World Bulletin of Education and Learning (WBEL)
ISSN (E): 3072-175X
Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

LINGUISTIC AND SOFTWARE CAPABILITIES OF AUDIO-TEXT CORPORATIONS

Mavluda Yangiboyevna Urazaliyeva
Independent researcher at the National University of Uzbekistan

Abstract
This article discusses the linguistic and technological significance of audio corpora. The role of audio corpora in empirically studying the phonetic, prosodic, morphological, and syntactic features of spoken language is substantiated. Using the example of large audio corpora created in English, Russian, Turkish, and German, their role in research and the development of artificial intelligence systems is analyzed. The importance of annotation, metadata, and standardization issues in creating audio corpora is also shown.

Keywords: Voice corpus, spoken speech, phonetic analysis, annotation, ASR, artificial intelligence, corpus linguistics.

Introduction

Abstract. V state rassmatrivaetsya nauchnaya and prikladnaya znachimost zvukovykh corpusov. Obosnovyvaetsya role audiocorpusov v empiricheskom izuchenii phoneticheskikh, prosodicheskikh, morphologicheskikh i syntakticheskikh osobennostey oral language. Na primere angliyskikh, russkikh, turetskikh i nemetskikh zvukovykh corpusov analiziruetsya ix vklad v lingvisticheskie issledovaniya i razvitie sistem iskusstvennogo intellecta. Osoboe vnimanie udelyaetsya voprosam annotatsii, metadannyx i standartizatsii pri sozdanii audiokorpusov.

Keywords: Sound corpus, oral speech, phonetic analysis, annotation, ASR, artistic intellect, corpus linguistics.

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | **Volume 01, Issue 03, December 2025**

https://worldbulletin.org/index.php/1

Abstract

This article examines the scientific and technological significance of speech corpora. The role of audio corpora in the empirical analysis of phonetic, prosodic, morphological, and syntactic features of spoken language is substantiated. Based on major English, Russian, Turkish, and German speech corpora, their contribution to linguistic research and the development of artificial intelligence systems is analyzed. Special attention is paid to annotation, metadata, and standardization issues in speech corpus construction.

Keywords: Speech corpus, spoken language, phonetic analysis, annotation, ASR, artificial intelligence, corpus linguistics.

Introduction

The scientific value of voice corpora lies, first of all, in the possibility of analyzing the phonetic, lexical, morphological and syntactic features of natural speech through them. Because, unlike written texts, spoken speech embodies intonational means, stress, tempo, pause, realistic acoustic representation of phonemes and other prosodic elements. Emphasizing this aspect, L. Ten Bosch writes: “Speech corpora not only provide an empirical basis for phonetic experiments, but also create the necessary resources for computational linguistics” [Ten Bosch L. 2008.].

The need for phonetic corpora in Turkish linguistics is explained by the development of computational linguistics, as well as the desire to study common phonetic and morphological patterns in the Turkic language family. The TRSpeech corpus, created at Boğaziçi University in Istanbul, is recognized as an important source for revealing the phonetic peculiarities of the Turkish language [Çöltekin Ç. 2014].

Corpora are becoming increasingly important not only for linguistic analysis, but also for training artificial intelligence systems. For example, popular models such as DeepSpeech, wav2vec 2.0, and Whisper have been tested on voice corpora created in dozens of languages and show high results. This puts voice corpora among the global scientific resources of the 21st century.

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

From international experience, we can see that the process of creating voice corpora is based on two main principles: firstly, they provide an empirical basis for linguistic research, and secondly, they serve as a large-scale training data for computer technologies. That is why today, voice corpora created in English, Russian, Turkish, and German are gaining importance not only for linguists, but also for programmers involved in artificial intelligence.

Methodology

The use of audio materials in corpus linguistics requires a specific methodological approach. First of all, in the process of collecting speech recordings, special attention is paid to the age, gender, regional affiliation and social stratification of respondents [Baevski A., Zhou Y., Mohamed, A., Auli, M. 2020]. Because these factors directly affect the use of language units, pronunciation variants and prosodic features. Therefore, great importance is attached to the annotation process in modern corpora, and there is an effort to implement phonetic, morphological and syntactic analysis based on a single standard.

Results

English speech corpora have been developed since the 1980s and have been used as a primary source for phonetic research and training automatic speech recognition (ASR) systems. Such corpora have also served as methodological models for later speech databases created for other languages.

One of the earliest voice corpora is ***the TIMIT Acoustic-Phonetic Continuous Speech Corpus*** [Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren N. L .]. This corpus was created in the late 1980s in collaboration with the US company Texas Instruments and the Massachusetts Institute of Technology and contains recordings of 10 sentences pronounced by 630 respondents. TIMIT is considered one of the most reliable sources in phonetic research, as it incorporates the possibility of phonetic transcription, segmentation and intonation marking. In particular, TIMIT has provided phonetic diversity in training ASR models and expanded the capabilities of

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

automatic speech analysis. Therefore, TIMIT is still used as experimental material in many scientific articles today.

One of the next major projects was ***the Switchboard Corpus*** , which was created in the 1990s. This corpus contains spoken conversations conducted over the telephone [Godfrey JJ, Holliman E. 1997]. The importance of Switchboard is that it recorded the sound of natural conversation, its sociolinguistic features, the use of language tools in real situations. This corpus became the basis for phonetic research, as well as pragmatic and discursive analysis. The Switchboard corpus was also later used to train many ASR systems, increasing the level of automatic recognition of telephone speech.

By the beginning of the 21st century, English audio corpora had expanded in size and scope. A prime example of this is ***the LibriSpeech Corpus*** . This corpus was introduced in 2015 and is based on over 1,000 hours of English audiobooks. LibriSpeech has become one of the most widely used resources in the field of automatic speech recognition, due to its open source nature. LibriSpeech serves as a key reference corpus for training deep learning models in particular. Since the corpus is high-quality, transcribed, and large, various neural networks have been tested and their results are being compared.

Among Russian researchers, the issue of systematic study of spoken speech is also widely covered in the works of Shmelev. For example, Shmelev notes: “Russian spoken speech is much more diverse in terms of syntactic constructions than written language, and these differences are clearly visible in corpus studies” [Shmelev AD 2003].

Another important project to create a Russian language voice corpus is **RUSLAN**. This corpus provides phonetic and morphological analysis of recordings of live Russian conversations. The RUSLAN project is more focused on phonetic studies, the study of the stress system, intonation and speech tempo. The corpus is also used to study the influence of regional dialects and social stratification on speech in Russian.

The creation of Russian language voice corpora is also a key resource in the development of speech technologies, in particular, automatic speech recognition (ASR) systems. For example, projects such as the SpeechOcean

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

Russian Corpus or OpenCorpora Speech are used to train artificial intelligence systems. These corpora contain hundreds of thousands of hours of audio recordings, which are used to train neural networks to distinguish phonetics.

Voice corpora created in the Russian language are important not only for phonetic studies, but also for sociolinguistic research. Because in live communication, it is clearly visible that speech is closely related to social factors. Therefore, corpora such as MKUR or RUSLAN also include detailed information about the age, gender, social status, and regional affiliation of respondents. This allows for a comparative analysis of speech variants.

The corpus of spoken Russian includes samples of mass and spontaneous Russian speech, as well as transcriptions of Russian films. Standard spelling rules were used when recording spoken samples. It is possible to create lexical, morphological and semantic queries. The user can create his own subcorpus (for this purpose, sociological parameters can also be used). The corpus includes samples belonging to various genres and types, as well as from different geographical regions (Moscow, St. Petersburg, Saratov, Ulyanovsk, Taganrog, Yekaterinburg, etc.). Below is information about about 15 thousand words from 4.5 thousand texts currently available in its composition. The Multimedia Russian Corpus ***(MURCO)*** is intended for the study of spoken speech in various genres. Initially planned as a corpus of film dialogues, a trial version of MURCO was launched in 2009-2010. Later, the corpus was enriched with samples of spoken speech in various genres. Currently, the corpus size is approaching 5.5 million words. MURCO includes the following sections (subcorpora).

One of the first major projects was the TRSpeech corpus. This corpus was developed at Boğaziçi University in Istanbul and aims to document the phonetic aspects of the Turkish language on a large scale. The TRSpeech corpus contains speech recordings from hundreds of respondents, and they were collected from individuals belonging to different regional dialects, age groups, and social classes. Çöltekin states: “The Turkish corpus has created an empirical foundation in linguistics and accelerated the development of

World Bulletin of Education and Learning (WBEL)
ISSN (E): 3072-175X
Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

computer linguistic resources for the Turkic languages." The corpus is also used in phonetic research, as well as in automatic speech recognition systems. Another important project is the Boğaziçi University Speech Corpus (BOSC). This corpus contains phonetic, morphological and syntactic analysis of Turkish spoken language and is made publicly available to the scientific community. With the help of BOSC, pronunciation variants, stress system and prosody features of Turkish have been studied in depth. The corpus is also used to adapt artificial intelligence models to Turkish.

Another important source in Turkish is the Turkish Broadcast News Corpus. This corpus is based on audio recordings from Turkish broadcast news programs. It was created in the 2000s and contains hundreds of hours of news discourse. This type of material has become important in the study of formal speech, syntactic structures, and prosodic features.

Today, a number of scientific works are being carried out on voice corpora. L. Dolores' doctoral dissertation "Spoken language corpora: Approaches for facilitating linguistic analysis" is devoted to the issues of creating, structural organization and adaptation of spoken language corpora to linguistic analysis [Lemmenmeier-Batinić D.]. The main object of the research is speech materials based on audio recordings, and the subject is methods of their scientific annotation and transformation into a corpus suitable for analysis. The author systematizes the process of corpus creation in the following stages: audio data collection (recording design); orthographic and phonetic transcription; time-alignment; multi-layered annotation (prosodic, pragmatic, discursive); corpus interface and search mechanisms. The study analyzes real spoken corpora in European languages (including GOS – spoken Slovene corpus) as examples. Classical and modern models of audio corpus creation are compared. The features that fundamentally distinguish oral speech from written corpus are scientifically revealed. The dissertation offers a complete methodological model for creating an audio corpus.

Another work in Russian linguistics is E.V. Sidorova's " Principles" creation multimedia corpus c pragmatic with markings emotional component speech

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

and ego use at artificial bilingualism ", this dissertation is dedicated to the creation of a multimedia speech corpus (audio + video) [Sidorova E. B.] .

The focus of the study is on live oral speech, especially its emotional and pragmatic components. In the work, the audio corpus is characterized by the following features: audio and video synchrony; time-based marking of speech segments; pragmatic tagging (communicative intention, addressee, situation); special annotation layers that characterize the emotional state. The author interprets the audio corpus as a system related to the communicative situation, unlike a simple phonetic or text corpus. He bases the audio corpus as a multimedia resource. He shows the possibilities of studying emotion and pragmatics in oral speech through the corpus.

The next work is the dissertation of H. Navar "Modern Standard Arabic Phonetics for Speech Synthesis", which is devoted to the creation of a phonetic audio corpus for the modern Arabic language and its application in speech technologies (speech synthesis, ASR) [Halabi N.]. The object of the dissertation is the phonetic units of audio recordings, and the subject is their modeling and systematization. The dissertation clearly describes the following stages: preparation of audio recordings at a professional level; phonetic and phonological transcription; segmentation and labeling system; adaptation of the audio corpus to machine learning models. The author analyzes in detail the problems of creating a standardized audio corpus for a language with limited resources. The relationship between the audio corpus and ASR/TTS models is revealed on a scientific basis. The dissertation belongs to the direction of speech corpus engineering.

The corpus of spoken text is also important because language in linguistic perception appears in the form of speech. The text in written form is represented in the form of graphics, that is, it has a material form. In oral speech, language undergoes changes such as abbreviations, omissions, renaming, and different pronunciations. The base included in the corpus is an important source for processing natural language. Linguistic analysis of text using a corpus is also important due to its accuracy compared to sensory analysis, richness of material, and ease of work. The researcher's knowledge of

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

collecting and analyzing linguistic knowledge is limited, and no matter how extensively he studies the language, he may not be aware of new findings generated in the speech process. The size and diversity of the linguistic base included in the corpus further expands the scope of study. The expression of the frequency of a particular word in the text in terms of indicators helps to draw conclusions about the style within which language or speech units are accepted, ready for use, or not accepted in the process of communication.

Conclusion

Audio corpora are an important scientific resource that allows for in-depth study of the phonetic, prosodic, lexical, and grammatical features of natural speech that cannot be fully reflected in written sources. They serve as a fundamental training base for the development of automatic speech recognition and other artificial intelligence technologies, as well as an empirical analysis of the real-world use of language units.

International experience (for example, in English, Russian, Turkish) shows that the scientific effectiveness of audio corpora is directly related to taking into account the composition of respondents, standardizing multi-layer annotation, and ensuring audio-text synchrony. In this regard, large audio corpora appear as a strategic resource of modern linguistics and speech technologies and serve as a solid scientific and methodological foundation for creating an audio corpus in the Uzbek language.

REFERENCES

1. Ten Bosch L. Building and using speech corpora for speech technology research. Speech Communication, 50(11–12), 2008. -B. 195.
2. Cöltekin Ch. A corpus of Turkish spoken language. Proceedings of LREC 2014. –B. 102.
3. Baevski A., Zhou Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems. -B. 87.

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

4. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium, –B. 12.
5. Zue V., Seneff S. Transcription and annotation of the TIMIT database. Speech Communication, 1996. 9(5–6), -B. 233.
6. Shmelev A. D. Russian conversational language: korpusnye issledovaniya. 2005. Moscow: Yazyki slavyanskoy kultury, – B . 119.
7. Alekseev M. E. Russian upper body: structure and application. 2008. Moscow: Izdatelstvo MGU, –B. 54.
8. https://ruscorpora.ru/en/corpus/spoken
9. Boersma B., Weenink, D. Praat: Doing Phonetics by Computer. – University of Amsterdam, 2020.
10. Schmid H. TreeTagger – A Language Independent Part-of-Speech Tagger. - Stuttgart: IMS, 1994. -B. 17.
11. Cöltekin Ch. A corpus of Turkish spoken language. Proceedings of LREC 2014, –B. 102
12. Abdurakhmonova NZ, Urazaliyeva MY Theoretical and practical issues of creating a corpus of oral texts in the electronic corpus of the Uzbek language (http://uzbekcorpus.uz/). Academic Research in Educational Sciences. 2022.
13. Lemmenmeier-Batinić D. Spoken language corpora: Approaches for facilitating linguistic analysis Cumulative thesis presented to the Faculty of Arts and Social Sciences of the University of Zurich for the degree of Doctor of Philosophy. Zurich, 2023. -B.39.
14. Sidorova E.V. Principles of the creation of a multimedia corpus with pragmatic marking of emotsionalnoy sostavlyayushchey rechi i ego ispolzovanie pri ikusstvennom bilingvizme (na materiale angliyskogo i russkogo zyzyka) tema dissertatsii i autoreferata po VAK RF. 2015.
15. Halabi N. Modern Standard Arabic Phonetics for Speech Synthesis. Thesis for the degree of Doctor of Philosophy. Southampton. 2016. -B.157.
16. Abdurakhmanova N. Z. Corps Linguistics . Textbook . Tashkent - 2023. – P. 260.

World Bulletin Publishing
Online Publishing Hub

World Bulletin of Education and Learning (WBEL)

ISSN (E): 3072-175X | Volume 01, Issue 03, December 2025

https://worldbulletin.org/index.php/1

17. DB Mengliev, NZ Abdurakhmonova, VB Barakhnin, RK Shirinova, AR Iskanderova, AZ Otemisov, "Building a Comprehensive Uzbek Lexicon: Bridging Dialects for Text Standardization", 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2440-2444, 2024.
18. Urazaliyeva MY Analysis of the problems of including audio texts in the corpus. Uzbekistan: Language and Culture . 2024. – P. 115-125 .