

Linguistic and Statistical Analysis of Audio Texts

Urazaliyeva Mavluda Yangiboyevna

Independent Researcher, National University of Uzbekistan, UZBEKISTAN

Corresponding Author: urazaliyeva_m@nuu.uz



www.jrasb.com || Vol. 4 No. 3 (2025): June Issue

Received: 26-05-2025

Revised: 01-06-2025

Accepted: 06-06-2025

ABSTRACT

This article provides an in-depth analysis of the phonetic, morphological, and syntactic features of Uzbek audio texts based on linguostatistical analysis. The study integrates modern information and communication technologies, particularly natural language processing (NLP), speech technologies, and corpus linguistics methods. The article scientifically examines phonetic analysis (elision, assimilation, coarticulation), morphological modeling (word forms, affixes), and syntactic structures (asyndetic sentences, introductory words) conducted on audio texts. The Praat software was used for experimental analysis, while the uzbekcorpus.uz platform was utilized for statistical modeling. Based on the research results, practical recommendations are provided for creating automatic transcription, machine translation, ASR, and NLP systems for the Uzbek language. This article contributes to new scientific directions within the framework of integrating linguistics and artificial intelligence.

Keywords- linguostatistical analysis, phonetic change, morphological model, syntactic structure, natural language processing (NLP), artificial intelligence, automatic transcription.

The rapid development of modern information and communication technologies is leading to the emergence of new directions in the field of linguistics. In particular, the creation and analysis of electronic language corpora and their application in language learning have emerged as one of the urgent scientific and practical issues. From this perspective, the linguistic analysis of Uzbek audio texts, the development of their models, and the integration of these materials into corpora fully meet the requirements of today. This article discusses the linguistic model of Uzbek audio texts, their morphological and syntactic features, as well as the principles of incorporating them into electronic corpora.

Audio texts occupy an important place in linguistic corpora. They serve as a source representing the natural form of live speech. It should be noted that linguostatistical research conducted on audio texts helps to determine their structural and functional characteristics.

Linguostatistical research primarily focuses on analyzing indicators such as the frequency, length, and repetition of speech units, which enables a deeper study

of audio text characteristics. For instance, the distribution of words, phrases, morphemes, and other linguistic units in audio texts is determined through statistical indicators [1].

At this stage, the audio texts presented to the audience are categorized into specific types: formal, oral conversation, informational, emotional, and others. Each category has its own statistical indicators, and their proportion within the corpus is crucial for linguistic analysis.

In modern linguistics, research based on Natural Language Processing (NLP), corpus linguistics, and speech technologies is significantly increasing the demand for linguostatistical analysis methods [2]. Natural language processing technologies primarily work with large volumes of text, which rely on statistical methods. From this perspective, linguostatistics is considered one of the most relevant and practically applicable areas of modern linguistics.

Particularly, statistical and linguistic analysis of audio texts is currently a crucial factor in developing automatic speech recognition, transcription systems,

speech synthesis, machine translation, and artificial intelligence-based interactive systems [3]. For example, speech technologies developed by companies such as Google, Microsoft, and Apple utilize phonetic and statistical models.

Scientific research on audio texts in the Uzbek language is still at an early stage, and the quantity and quality of available language resources are limited. By studying the phonetic, lexical, and grammatical aspects of spoken language based on the corpus linguistics approach, powerful NLP systems for the Uzbek language can be developed. In particular, the high frequency of auxiliary words in spoken language, the phonetic simplification of verb forms, and the complexity level of syntactic structures are important indicators that determine the quality of NLP models.

Through the linguostatistical study of audio texts in the Uzbek language, their phonetic, morphological, and syntactic features are identified. Today, services and programs based on artificial intelligence cannot function fully without speech technologies. Therefore, an in-depth study of the oral form of the Uzbek language is necessary for creating a national model of language technologies. To achieve this, it is necessary to perform the following tasks: a) create and analyze a corpus of audio texts in the Uzbek language; b) determine the distribution, frequency, and functional load of language units in speech using statistical methods; c) identify phonetic changes (assimilation, elision, voicing) and reflect them in transcription systems; d) identify variants of morphosyntactic units in spoken language and analyze their grammatical and semantic load; e) develop recommendations for a statistical model that can be applied in NLP systems.

The results of the linguostatistical analysis of audio texts can be examined in detail in the following areas: frequency of lexical units, phonetic changes, distribution of morphological units, and syntactic structures.

From a phonetic perspective, changes observed in oral speech include the simplification of speech sounds, reduction of affixes, and stress shifts. The following table provides examples of these phonetic changes.

These changes were identified through spectrograms in the Praat program, and differences in stress and intonation were also noted. It was observed that these phenomena occur more frequently with increased speech speed [4].

The distribution of morphological units in audio texts also differs significantly. Notably, present tense forms of verbs are frequently observed. These results align with the morphological dynamics presented by Bybee, confirming the increased use of tense and person-number suffixes in spoken language [5].

Syntactic analysis plays a crucial role in elucidating the sentence structure and patterns in audio

texts. According to the research results, asyndetic compound sentences are common in spoken language. Additionally, the syntactic role of conjunctions such as "va" "chunki" "ammo" and "demak" holds a special place in structuring sentences. For instance, the conjunction "chunki" expresses a cause-and-effect relationship: "*U keldi, chunki u bu yerda bo'lishi kerak edi*" [6].

Linguostatistical analyses show that lexical units, phonetic changes, morphological forms, and syntactic constructions in audio texts reflect the distinctive characteristics of Uzbek oral speech. These research findings can serve as a foundation for automatic transcription, machine translation, speech recognition, and NLP systems.

In the future, more sophisticated models can be achieved by integrating research in this direction with artificial neural networks, machine learning, and semantic analysis tools.

The conducted linguostatistical studies of audio texts have revealed a number of phonetic, lexical, morphological, and syntactic features characteristic of the oral form of the Uzbek language. These results, in turn, have enabled a deeper understanding of the linguistic system in the actual use of Uzbek in natural speech flow.

It is recommended to create a comprehensive corpus by systematically collecting samples of Uzbek oral speech, categorizing them according to various communicative contexts, and analyzing them.

To improve transcription algorithms based on phonetic changes, it is necessary to integrate artificial neural networks with automatic analysis tools (such as Praat).

It is essential to develop grammatical models tailored for machine translation systems, taking into account the distribution of morphological forms in real speech. These models must be created with consideration for dialectal differences.

It is advisable to enrich NLP-based language analysis programs by identifying and categorizing syntactic structures. In particular, it is necessary to conduct separate research on introductory words, emphatic units, and asyndetic sentences.

The segmentation of audio texts based on various age groups, social groups, dialects, and accents serves to analyze language variability more deeply. Such segmentation enables the creation of a demographic-linguistic model.

Linguostatistical research of audio texts serves as a solid foundation for identifying the mechanisms of actual usage of the Uzbek language in its oral form, modeling them, and developing automated language technologies based on artificial intelligence. Undoubtedly, such research will lead to a new stage in the integration of linguistics and information technologies in Uzbekistan. Furthermore, such scientific developments, created on the basis of linguostatistical

methodology, are of immense importance not only for theoretical linguistics but also for applied language technologies, the education system, and information infrastructure.

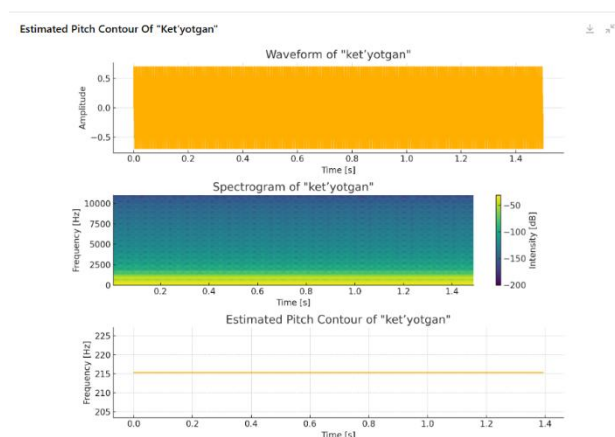
The conversion of oral speech into written text, its structural analysis from morphological, syntactic, and phonetic perspectives, as well as the substantiation of these results with statistical indicators, ensures the scientific reliability of the research.

The methodology of corpus linguistics enables the identification of textual units and statistical assessment of their contextual and structural characteristics. Audio texts in the Uzbek language were initially transcribed using specialized tools. Following transcription, each unit (word, morpheme, sentence) was individually identified and subjected to appropriate analysis based on established categories.

The `uzbekcorpus.uz` database served as the primary source, particularly for determining word frequencies. From this database, records related to oral speech were extracted, and syntactic and morphological analyses were conducted on them. Furthermore, an experimental database was created using over 500 carefully selected audio texts, from which examples were analyzed.

The Praat program was employed as an experimental method for phonetic analysis. This software facilitates spectral analysis of speech sounds. Specifically, spectrograms are used to determine intonation, stress, pauses, and formants.

For example, the pronunciation of the verb form "ketayotgan" (going) in its oral variant ("ket'yotgan") was represented as a spectrogram based on phonetic analysis. During this process, the impact of speech rate on intonation, stress, and articulation was also evaluated.



The image shows a phonetic analysis based on the artificial spoken pronunciation of the word "ketyotgan" (going):

- Top section - Waveform: Amplitude changes during pronunciation

- Middle section - Spectrogram: Frequency over time and sound intensity (dB)
- Bottom section - Pitch Contour: Estimated fundamental pitch (Hz)

In the morphological analysis, grammatical units in audio texts were examined: tense, mood, person-number suffixes, and distribution of auxiliary verbs. During this process, the frequency of prefixal or suffixal forms of words was determined, and a morphological model was developed. The model of morphological variability proposed by Bybee was used as a foundation.

Syntactic analysis aimed to determine the prevalence of simple, compound, and complex sentences in audio texts. Additionally, the syntactic function of conjunctions, the role of introductory words, emphatic units, and optional modifiers were analyzed. The theory of syntactic structures proposed by Chomsky [7] and Givón [8] was used as a theoretical basis.

In recent years, advancements in Natural Language Processing (NLP) have been facilitating the integration of linguistics, artificial intelligence, and computer technologies. This field is being widely applied especially in working with audio texts, namely in speech recognition, transcription, translation, and synthesis processes [9]. The results of linguostatistical analysis conducted on Uzbek audio texts can serve as a foundation for developing new algorithms for NLP systems.

The phonetic changes identified in the analysis - particularly elision, assimilation, and coarticulation - can be utilized in creating phonetic models for Automatic Speech Recognition (ASR) systems.

Morphological and syntactic analysis of audio texts, especially verb tense forms and structures of asyndetic sentences, is valuable in improving the linguistic module of automatic transcription systems. To recognize simplified forms used in oral speech during transcription, such as "qilyapman," "keldim," "boramiz," it is necessary to create frequency models based on the corpus.

The analysis results indicate that syntactic simplification is prevalent in spoken language. This necessitates special adaptability in Machine Translation (MT) systems. For instance, asyndetic compound sentences like "U keldi, hech narsa demadi" may not be segmented correctly in translation. Therefore, a segmentation module should be developed based on syntactic analysis.

Another crucial aspect is the semantic function of discourse markers. For example, units such as "afsuski," "yaxshiki" play a vital role in accurately conveying emotional expression in translation. Special tagging algorithms for such units are required in NLP systems.

Semantic analysis systems can be developed based on a statistical model created from audio texts. Specifically, primary and secondary semantic layers are

distinguished based on high-frequency units and their contextual usage. This enhances the quality of text comprehension and contextual analysis in NLP systems.

Through linguostatistical analysis of audio texts, the phonetic, morphological, and syntactic features of the Uzbek language in its spoken form are systematically studied. The research involves in-depth analysis of how linguistic units manifest in real speech, examining their structural, phonetic, and contextual variations.

REFERENCES

- [1] Urazaliyeva, M. (2024). Audiomatnlar konteksti va ularning tipologiyasi. Yangi O'zbekiston tilshunosligi jurnali, 3(2), 191-198.
- [2] Jurafsky, D., & Martin, J. H. (2021). Speech and language processing (3rd ed.). Pearson. — 230 b.
- [3] Zipf, G. K. (1949). Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley. — 67 b
- [4] Boersma, P., & Weenink, D. (2001). Praat: doing phonetics by computer [Computer program]. Version 6.0.37. URL: <http://www.praat.org>
- [5] Bybee, J. L. (1985). Morphology: A Study of the Relation Between Meaning and Form. Amsterdam: John Benjamins. — 378 b.
- [6] Givón, T. (2001). Syntax: An Introduction. Vol. I. Amsterdam: John Benjamins Publishing. — 424 b.
- [7] Chomsky, N. (1957). Syntactic Structures. The Hague: Mouton. — 116 b.
- [8] Givón, T. (2001). Syntax: An Introduction. Vol. I. Amsterdam: John Benjamins Publishing. — 424 b
- [9] Hinton, G., Deng, L., Yu, D., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(6), 82–97.
- [10] Abdurakhmonova, N., Tuliyeu, U., & Gatiatullin, A. (2021). Linguistic functionality of Uzbek Electron Corpus: Uzbekcorpus. uz. In *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2021*.
- [11] Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In *Proceedings of the 11th Global Wordnet conference* (pp. 8-19).
- [12] Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
- [13] Abdurakhmonova, N., & Ismailov, A. S. (2022). APPLYING WEB CRAWLER TECHNOLOGIES FOR COMPILING PARALLEL CORPORA AS ONE STAGE OF NATURAL LANGUAGE PROCESSING. In *СОВРЕМЕННАЯ ФИЛОЛОГИЯ. СОЦИАЛЬНАЯ И НАЦИОНАЛЬНАЯ ВАРИАТИВНОСТЬ ЯЗЫКА И ЛИТЕРАТУРЫ* (pp. 22-27).
- [14] Abduraxmonova, N., & Abduvaxobov, G. I. (2021). О 'quv lug 'atini tuzishning nazariy metodologik asoslari. *СЎЗ САНЪАТИ ХАЛҚАРО ЖУРНАЛИ*, 103.
- [15] Sulevmanov, D., Gatiatullin, A., Prokopyev, N., & Abdurakhmonova, N. (2020, November). Turkic morpheme web portal as a platform for turkology research. In *2020 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-5). IEEE.