologies, develop automatic speech recognition and synthesis systems, as well as use this corpus to create dialectal atlases. As a result, a solid scientific and practical platform will be created to preserve the dialect richness of the Uzbek language, introduce it into scientific circulation, and widely promote it within the framework of international linguistic research.

References:

1. Ashirboyev S. Uzbek dialectology. – Tashkent: Fan, 2016. – 256 p.
2. Abdullayev F. Uzbek dialects and their classification. – Tashkent: O'qituvati, 2014. – 198 p.
3. Hojiyev A. Phonetics of the Uzbek language. – Tashkent: O'qituvati, 2006. – 224 p.
4. Nurmonov A., Mahmudov N. Theoretical grammar of the Uzbek language. – Tashkent: O'qituvati, 2015. – 312 p.
5. Mengliyev B. The current Uzbek literary language. – Tashkent: Tafakkur Bo'stoni, 2018. – 408 p.
6. Yuldoshev B. Introduction to linguistics. - Tashkent: Akademnashr, 2017. - 286 p.
7. Jorayev M. Uzbek folk dialects. - Tashkent: Science and Technology, 2013. - 240 p.
8. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. - Cambridge: Cambridge University Press, 2012. - 294 p.
9. McEnery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 2001. – 238 p.