STORAGE AND RETRIEVAL MECHANISMS OF AUDIO FILES

Urazaliyeva Mavluda Yangiboyevna

Independent researcher at the National University of Uzbekistan

urazaliyeva_m@nuu.uz

Abstract. This study focuses on the development and annotation of an Uzbek audio corpus based on speech recordings from young speakers aged 18–27. The corpus includes monologic, dialogic, and polylogic audio texts reflecting natural spoken Uzbek. Audio segmentation, transcription, and multi-layer annotation were conducted using the Label Studio. The annotation process incorporated speaker diarization, time-aligned transcription, and selected sociolinguistic metadata. The resulting corpus provides a reliable resource for linguistic research as well as for training and evaluating automatic speech recognition and artificial intelligence–based language models.

Keywords: audio corpus, annotation, transcription, youth speech, speaker diarization, Uzbek spoken language.

Within contemporary linguistics – especially in digital linguistics and the domains of automatic speech processing – the systematic analysis of audio texts occupies a central position. Spoken discourse, as a complex linguistic unit, is examined with respect to its internal structure, grammatical organization, and semantic stratification. Such analyses enable researchers to capture the dynamic and context-dependent nature of oral language, which cannot be fully represented through written texts alone. In particular, for Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) systems, the accurate identification and formal representation of grammatical structures in audio data constitute a methodological and technological foundation. The effectiveness of storage and retrieval mechanisms for audio files therefore directly influences the reliability of linguistic annotation, the precision of automatic transcription, and the overall performance of speech-based language technologies (Abduraxmonova N. 2021).

The morphological analysis of audio texts constitutes a crucial stage in identifying their grammatical structure. At this stage, audio data are first transcribed, after which words are segmented into morphemes and each morpheme is annotated in terms of its grammatical categories. This process enables a systematic description of linguistic forms as they occur in spoken discourse.

It should be emphasized that, due to the agglutinative nature of the Uzbek language, affixal markers play a decisive role in morphological analysis. In audio texts, the morphological analysis stage serves to identify their internal components, reveal grammatical units and the relationships between them, and facilitate the automated segmentation of morphemic structures. This procedure reflects the core linguistic characteristics of Uzbek in a comprehensive manner, contributes to the enrichment of language corpora, and provides a solid foundation for the development of artificial intelligence – based language models.

Based on the results of morphological analysis, morphological tags are assigned to the corpus data. This, in turn, enhances the informational richness of the corpus and supports the further improvement of automatic analysis systems, thereby increasing their accuracy and applicability in computational linguistics and speech technologies.

The morphological analysis stage in audio texts represents one of the most critical phases for linguistics, computational linguistics, and artificial intelligence–based systems, as it enables the identification of grammatical and structural properties of linguistic material. The primary objective of this analysis is to transcribe language units formed in spoken discourse, segment them into morphemes, and determine their functions through grammatical tagging based on established morphological categories.

The initial phase of this process is transcription, during which audio material is converted from its phonetic form into an orthographic (written) representation. This stage requires careful consideration of prosodic and articulatory features present in natural speech, including intonation patterns, stress placement, reductions, elisions, and other pronunciation-related phenomena. For instance, in the audio texts available on the uzbekcorpus.uz platform, the verb *"ketayotgan"* ("going") may occur in spoken form as [ket'yotgan] and is transcribed accordingly to reflect actual pronunciation. Speech rate, stress variation, and regional pronunciation features directly influence the transcription process. Therefore, alongside automated transcription tools, linguistically informed expert verification remains essential to ensure accuracy and reliability at this stage.

The development of audio corpora involves several essential components, which can be outlined as follows. First, audio texts constitute recorded speech samples representing a variety of genres, such as interviews, dialogues, monologues, radio broadcasts, narratives, and other forms of spoken discourse. These materials provide authentic linguistic data reflecting real communicative situations. Second, transcription refers to the conversion of recorded speech into a written form. This process is typically carried out in two main formats: phonetic transcription, often using the symbols of the International Phonetic Alphabet (IPA), and orthographic transcription, which represents speech in standard written form. Each type serves different analytical purposes within linguistic research. Third, annotation involves the additional labeling of speech data with information on morphological, syntactic, and prosodic features. Such annotations enable deeper linguistic analysis and facilitate the use of audio corpora in computational applications. Fourth, metadata include demographic and sociolinguistic information about the speaker, such as age, gender, region, or social background. These data are crucial for contextualizing speech samples and for conducting variationist and sociolinguistic studies. Finally, technical specifications describe the recording conditions and parameters, including audio quality, file format, sampling rate, and other relevant technical characteristics.

The integration of these components distinguishes audio corpora from other types of linguistic resources and ensures their readiness for both scientific research and technological applications (see Figure 3.2.1).

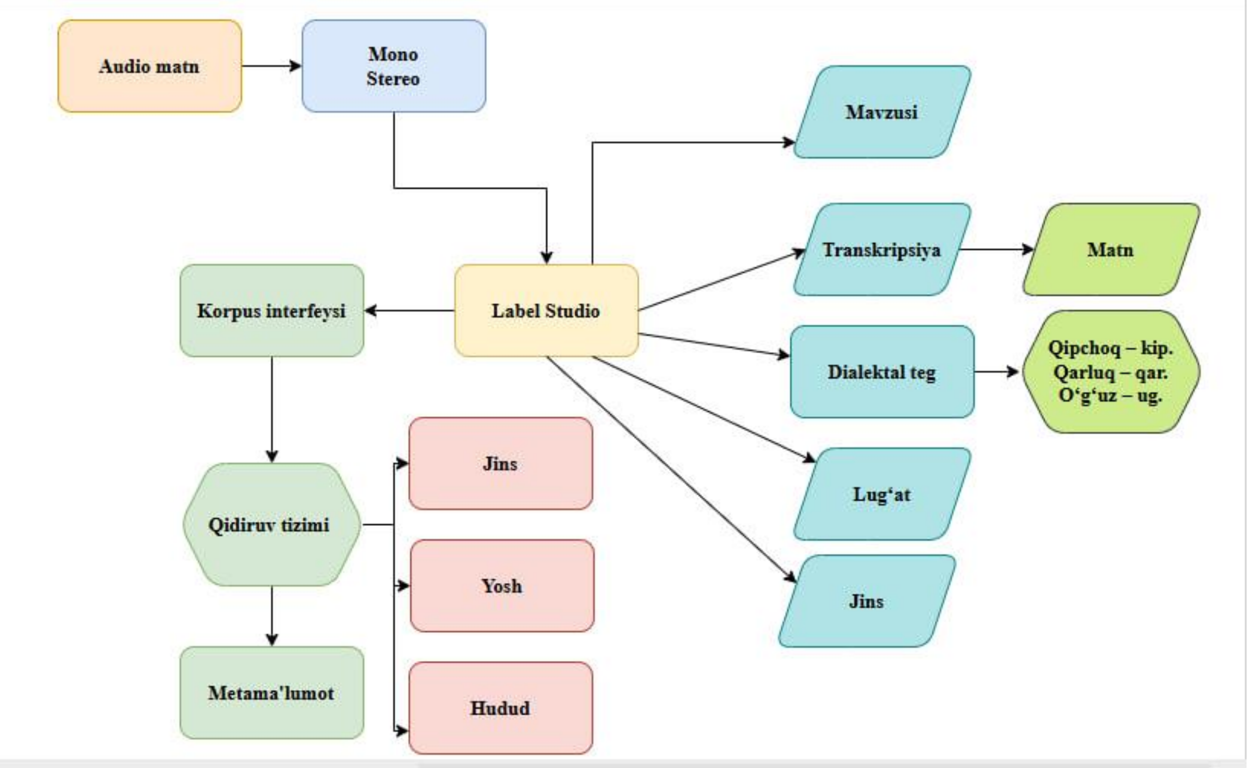
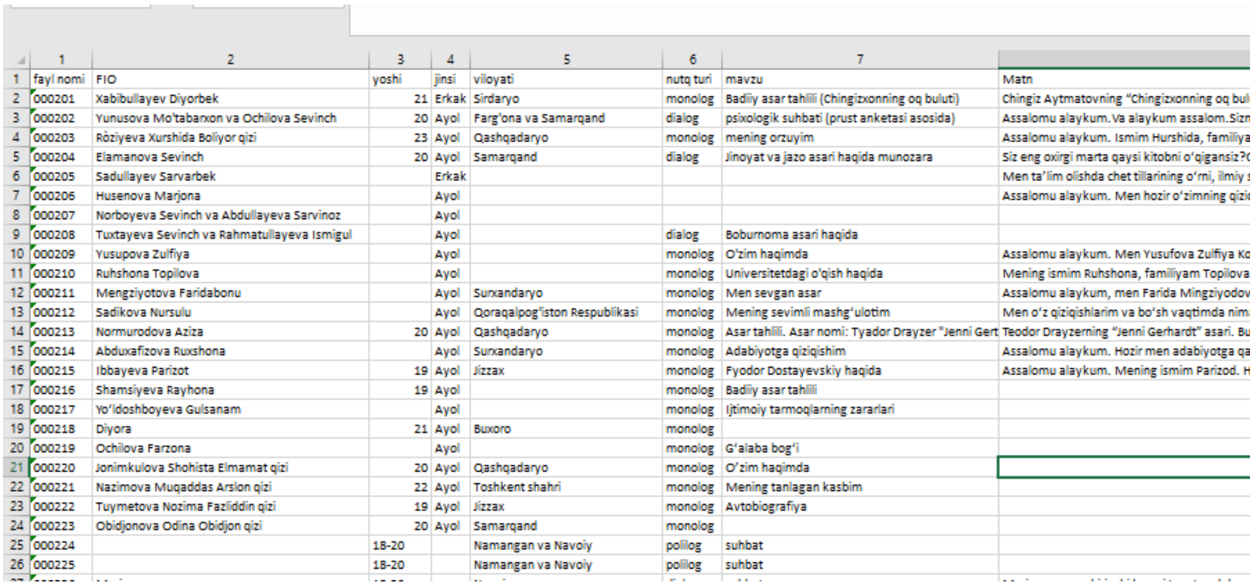


Figure 3.2.1. The process of annotation and processing of an audio text corpus

The morphological analysis of audio texts constitutes a crucial stage in determining their grammatical structure. At this stage, audio materials are first transcribed, after which words are segmented into morphemes, and each morpheme is systematically annotated in accordance with relevant grammatical categories.



	1	2	3	4	5	6	7	
1	fayl nomi	FIO	yoshi	jinsi	viloyati	nutq turi	mavzu	Matn
2	000201	Xabibullayev Diyorbek	21	Erkak	Sirdaryo	monolog	Badiiy asar tahlili (Chingizxonning oq buluti)	Chingiz Aytmatovning “Chingizxonning oq bul
3	000202	Yunusova Mo'tabarxon va Ochilova Sevinch	20	Ayol	Farg'ona va Samarqand	dialog	psixologik suhbati (prust anketasi asosida)	Assalomu alaykum.Va alaykum assalom.Sizn
4	000203	Ròziyeva Xurshida Boliyor qizi	23	Ayol	Qashqadaryo	monolog	mening orzuyim	Assalomu alaykum. Ismim Hurshida, familiya
5	000204	Elamanova Sevinch	20	Ayol	Samarqand	dialog	Jinoyat va jazo asari haqida munozara	Siz eng oxirgi marta qaysi kitobni o‘qigansiz?(
6	000205	Sadullayev Sarvarbek		Erkak				Men ta’lim olishda chet tillarining o‘rni, ilmiy s
7	000206	Husenova Marjona		Ayol				Assalomu alaykum. Men hozir o‘zimning qizi
8	000207	Norboyeva Sevinch va Abdullayeva Sarvinoz		Ayol				
9	000208	Tuxtayeva Sevinch va Rahmatullayeva Ismigul		Ayol		dialog	Boburnoma asari haqida	
10	000209	Yusupova Zulfiya		Ayol		monolog	O'zim haqimda	Assalomu alaykum. Men Yusufova Zulfiya Ko
11	000210	Ruhshona Topilova		Ayol		monolog	Universitetdagi o'qish haqida	Mening ismim Ruhshona, familiyam Topilova
12	000211	Mengziyotova Faridabonu		Ayol	Surxandaryo	monolog	Men sevgan asar	Assalomu alaykum, men Farida Mingziyodov
13	000212	Sadikova Nursulu		Ayol	Qoraqalpog'iston Respublikasi	monolog	Mening sevimli mashg‘ulotim	Men o‘z qiziqishlarim va bo‘sh vaqtimda nim
14	000213	Normurodova Aziza	20	Ayol	Qashqadaryo	monolog	Asar tahlili. Asar nomi: Tyador Drayzer "Jenni Gert	Teodor Drayzerning “Jenni Gerhardt” asari. Bu
15	000214	Abduxafizova Ruxshona		Ayol	Surxandaryo	monolog	Adabiyotga qiziqishim	Assalomu alaykum. Hozir men adabiyotga qa
16	000215	Ibbayeva Parizot	19	Ayol	Jizzax	monolog	Fyodor Dostayevskiy haqida	Assalomu alaykum. Mening ismim Parizod. H
17	000216	Shamsiyeva Rayhona	19	Ayol		monolog	Badiiy asar tahlili	
18	000217	Yo‘ldoshboyeva Gulsanam		Ayol		monolog	Ijtimoiy tarmoqlarning zararlari	
19	000218	Diyora	21	Ayol	Buxoro	monolog		
20	000219	Ochilova Farzona		Ayol		monolog	G‘alaba bog‘i	
21	000220	Jonimkulova Shohista Elmamat qizi	20	Ayol	Qashqadaryo	monolog	O’zim haqimda	
22	000221	Nazimova Muqaddas Arslon qizi	22	Ayol	Toshkent shahri	monolog	Mening tanlagan kasbim	
23	000222	Tuymetova Nozima Fazliddin qizi	19	Ayol	Jizzax	monolog	Avtobiografiya	
24	000223	Obidjonova Odina Obidjon qizi	20	Ayol	Samarqand	monolog		
25	000224		18-20		Namangan va Navoiy	polilog	suhbat	
26	000225		18-20		Namangan va Navoiy	polilog	suhbat	

Figure 3.2.2. Table of spoken speech materials collected from students and their metadata

In the collected materials, the speakers' ages predominantly range from 18 to 27, which forms the youth speech layer of the audio corpus. The regional dimension of this layer is particularly significant, as it includes participants from multiple areas, namely Samarqand, Qashqadaryo, Surxondaryo, Buxoro, Namangan, Navoiy, Jizzax, and Toshkent. Such broad geographic coverage provides a solid empirical basis for the comparative analysis of regional pronunciation features and lexical variation within the corpus, while simultaneously reflecting the diversity of contemporary youth speech in Uzbek.

In terms of speech-type distribution, monologic speech predominates. Students expressed their views freely on topics such as personal background, the educational process, chosen profession, hobbies, literature, and broader social issues. In addition, a number of recordings include dialogic and polylogic speech. These materials are of particular value for analyzing turn-taking mechanisms, the use of discourse markers, and pause structures in spontaneous interaction.

From a thematic perspective, student speech is largely grounded in everyday life and personal experience. As a result, it exhibits characteristic features of Uzbek spontaneous spoken discourse, including relatively simple syntactic constructions, recurrent lexical units, forms of address, and discourse-organizing elements. These properties clearly distinguish the data from scripted or read-aloud audio texts and substantially enrich the audio corpus with authentic samples of natural speech.

The availability of metadata enables these speech recordings to be annotated, filtered, and selectively retrieved in subsequent stages in accordance with specific research objectives. For example, it becomes possible to isolate speech samples by age group or region, to compare monologic and dialogic speech, or to calculate the frequency of lexical units associated with particular topics. At the same time, these materials are of considerable importance as training and test data for automatic speech recognition systems.

In the transcription of vowel sounds, the symbol system of the “New Uzbek Transcription,” as accepted in Uzbek dialectology, was consistently applied. This system is based on a descriptive phonetic approach (Ashirboyev S. 2013). During the transcription process, strict attention was paid to vowel features such as tongue position (front vs. back), degree of openness (low, mid, high), lip rounding, and vowel length.

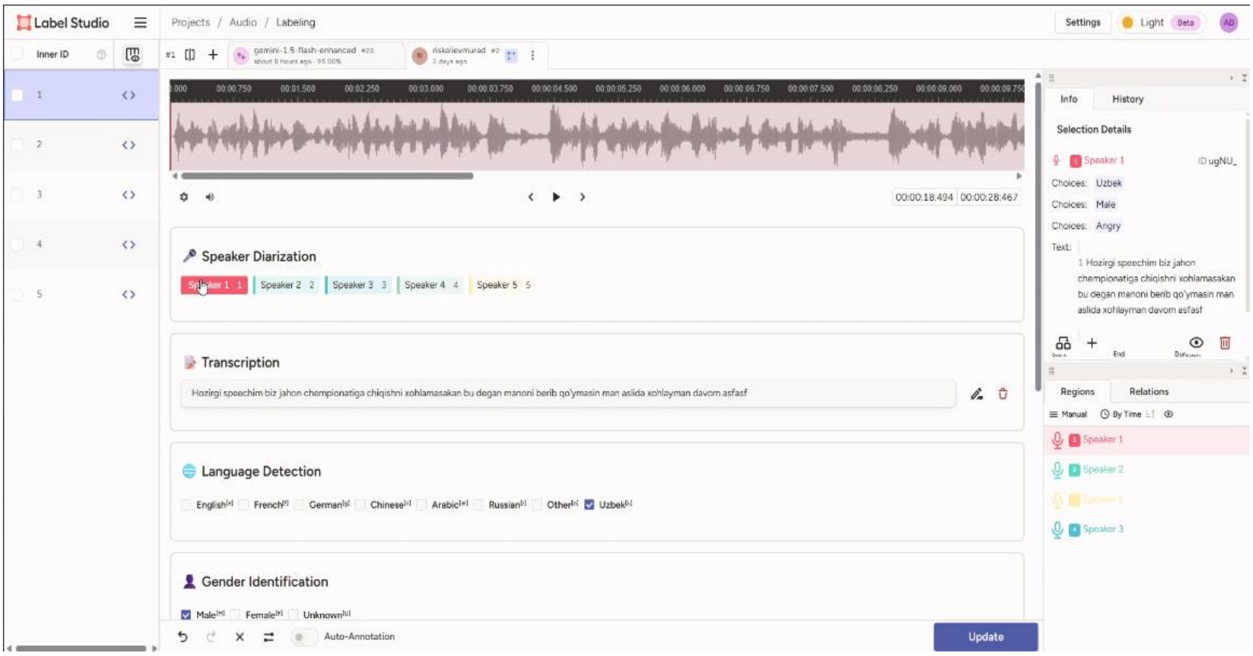
As analytical material, a speech fragment recorded in an urban-type dialect characteristic of the Navoiy viloyati was selected (audio 000226). In this dialect, incomplete preservation of vowel harmony is observed, alongside the presence of vowels with an indifferent character and the active use of long vowels. In particular, the long back low vowel ā, typical of urban dialects (ānča, yāxši), the front low unrounded vowel ä (o‘yläyman, ko‘rgänmän), as well as the front high rounded vowel ü (bo‘lgünimča), occur regularly in speech. This evidence indicates that the urban dialect of the Navoiy region retains a phonetic system that, while aligned with the standard literary language, also preserves distinct dialectal features. The following text was recorded in accordance with the transcription symbols outlined above.

Kitob o‘qish… Men doim kitob o‘qisa, ānča foyda bo‘ladi deb o‘yläyman. Talaba bo‘lgünimča, men kitob o‘qishni juda ham yāxši ko‘rgänmän. Kitoblärniŋ bittäsini tugätib, ikkinčisini darrov bošlardim.

In the given transcription, long vowels were marked by a macron (ā, *yāxši*), while the *ng* sound was represented by the symbol ŋ (η) in accordance with Turkological tradition (*kitoblärniŋ*). In addition, the use of the symbols *š* and *č* made it possible to reflect more accurately the dialect-specific pronunciation features of consonants. As a result, the speech material was recorded in a form that closely approximates actual pronunciation, thereby providing a reliable empirical basis for subsequent phonetic and morphological analyses.

The speech recordings collected from students, together with their associated metadata, significantly expand the sociolinguistic and communicative scope of the audio corpus. This layer, which reflects authentic spoken Uzbek, serves as an important empirical resource for phonetic, lexical, and discourse-oriented research. The corpus's multi-source structure enhances its value as a scientific resource, enabling the investigation not only of normative language use but also of natural spoken language processes in real communicative contexts.

In this study, Label Studio was employed for audio classification, transcription, and the annotation of audio texts. The platform enabled systematic labeling of speech data, facilitating the structured preparation of audio materials for subsequent linguistic analysis and computational processing.



Transcription: Hozirgi speechim biz jahon chempionatiga chiqishni xohlamasakan bu degan manoni berib qo'ymasin man aslida xohlayman davom asfasf

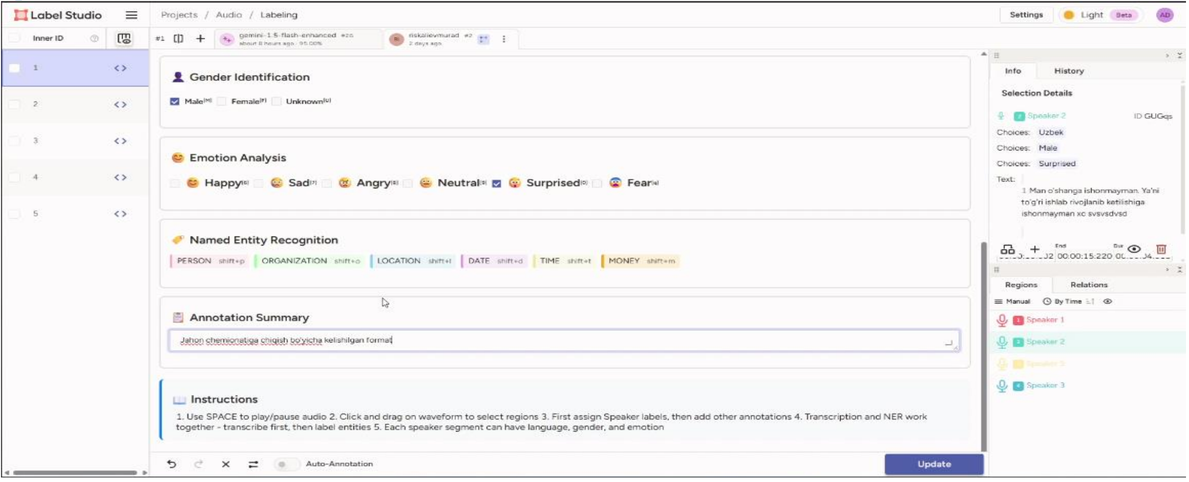
Language Detection: □ English □ French □ German □ Chinese □ Arabic □ Russian □ Other ☑ Uzbek

Gender Identification: ☑ Male □ Female □ Unknown

Selection Details: Speaker 1 ID ugNU_
Choices: Uzbek
Choices: Male
Choices: Angry
Text: 1 Hozirgi speechim biz jahon chempionatiga chiqishni xohlamasakan bu degan manoni berib qo'ymasin man aslida xohlayman davom asfasf

Figure 3.2.2. Linguistic annotation of the audio corpus in the Label Studio environment

Speech rate, stress variation, and regional pronunciation directly influence this stage. Therefore, in addition to automated tools, transcription necessarily requires verification based on linguistic expertise to ensure accuracy and analytical reliability.



Gender Identification

☑ Male[M] □ Female[F] □ Unknown[U]

Emotion Analysis

□ Happy[6] □ Sad[7] □ Angry[8] □ Neutral[9] ☑ Surprised[0] □ Fear[4]

Named Entity Recognition

PERSON shift+p | ORGANIZATION shift+o | LOCATION shift+l | DATE shift+d | TIME shift+t | MONEY shift+m

Annotation Summary

Jahon chemionatiga chiqish bo'yicha kelishilgan format

Instructions

1. Use SPACE to play/pause audio 2. Click and drag on waveform to select regions 3. First assign Speaker labels, then add other annotations 4. Transcription and NER work together - transcribe first, then label entities 5. Each speaker segment can have language, gender, and emotion

Selection Details

Speaker 2 ID GUGqs

Choices: Uzbek

Choices: Male

Choices: Surprised

Text: 1 Man o'shanga ishonmayman. Ya'ni to'g'ri ishlab rivojlanib ketilishiga ishonmayman xc svsvsdvsd

9

Figure 3.2.3. Multi-layer linguistic annotation of the audio corpus in the Label Studio environment

Within the Label Studio environment, audio texts were annotated across several core layers. First, the audio signal was visualized along a time axis in the form of a waveform, which enabled precise identification of speech onset and offset points. This procedure made it possible to segment audio texts into analytically relevant units for subsequent phonetic and statistical analyses. Segmentation was carried out with careful consideration of natural pauses and intonational boundaries inherent in spontaneous speech.

One of the key stages of annotation was the process of speaker diarization, that is, the identification and differentiation of individual speakers. The Label Studio interface allows multiple speakers (e.g., Speaker 1, Speaker 2, etc.) to be defined, which is particularly important for the analysis of dialogic and polylogic audio texts. Speaker separation made it possible to identify turn-taking patterns, speech duration, and indicators of each participant's linguistic activity.

After transcription was completed, a brief description of the overall topic of the text was added at the final stage of annotation. In addition, named entity recognition (NER) elements and emojis relevant to sentiment analysis, when present in the audio material, were tagged under specific labels.

The next annotation layer involved transcription. For each segmented unit, a corresponding written text was entered, ensuring precise time-aligned correspondence between the audio signal and the textual representation. Transcriptions were predominantly produced in orthographic form, while certain phonetic simplifications characteristic of spoken language were normalized within the framework of standard literary norms. This approach facilitates the subsequent use of audio texts as training material for automatic speech recognition (ASR) systems.

Furthermore, metadata such as language detection and gender identification were recorded in the Label Studio environment. These parameters expand the sociolinguistic description of audio texts and, together with age, region, and speech type, enable comprehensive multivariate analysis. During annotation, Uzbek was specified as the language of the recordings, along with the speaker's gender and, where necessary, speech style.

The multi-layered organization of the annotation process ensures that the audio corpus is suitable not only for phonetic and lexical analysis, but also for training machine learning models. In particular, the integration of speaker diarization and transcription layers contributes to improving ASR accuracy and reducing speaker-related recognition errors.

Overall, the annotation process carried out using the Label Studio platform provided a systematic structural, linguistic, and sociolinguistic characterization of the audio texts. This stage represents a crucial intermediate step in preparing the audio corpus for scientific research, forming a solid empirical foundation for subsequent linguostatistical analyses and the development of automatic speech processing models.

REFERENCES

1. Chomsky, N. (1957). Syntactic Structures. The Hague: Mouton. — 116 b. Givón, T. (2001). Syntax: An Introduction. Vol. I. Amsterdam: John Benjamins Publishing. – 424 b
2. 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.
3. Abdurakhmonova, N., Tuliyev, 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.

4. 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).

5. Mengliev, D., Abdurakhmonova, N., Barkhnin, V., Ibragimov, B., Jurakulova, M., Urazaliyeva, M., Islombekov, B. (2025, October). Integrating morphological stemming and syntactic parsing for low-resource Uzbek texts. In AIP Conference Proceedings (Vol. 3377, No. 1, p. 040003). AIP Publishing LLC.

6. 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).

7. 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 литературы (pp. 22-27).

8. Abduraxmonova, N., & Abduvaxobov, G. I. (2021). O‘quv lug‘atini tuzishning nazariy metodologik asoslari. Сўз санъати халқаро журнали, 103. 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.

9. Urazaliyeva, M. Y. (2025). Linguistic and software capabilities of audio-text corporations. World Bulletin of Education and Learning, 1(03), 578-587.