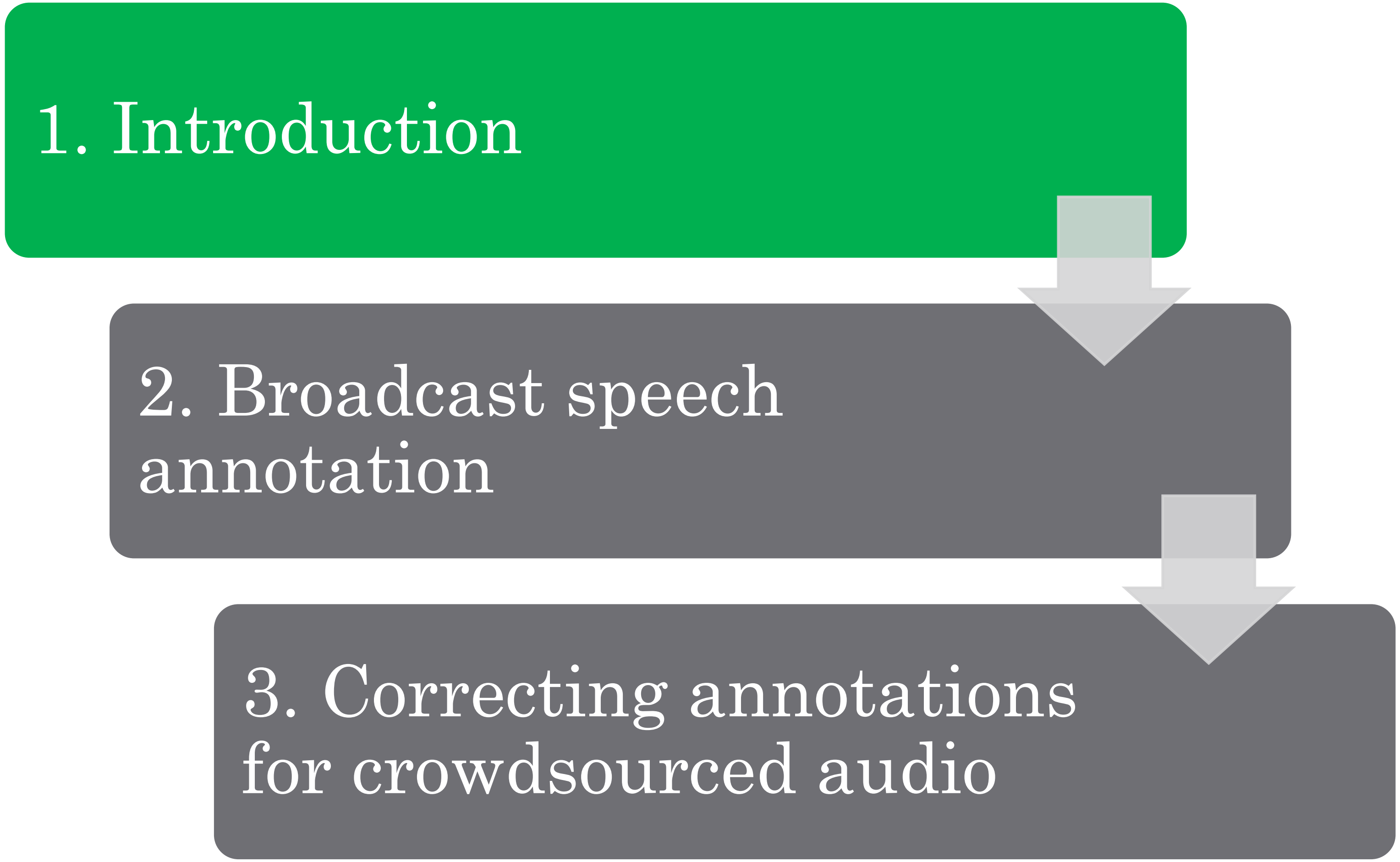
Crowdsourcing tool for annotated speech corpora creation

KHUSAINOV AIDAR
INSTITUTE OF APPLIED SEMIOTICS
KAZAN, RUSSIA

Outline



1. Introduction

2. Broadcast speech annotation

3. Correcting annotations for crowdsourced audio

1. Introduction

«Classical» approach:

- Acoustic models => phonemes
- Pronunciation model => words
- Language model => phrase

End-to-end approaches:

- Better accuracy, require a large amount of training data
- Using data for related languages; pre-trained models

Wav2vec2:

- NLP/Computer Vision fields benefited from using self-supervised pretraining
- Allows to learn **robust** audio representations based **on unlabeled data**
- Masking fragments; model tries to distinguish the true speech representation from distractors (uniformly sampled from other masked fragments)

1. Introduction

The main benefits:

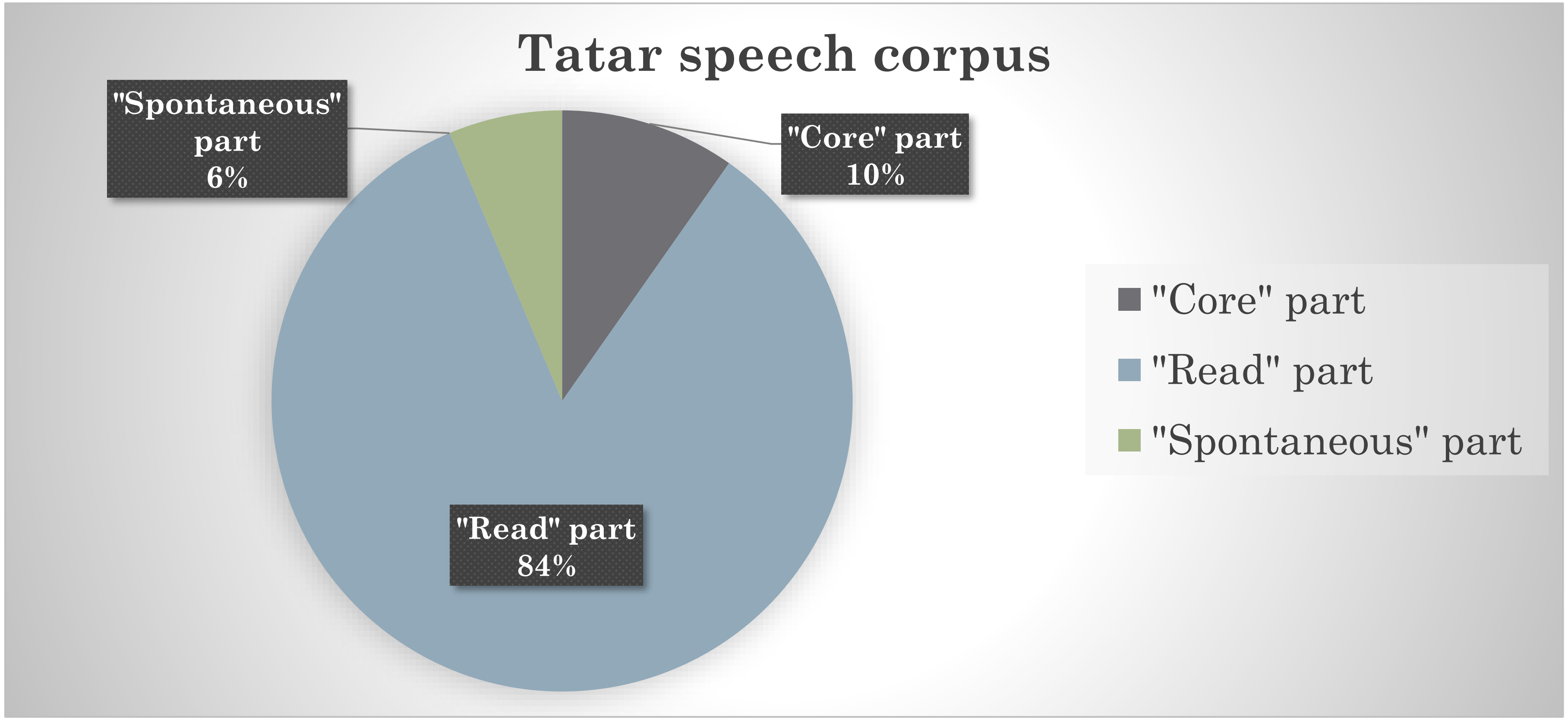
- For low-resourced language it’s much easier to find unlabeled data
- ASR system becomes more robust to background noises, dialects, speakers

Main goals:

- Collect required unlabeled and labeled Tatar speech corpora
- Try the approach with iterative self-supervised pretraining steps on audio data that is increasingly closer to the target domain

1. Introduction

• Recordings' format:	16 kHz, 16 bps mono WAV PCM
• Speakers:	native speakers, Kazan dialect
• Speech type:	read speech



1. Introduction

- Core part
  - Manually collected separate words and phrases
  - Phonetically full, max context
  - 251 speaker, average duration – 0:01:58
  - Total duration – 8:12:16
- Read part:
  - Rule-based selection from text corpus
  - 190 speakers, average duration – 0:22:18
  - Total duration – 70:39:00
- Spontaneous part:
  - Non-overlapping dialogues
  - Total duration – 5:19:33

1. Introduction

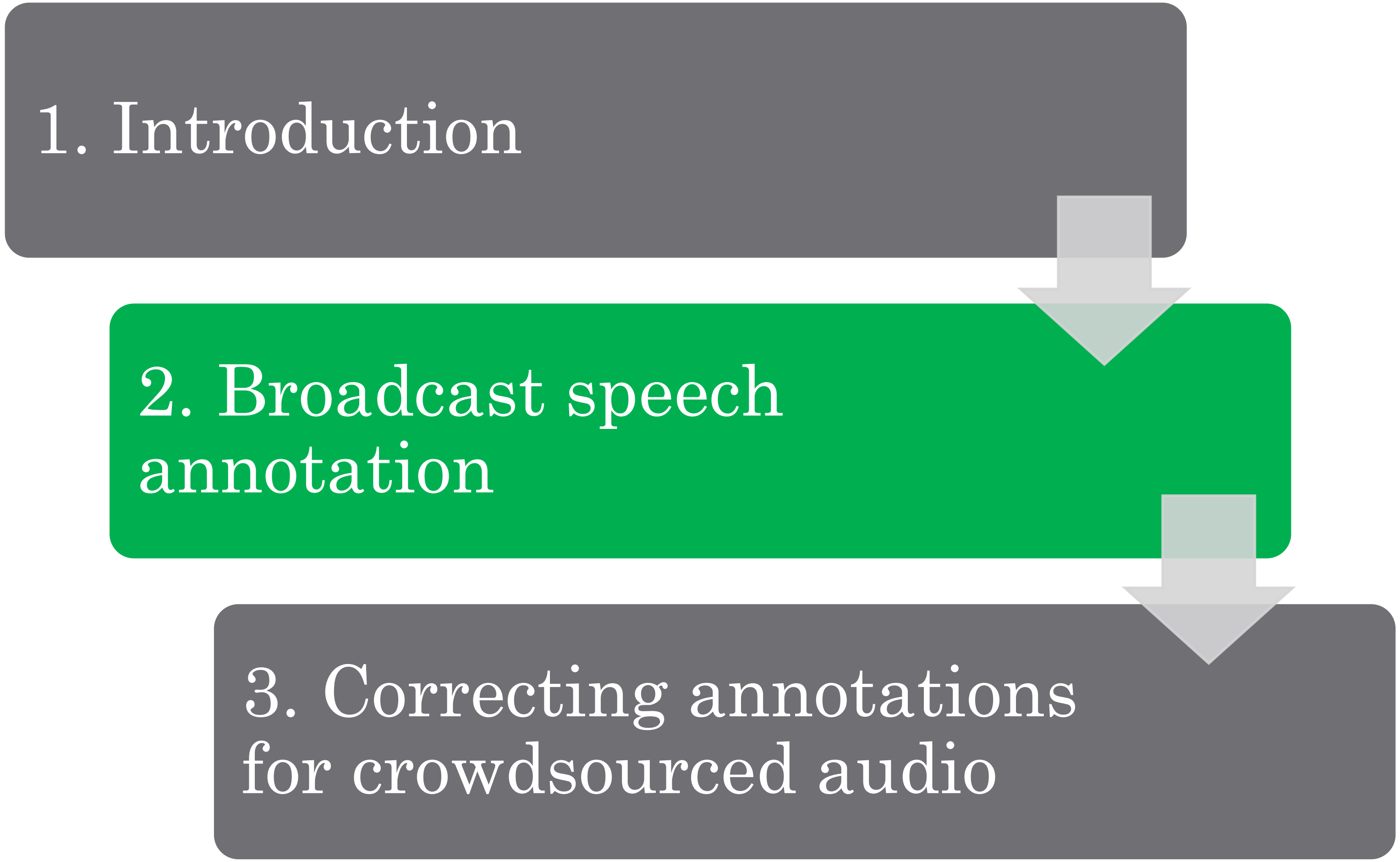
Speech corpus	
# speakers	**499**
Duration	**99:09:59**
Male / Female	30% / 70%
*Spontaneous speech**	*5:19:33*

* We’re recording spontaneous speech too, but it’s not annotated

1. Introduction

- **Annotation**
  - Speaker’s name
  - Age
  - Gender
  - Native language
  - Nationality
  - Speech quality (expert’s mark from 1 to 5)
  - Dialect
  - Microphone model
  - Comment

Outline



1. Introduction

2. Broadcast speech annotation

3. Correcting annotations for crowdsourced audio

2. Project description

Main goal – tools for corpus creation.

1. **Broadcast speech annotation web-tools**

[self-supervised approaches]

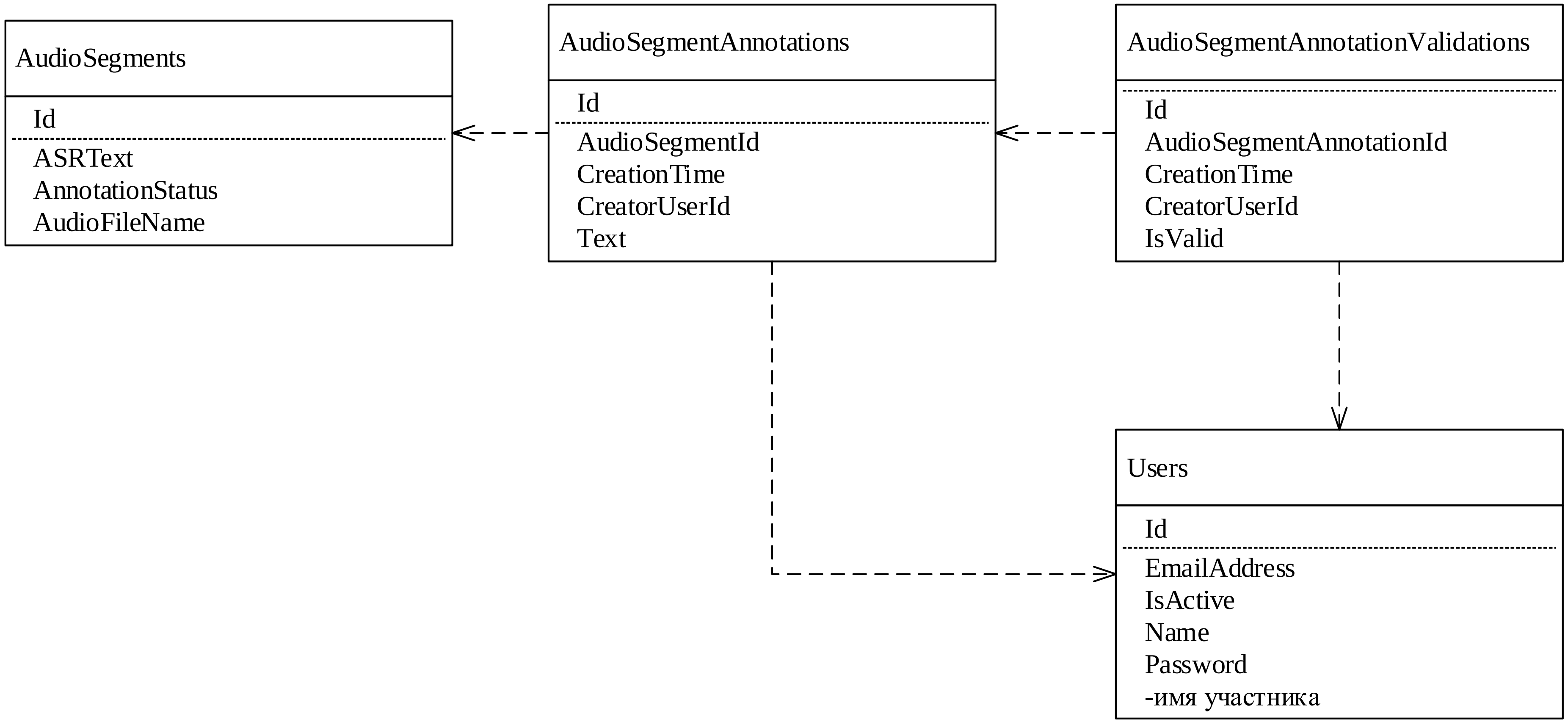
2. Tool to check and correct annotations

2. Project description

- ASP.Net Core
- React.js
- DDD (Domain Driven Design):
  - Infrastructure Layer
  - Domain layer
  - Application Layer
  - Service Layer
  - Presentation Layer
  - Client Applications

2. Project description

PostgreSQL



AudioSegments

Id

ASRText
AnnotationStatus
AudioFileName

AudioSegmentAnnotations

Id

AudioSegmentId
CreationTime
CreatorUserId
Text

AudioSegmentAnnotationValidations

Id
AudioSegmentAnnotationId
CreationTime
CreatorUserId
IsValid

Users

Id

EmailAddress
IsActive
Name
Password
-имя участника

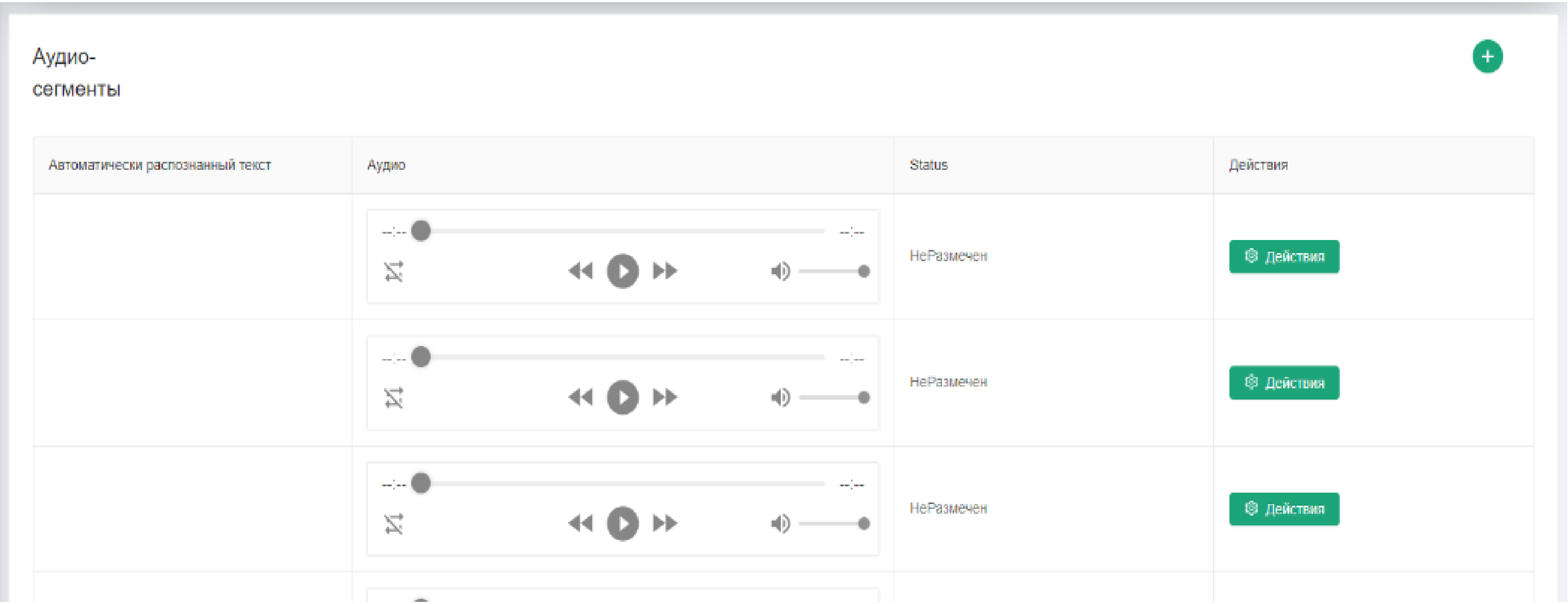
2. Project description

Basic functionality:

- Audio files upload;
- VAD and splitting uploaded files into fragments;
- Web-form for annotating fragment;
- Web-form for validating made annotations;
- View status of annotation of all segments;
- Downloading the annotations.

2. Project description

View fragments’ statuses



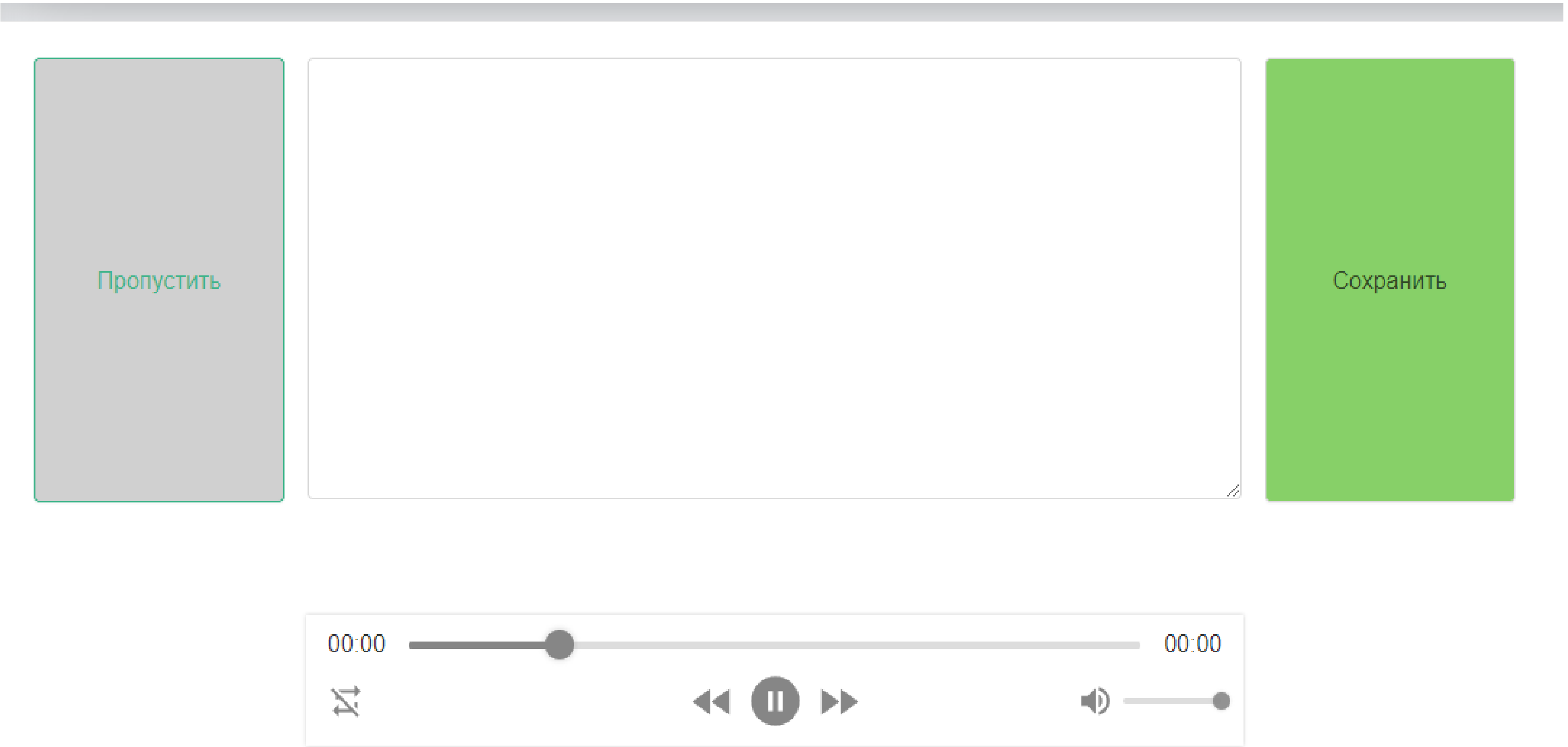
Аудио-сегменты

+

Автоматически распознанный текст	Аудио	Status	Действия
	--:-- --:--	НеРазмечен	Действия
	--:-- --:--	НеРазмечен	Действия
	--:-- --:--	НеРазмечен	Действия

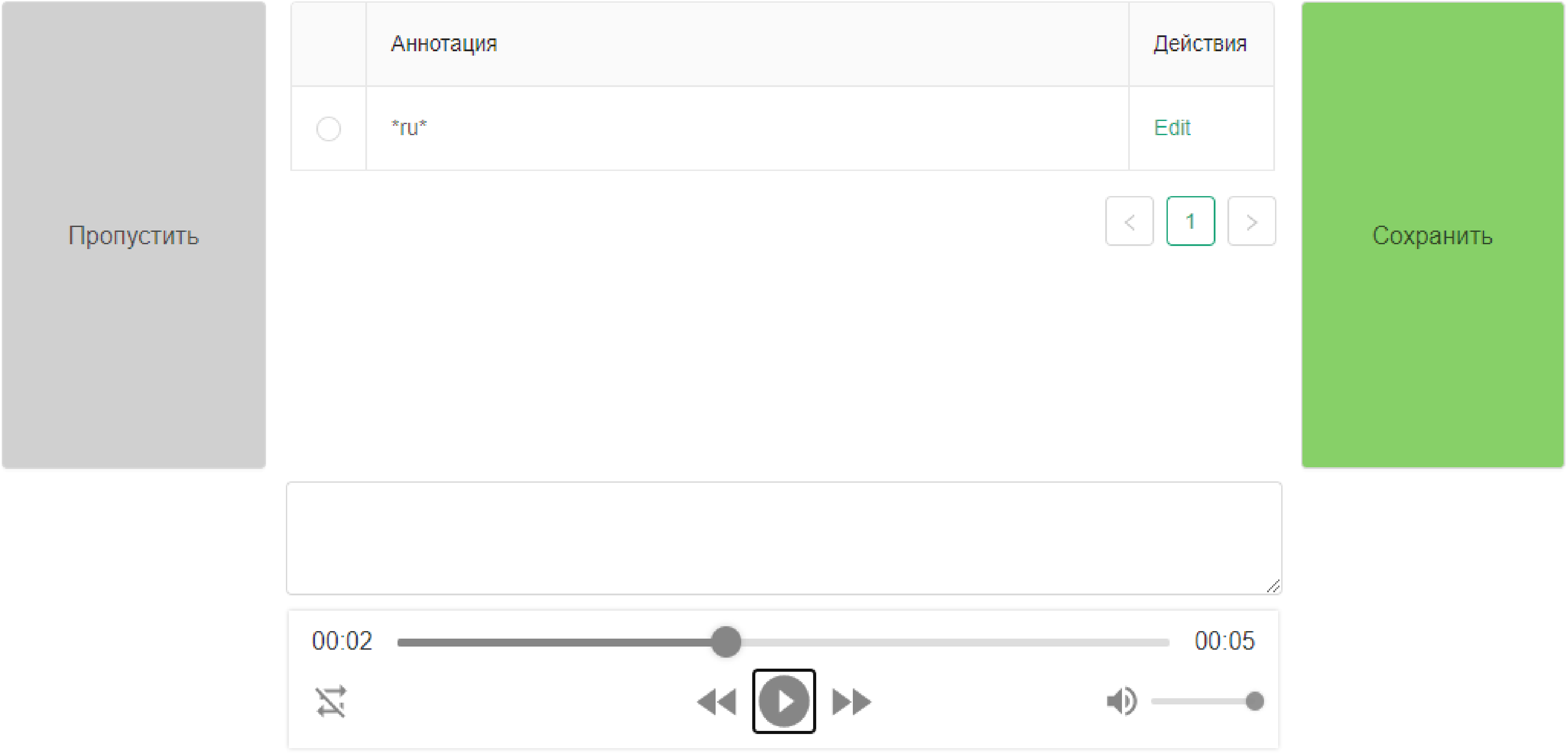
2. Project description

Annotating fragments



2. Project description

Validating fragments



	Аннотация	Действия
○	*ru*	Edit

2. Project description

Initial data:

- From TNV Planeta broadcast company;
- Recordings from December 2019;
- AVI video with mp3 96 kB/s stereo audio signal;
- Converted to 16 bps 16 kHz WAV;
- Total duration – 733 hour.

2. Project description

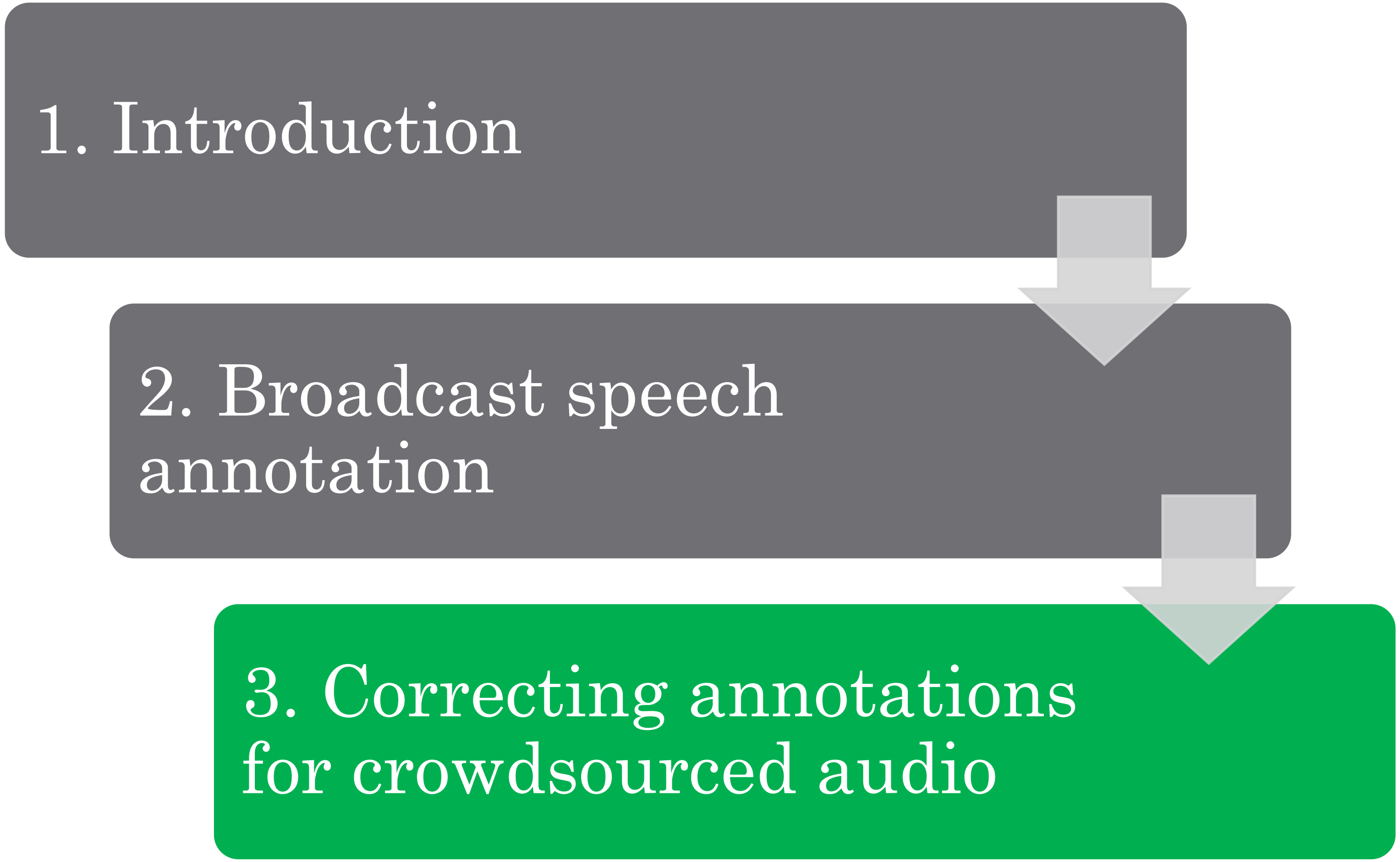
We manually selected segments for the first stage annotation:

- News programs;
- Interviews;
- Talk-shows.

In total 40 segments (23 hours 21 minutes) have been uploaded to the system.

This gave us 22 432 audio fragments with a duration less than 15 seconds.

Outline



1. Introduction

2. Broadcast speech annotation

3. Correcting annotations for crowdsourced audio

3. Telegram bot

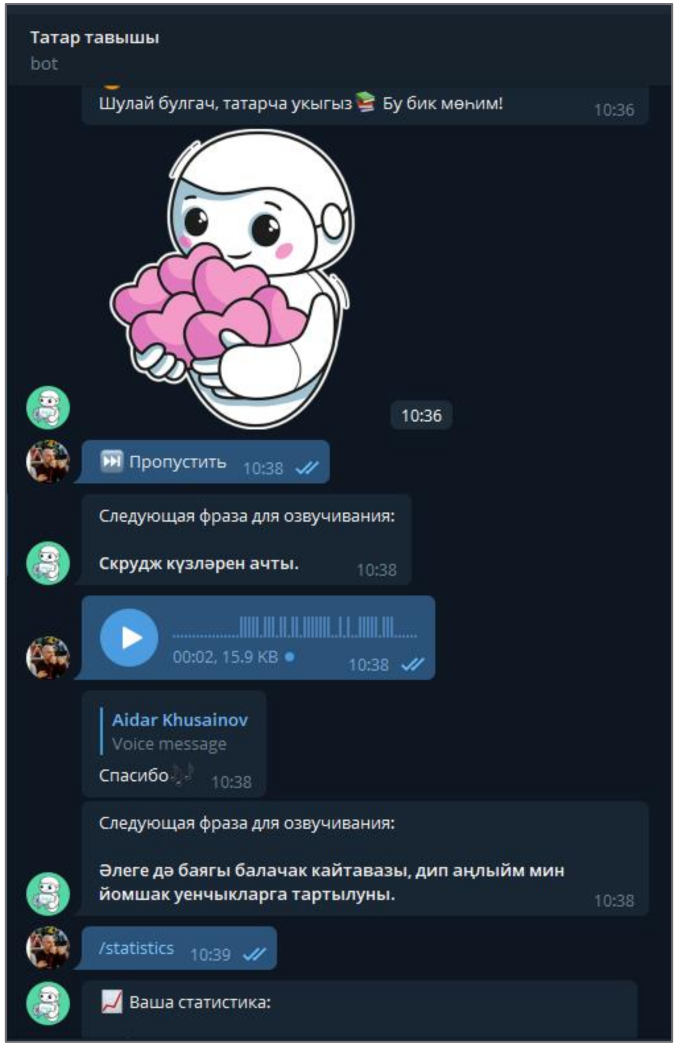
@TatarVoiceBot

Goal – 500 hours

- 408 speakers
- 9 hours 28 minutes

Commands:

1. Next – new phrase to read and send as voice message;
2. Correct – record previous phrase again;
3. Skip – to skip current phrase;
4. Statistics – show user's and overall statistics;
5. Age – select age interval;
6. Russian/Tatar/Help.



Татар тавышы
bot

Шулай булгач, татарча укыгыз 📚 Бу бик мөһим! 10:36

10:36

⏭ Пропустить 10:38

Следующая фраза для озвучивания:

Скрудж күзләрен ачты. 10:38

00:02, 15.9 KB 10:38

Aidar Khusainov
Voice message
Спасибо 10:38

Следующая фраза для озвучивания:

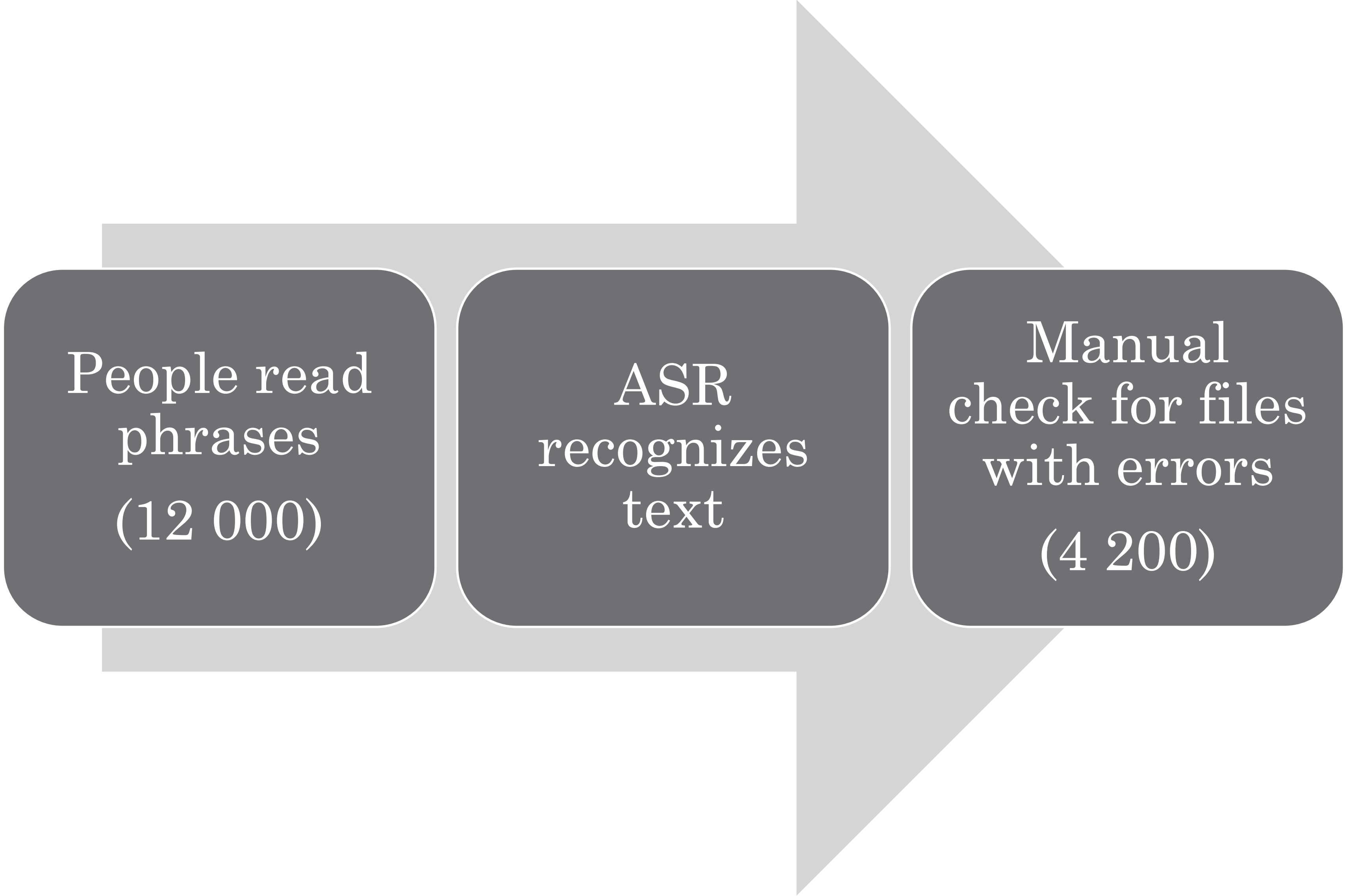
Әлеге дә баягы балачак кайтавазы, дип аңлыйм мин йомшак уенчыкларга тартылуны. 10:38

/statistics 10:39

📈 Ваша статистика:

@TatarVoiceBot

3. Process



People read phrases

(12 000)

ASR recognizes text

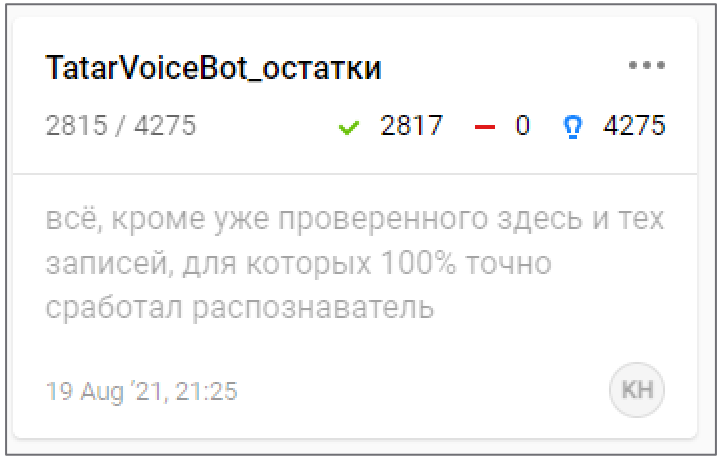
Manual check for files with errors

(4 200)

3. Data collection

https://github.com/heartexlabs/label-studio:

- Allows to build universal platform for all Institute annotation tasks;
- Easy to configure for each task (interface, data, hotkeys, etc.);
- Local, free.

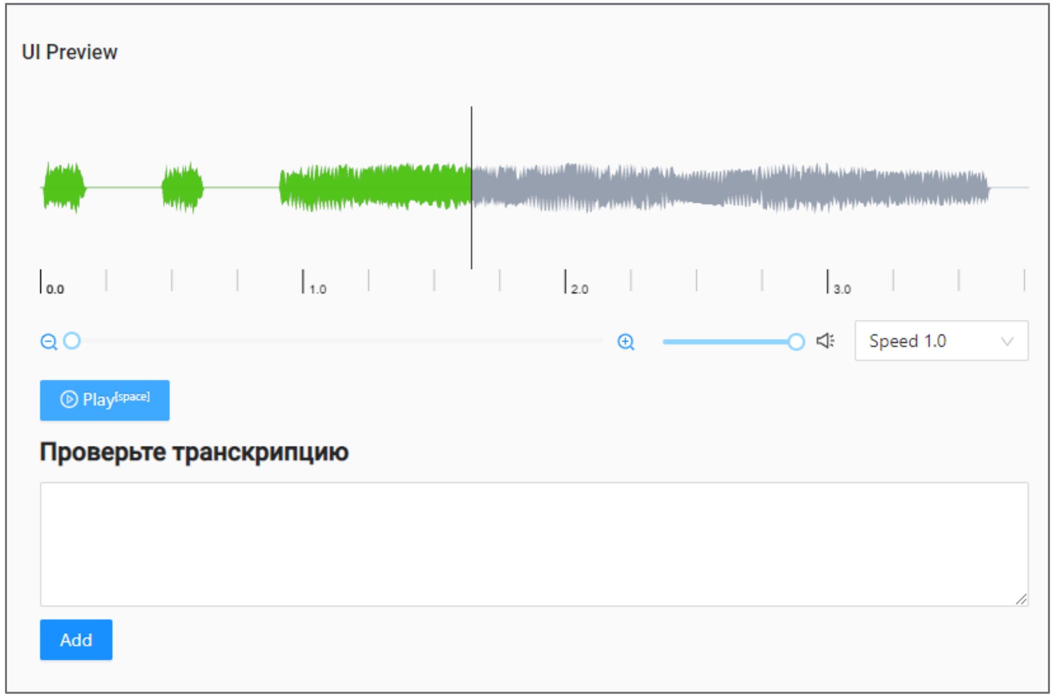


TatarVoiceBot_остатки

2815 / 4275 ✓ 2817 — 0 ? 4275

всё, кроме уже провереннного здесь и тех записей, для которых 100% точно сработал распознаватель

19 Aug '21, 21:25



UI Preview

Проверьте транскрипцию

3. Data collection

Requirements:

- Unlabeled dataset for SS steps;
- Labeled dataset for FT steps.

Labeled datasets:

1. «Tatar Corpus»;
2. «Common Voice»;
3. [new] TV broadcasting;
4. [new] TatarVoiceBot.

3. Data collection

Unlabeled dataset:

1. Audiobooks (read speech, recording studio) – 114 hours;
2. TV broadcasting recordings for 1 month (spontaneous speech, bg noises, music) – 733 hours;
3. 2 radio stations archives (read and spontaneous, bg music) – 215 hours;
4. 100 scientific lectures from YouTube (good SNR) – 87 hours.

Preprocessing:

1. Audio track extraction;
2. Audio conversion to 16 bps 16 kHz mono format;
3. VAD;
4. Filtering short (<4.5 sec) and long (>30 sec) fragments.

3. Data collection

Statistics of final unlabeled dataset

Subcorpus	Initial	After VAD	After filtering short and long fragments
Audiobooks	114 hours (520 fragments)	105 hours (36 712 fragments)	58 hours (17 563 fragments)
TV	733 hours (62 fragments)	472 hours (263 466 fragments)	202 hours (67 065 fragments)
Radio stations	215 hours (398 fragments)	146 hours (29 778 fragments)	29 hours (8 941 fragments)
YouTube videos	87 hours (100 fragments)	81 hours (31 437 fragments)	39 hours (12 764 fragments)
Corpus	**1 150 hours**	**804 hours**	**328 hours**

Thank you

Khusainov Aidar

khusainov.aidar@gmail.com