Tool for distributed creation of annotated speech corpora

KHUSAINOV AIDAR
INSTITUTE OF APPLIED SEMIOTICS
KAZAN, RUSSIA

Outline

1. Previous work

2. Project results and plans

Outline

1. Previous work

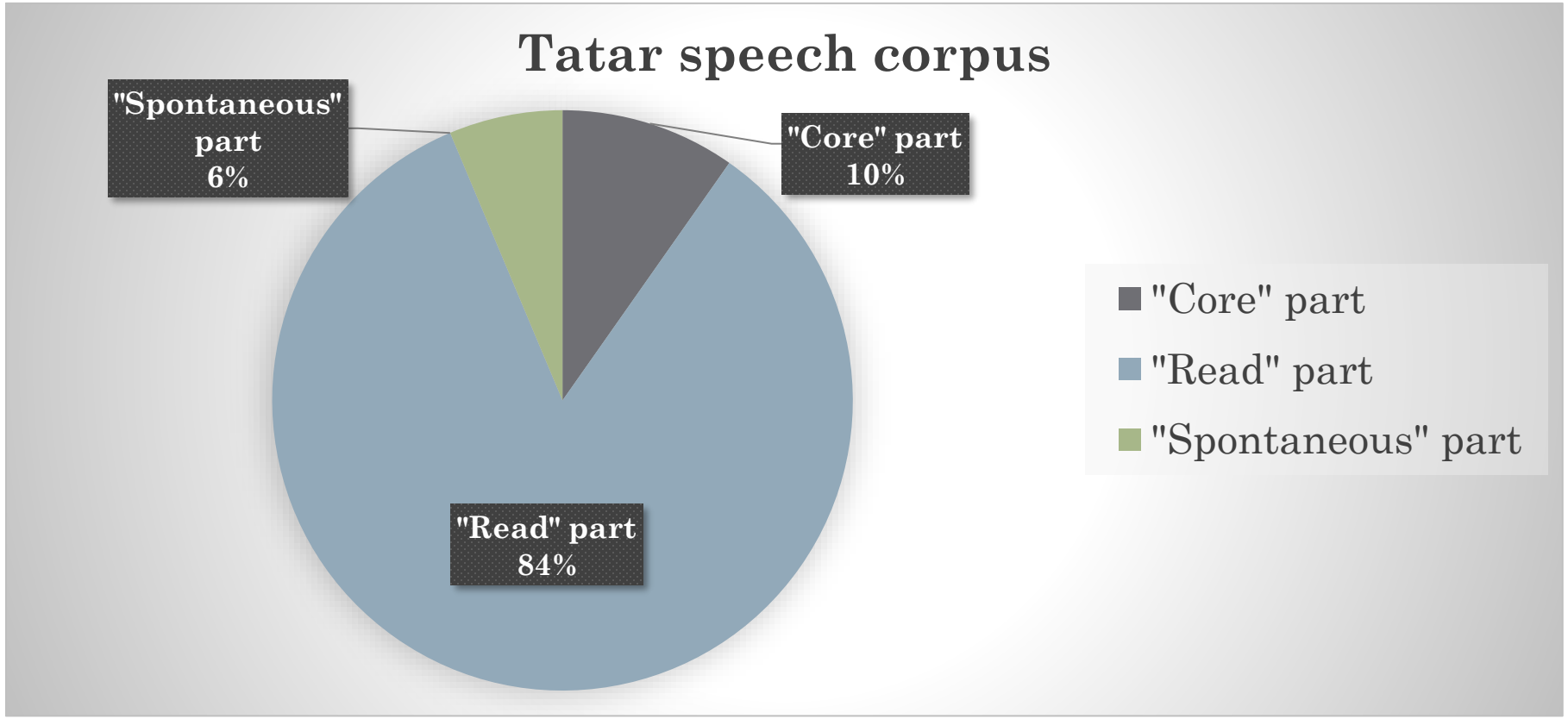
2. Project results and plans

1. Previous work

- Command recognition
  - Dynamic programming algorithm (DTW)
  - Template-matching
  - Tiny one-speaker corpus with several copies for each command
- Isolated word recognition:
  - One-speaker corpus of several hours
  - VAD algorithm based on zero-cross count and signal energy
- Read speech recognition:
  - Multi-speaker corpus of 100 hours
  - ~88% accuracy on the test set
- Spontaneous speech recognition: ???

1. Previous work. Read speech corpus

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



1. Previous work. Read speech corpus

- 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. Previous work. Read speech corpus

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. Previous work. Read speech corpus

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

1. Previous work

- All systems built using the Kaldi toolkit (based on librispeech recipe)

Systems	Acoustic unit	Training audio data	Features	Language models
Mono, MonoSW	monophone	separate words	MFCCs	small 3-gram
Tri1, Tri1SW	triphone	separate words	+ delta, delta-delta	+ 3-gram full
Tri2, Tri2SW	triphone	“Core” part	+ LDA / MLLT	as above
Tri3, Tri3SW	triphone	“Core” part	+ fMLLR	as above
Tri4, Tri4SW	triphone	full training corpus	as above	+ 4-gram
NN, NNSW	triphone	full training corpus	as above	as above

1. Previous work

System	LM	WER		SER		CER	
		Word	Sub-word	Word	Sub-word	Word	Sub-word
Mono	Pruned 3-gram	52,06	-	39,65	-	27,70	-
Tri1	Pruned 3-gram	28,80	24,08	18,32	13,98	12,54	8,18
Tri1	3-gram	22,59	18,42	14,09	10,84	9,78	6,44
Tri2	Pruned 3-gram	24,14	21,20	13,95	11,46	8,69	6,40
Tri2	3-gram	19,08	16,17	10,86	9,11	6,91	5,82
Tri3	Pruned 3-gram	21,16	18,67	11,35	9,74	6,67	5,33
Tri3	3-gram	17,21	14,90	9,04	7,81	5,37	4,91
Tri4	Pruned 3-gram	18,57	19,70	9,29	10,08	5,24	5,54
Tri4	3-gram	15,19	16,09	7,46	8,29	4,18	4,59
Tri4	4-gram	15,10	15,71	7,41	8,05	4,15	4,44
NN	Pruned 3-gram	16,47	17,17	8,27	8,13	4,94	4,29
NN	3-gram	12,99	13,25	6,44	6,37	3,86	3,41
NN	4-gram	12,89	**12,79**	6,38	**6,14**	3,83	**3,29**

Outline

1. Previous work

2. Project results and plans

2. Project description

Main goal – to build first annotated Tatar corpus for broadcast speech.

The main task is to create required tool for annotation process.

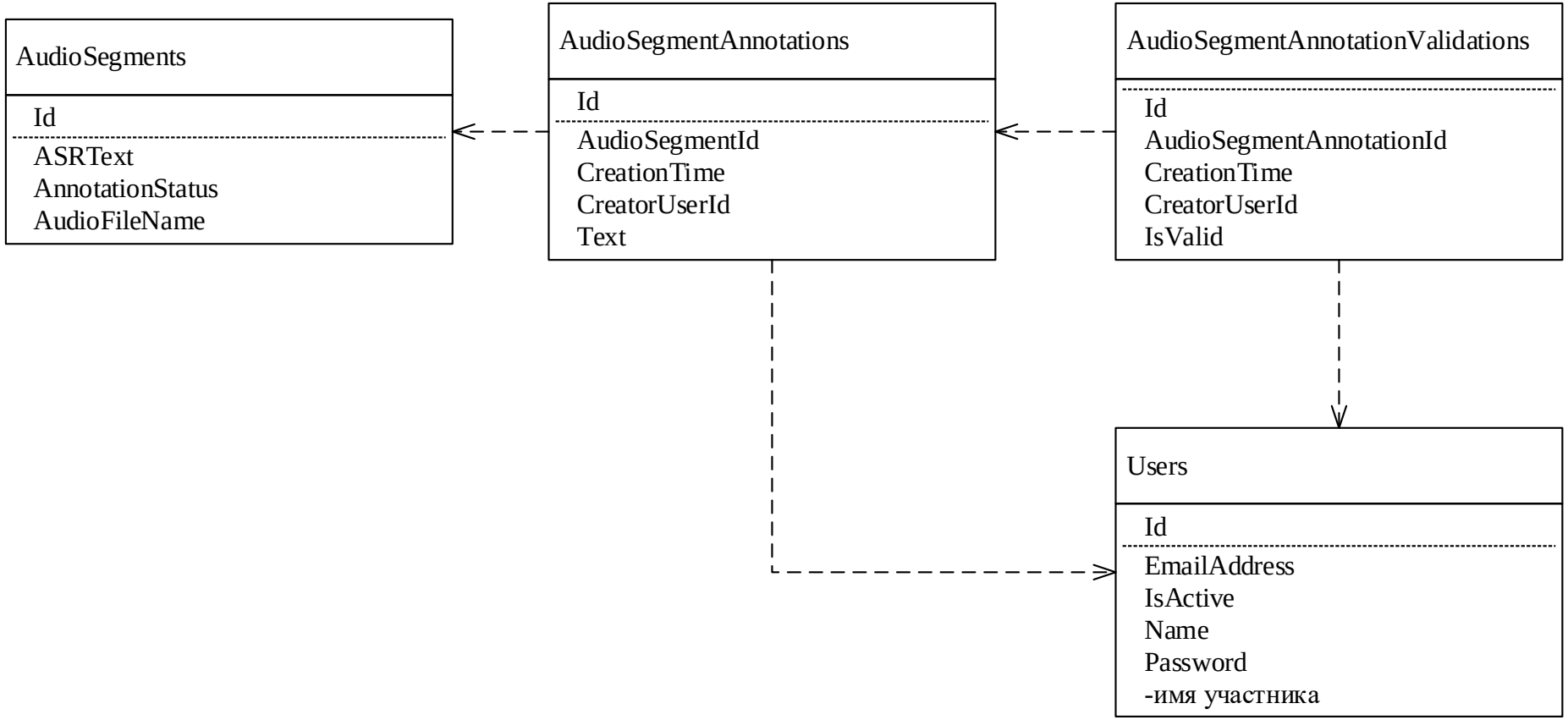
- Software architecture;
- DB creation;
- Software development;
- Audio analysis tools.

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



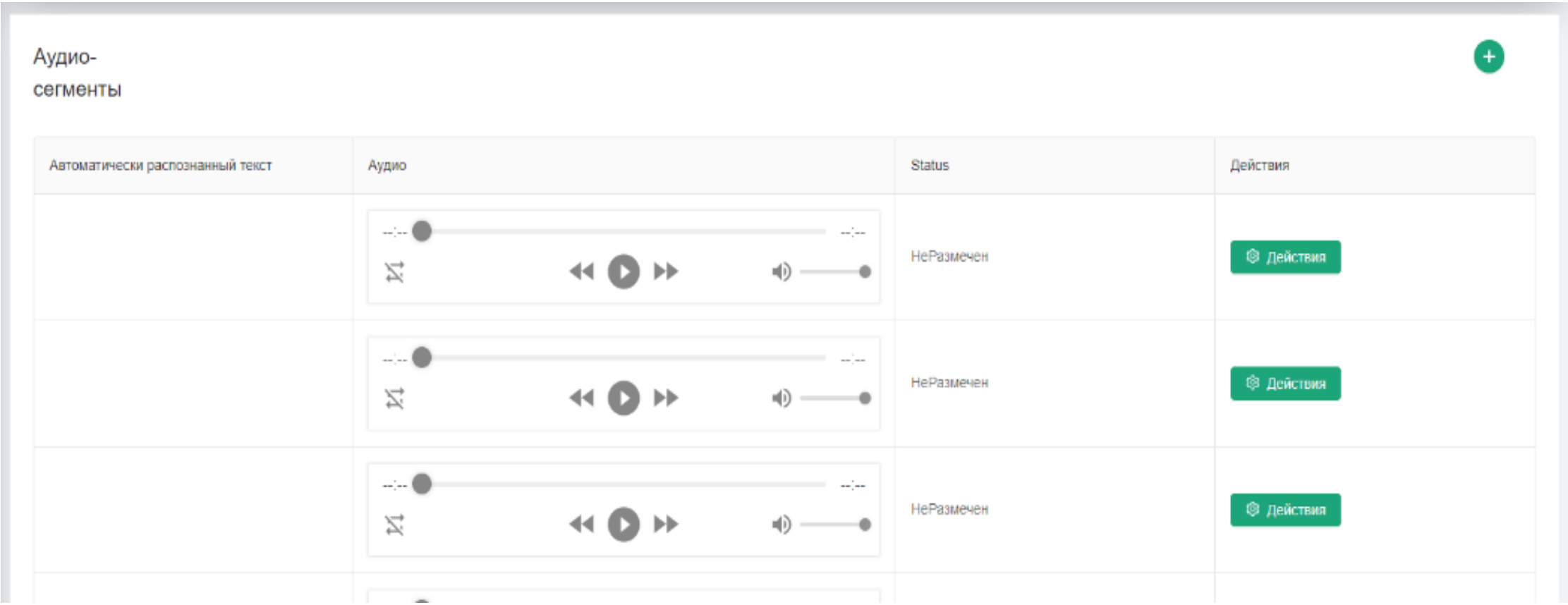
2. Project description

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.

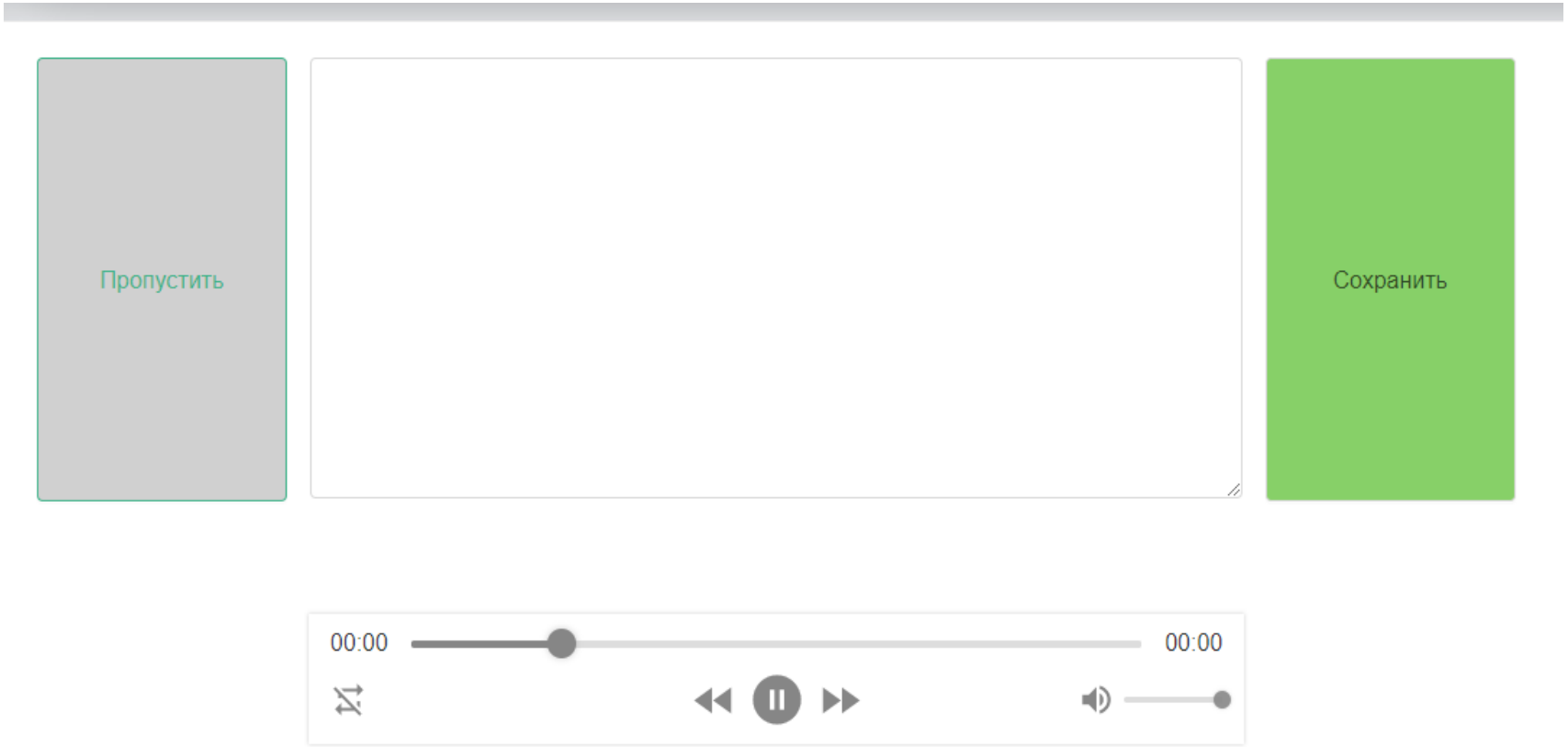
2. Project description

View fragments' statuses



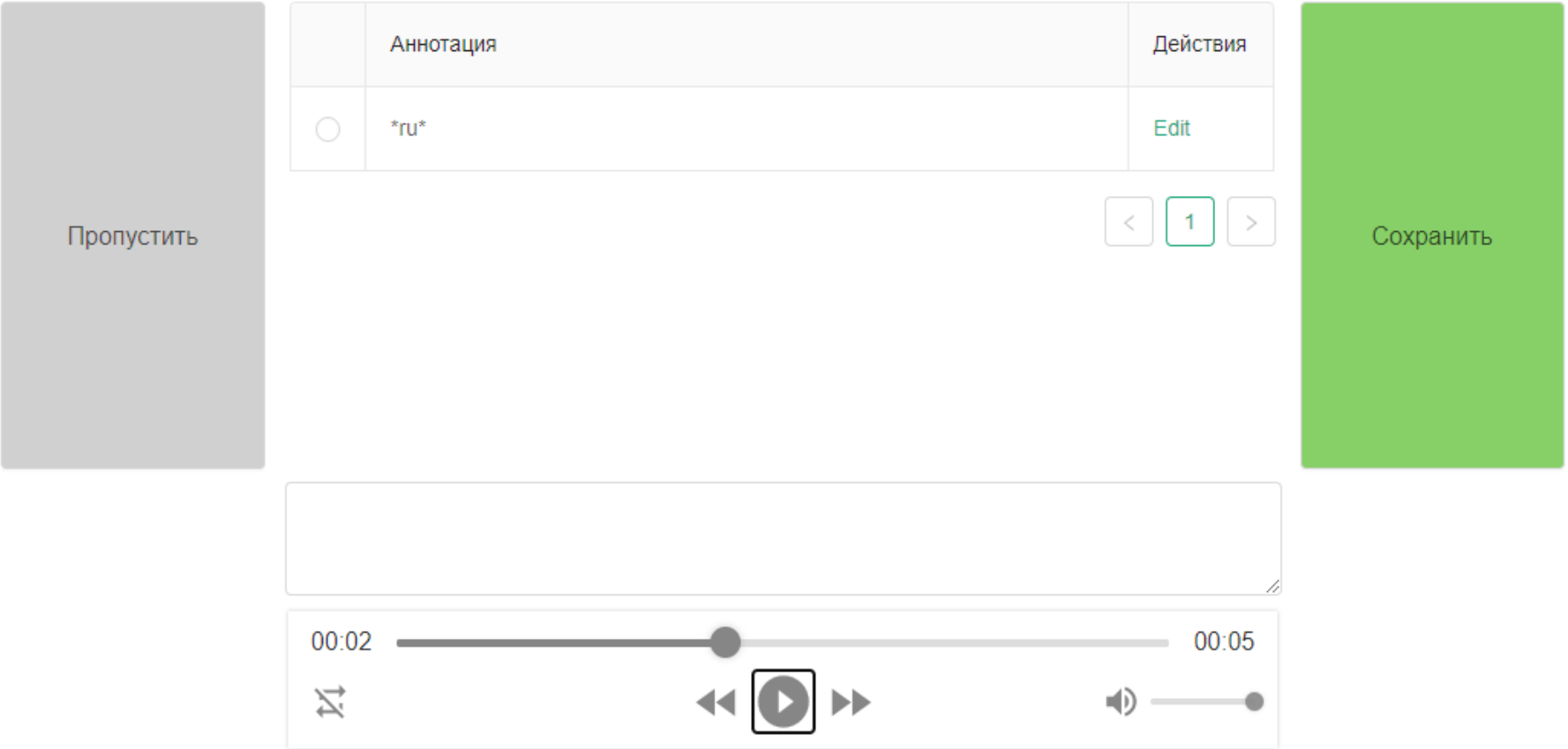
2. Project description

Annotating fragments



2. Project description

Validating fragments



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.

2. Project description

Plans:

- Start working with annotators in University;
- Train an ASR system;
- Use this system to generate hypothesis to speed up the annotation process.

- 2000 hours…

Thank you

Khusainov Aidar

khusainov.aidar@gmail.com