(Infra)structural considerations for high quality ASR for a variety of research domains

25 June 2024
Henk van den Heuvel, Christoph Draxler, Arjan van Hessen

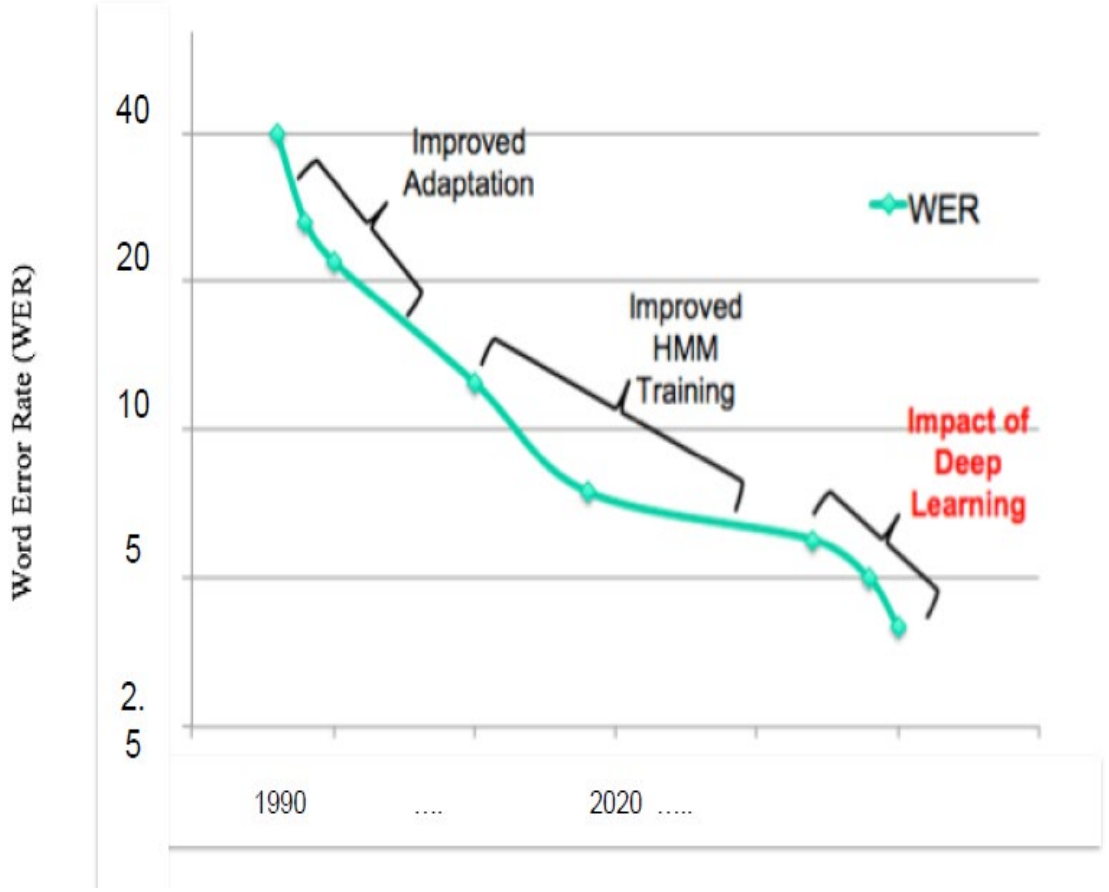
Seminar SURF / Stichting Open Spraaktechnologie:
Exploring infrastructure for Dutch speech recognition

OUTLINE

1. ASR paradigm shift
2. Webservice at Radboud University
3. Webservices at BAS
4. Do it yourself with Whisper
5. Commercial solutions
6. Remaining challenges

1. ASR IN (ORAL) HISTORY

- ASR traditionally was knowledge based built on three components: Acoustic Models, Pronunciation Lexicon, Language Model
- Since 2020: Data-driven Deep Learning, Transformers
- Large Speech models are **not** Large Language Models
  - LSM: Speech input, text output
  - LSMs: Whisper, MMS, Chirp
- LSMs and LLMs can collaborate in applications, e.g. for oral history



TRADITIONAL ASR OUTPUT

mmm <unk> <unk> <unk> linkerhand dan zit je hier de gewrichten van hans en is 't eigenlijk hoef je niet uh ruimte loopt zijn om 't te zien hè want die is kapot dus dat is deze hè van hun rechterhand <unk> links is ook beschadigd hè want die ziet ook dat hier eigenlijk geen met tussen zit uhm en dat die ook een beetje hij staat ook een beetje scheef <unk> mooie recht op mekaar dat klopt dat is dat dat u uh aan de buitenkant mmm <unk> uh vertellen hoe d'r maar 'ns uh begonnen dat dus uh dat gebeurt uh eerst leek 't op colitis ulcerosa <unk> <unk> boek bloed geconstateerd twintigste-eeuwse definitief eigenlijk vastgesteld <unk> <unk> uh <unk> uh <unk> <unk> <unk> <unk> waardoor de dieven invloed met hoofd en drummer klachten bleven wel uh bestaan voornamelijk in mijn uh <unk> mijn ervaringen met het voorschrijven van <unk> zijn heel positief uhm 't belangrijkste punt daarin is dat ze goed werken dat wil zeggen dat mensen uh zelf minder pijn ervaren en dat wij als dokter ook minder ontstekingen aan die vent richten zien 't voordeel uh van het gebruik van <unk> ten opzichte van de conventionele middelen is dat ze uhm uh voor een goed patiënten die niet reageert op conventionele middelen dus beter werken uhm daarnaast is het zo dat uh de patiënten ook aangeven dat ze vaak sneller werken is dat mensen sneller van hun klachten af zijn uh <unk> die ik nu krijgt 't is uh <unk> één keer een half jaar en uh twee maal daar is honderd milligram <unk> en nou gebruikte dieren turkse maar aan een aantal jaren sinds tweeduizend zes daarvoor heeft u ook een hele hoop medicijnen verbruid waaronder ook uh uh de <unk> uh die vond u dat hoef je in de hand dat hier ook wel goed we op een gegeven moment denk ik dat 't toch niet helemaal uh voldoende meer hielp <unk> die een bril gegaan dat ging ook wel wel wel goed maar 'k heb toch 't idee dat dit 't beste loopt dus je bent echt gewisseld van de medicijnen omdat het onvoldoende werkte op tv en had u

MODERN ASR WHISPER OUTPUT WITH PUNCTUATION AND SPEAKER DIARIZATION

Speaker 1: Kijk, hier ziet u hem zien met uw hand. Dus dit is uw rechterhand, dit is uw linkerhand. Dan ziet u hier de gevrichte van de hand en u ziet dat u eigenlijk niet een rheumatoloog moet zijn om het te zien, want die is kapot. Dus dat is deze van uw rechterhand. Ja, die is kapot. Links is ook beschadigd, want u ziet ook dat hier eigenlijk geen kraakbeerhandje meer tussen zit. Hij staat ook een beetje scheef, ze staan niet mooi recht op elkaar. Dus dat klopt, dat is dat, dat u aan de buitenkant aan uw hand ziet.

Speaker 0: Ik kan vertellen hoe de reum is begonnen. Op mijn twaalfde is dat gebeurd. Eerst leek het op colitis ulcerosa, maar er waren van alle rheumafactoren in het bloed geconcentreerd. Op mijn 21e is definitief eigenlijk vastgesteld dat het om weg tref ging. Dit was nadat ik eigenlijk mijn colitis ulcerosa, dat is verholpen door mijn dikke darm te verwijderen, waardoor die geen invloed meer had. De rheumaklachten bleven wel bestaan, voornamelijk in mijn wervelkolom.

Speaker 1: Mijn ervaringen met het voorschrijf van TNF-alpha-remmers zijn heel positief. Het belangrijkste punt daarin is dat ze goed werken. Dat wil zeggen dat mensen zelf minder pijn ervaren en dat wij als dokter ook minder ontstekingen aan de gewrichten zien. Het voordeel van het gebruik van biologicals ten opzichte van de conventionele middelen is dat ze voor een groep patiënten die niet reageert op conventionele middelen, dus beter werken. Daarnaast is het zo dat patiënten ook aanreven dat ze vaak sneller werken. Dus dat mensen sneller van hun klachten af zijn.

WHISPER OUTPUT WITH TIMESTAMP AND SPEAKER DIARIZATION

```
1
00:00:06,780 --> 00:00:09,989
Speaker 1: Kijk, hier ziet u hem zien met uw hand.

2
00:00:10,009 --> 00:00:13,031
Speaker 1: Dus dit is uw rechterhand, dit is uw linkerhand.

3
00:00:13,071 --> 00:00:20,815
Speaker 1: Dan ziet u hier de gevrichte van de hand en u ziet dat u eigenlijk niet een
rheumatoloog moet zijn om het te zien, want die is kapot.

4
00:00:20,855 --> 00:00:25,338
Speaker 1: Dus dat is deze van uw rechterhand.
```

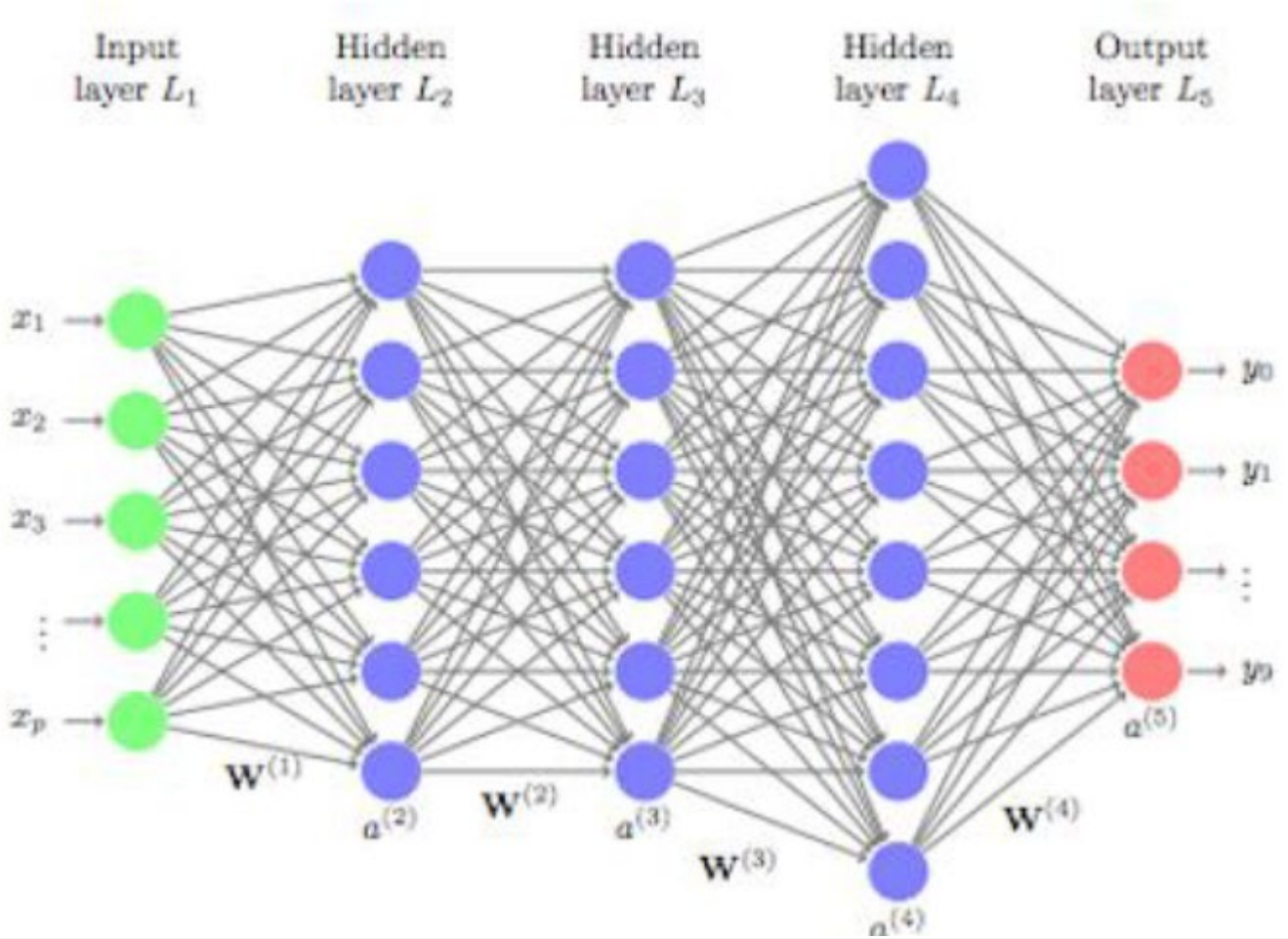
SRT (SubRip Subtitle file) output:
suitable for subtitling videos, and time aligned research

NLP & MORE

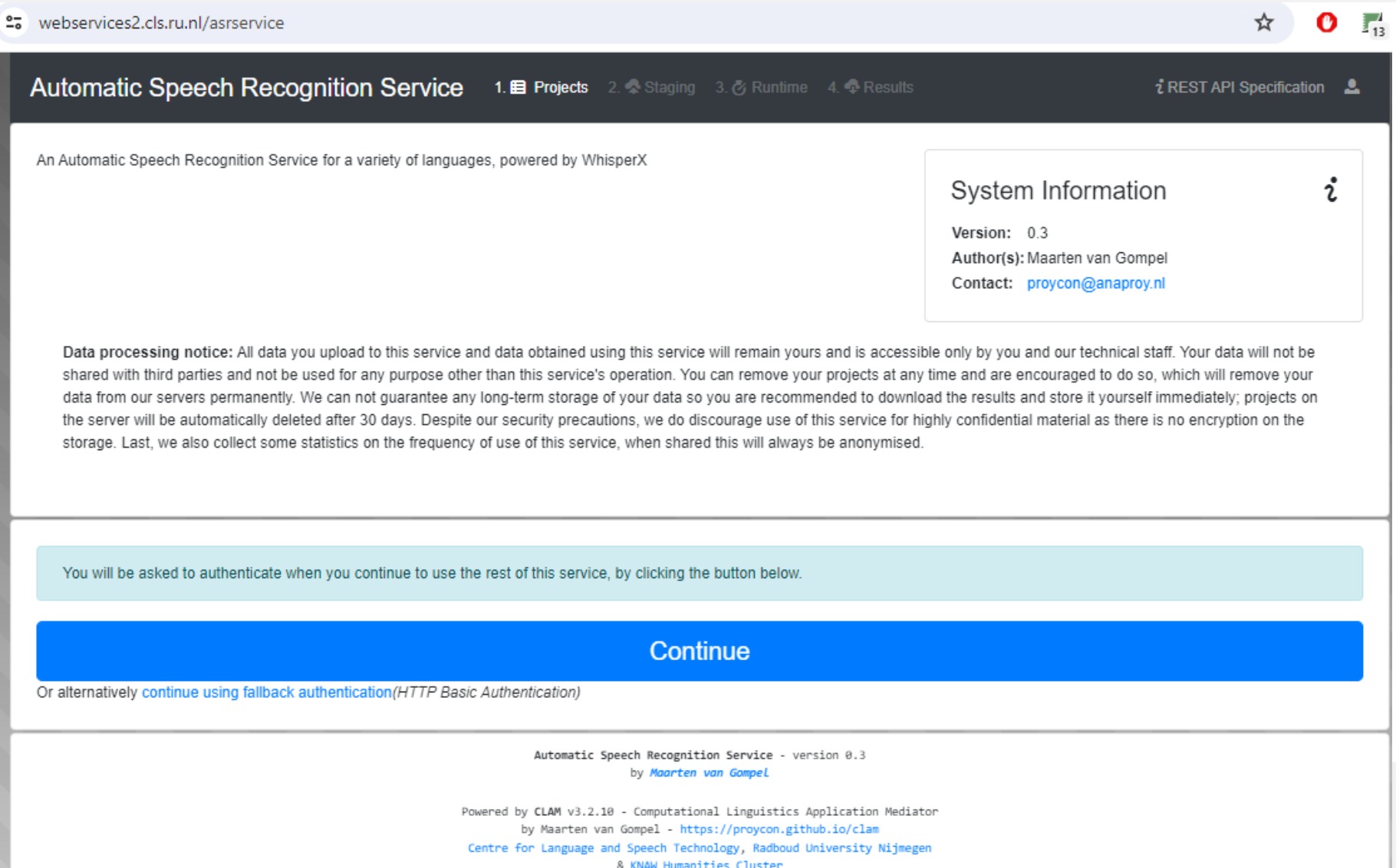
- Text Summarisation
- Topic modelling
- Sentiment analysis
- Text simplification
- Text generation
- Translation
- Literature review (wider coverage)

LLM finetuning of LSM output:

- Conversational prompting
- Prompt design
- Prompt engineering



2. WEBSERVICES AT RADBOUD UNIVERSITY: HTTPS://WEBSERVICES2.CLS.RU.NL/ASRSERVICE



- Login (CLARIN)
- -Create project
- Upload audio file(s)

Input

Input files

Show 10 entries

Search:

Input File	Template	Format	Actions
KNMP2013najr.wav	Wav audio file	Wave Audio File	

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Upload a file from disk

Parameters

Global

Language
The language to recognize
Dutch / Nederlands

Model
The ASR model to use
large-v2

GPU
Use GPU (improves performance but may not always be available)


Diarization

Diarization
Enable speaker diarization?


Minimum speakers
Minimum number of speakers (this helps diarization)
2

Maximum speakers
Minimum number of speakers (this helps diarization)
3

Start

Done.
Processing files
Starting...

Cancel and delete project

Discard output and restart

Show input files

Output files

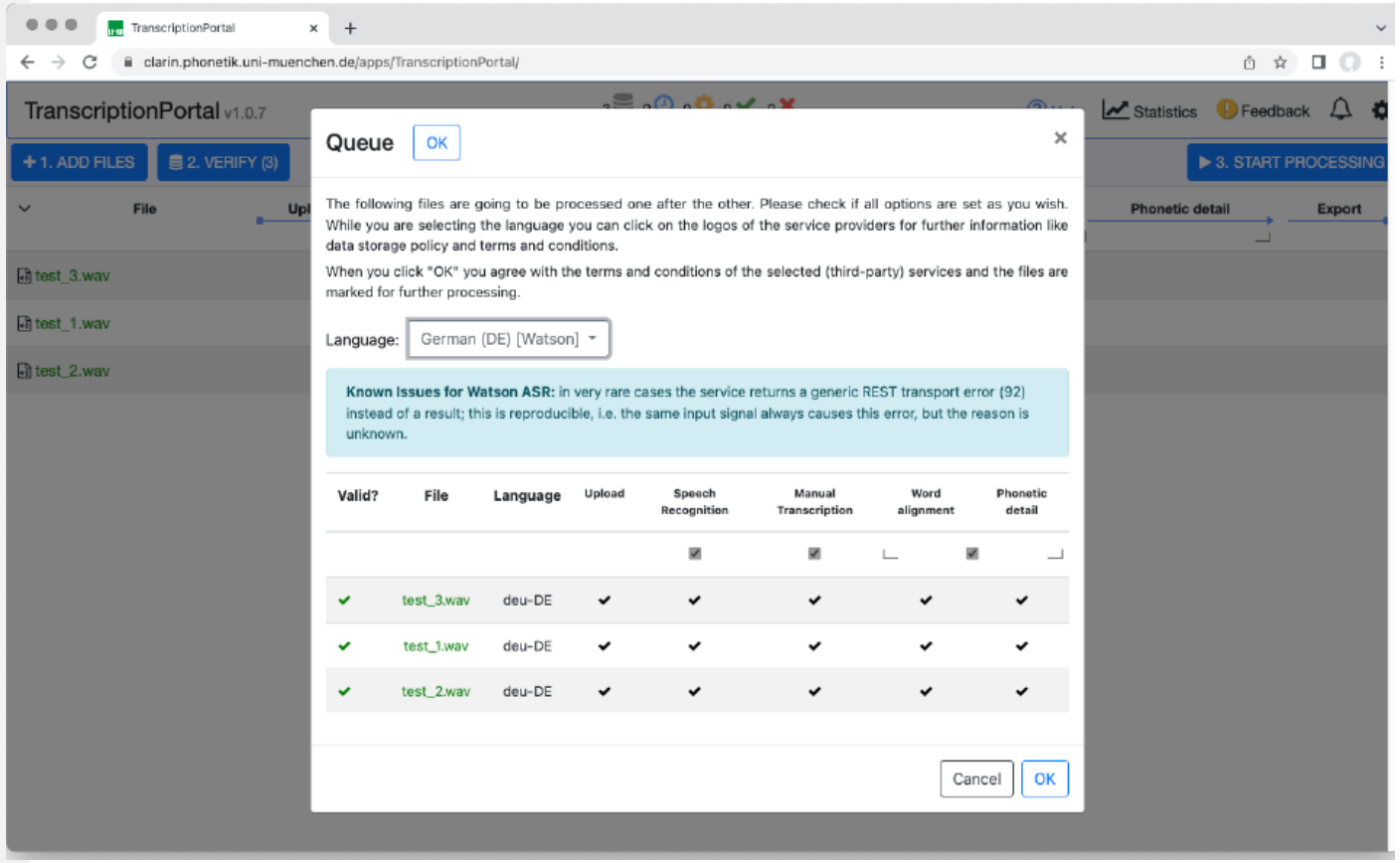
(Download all as archive: zip | tar.gz | tar.bz2)

Show 10 entries

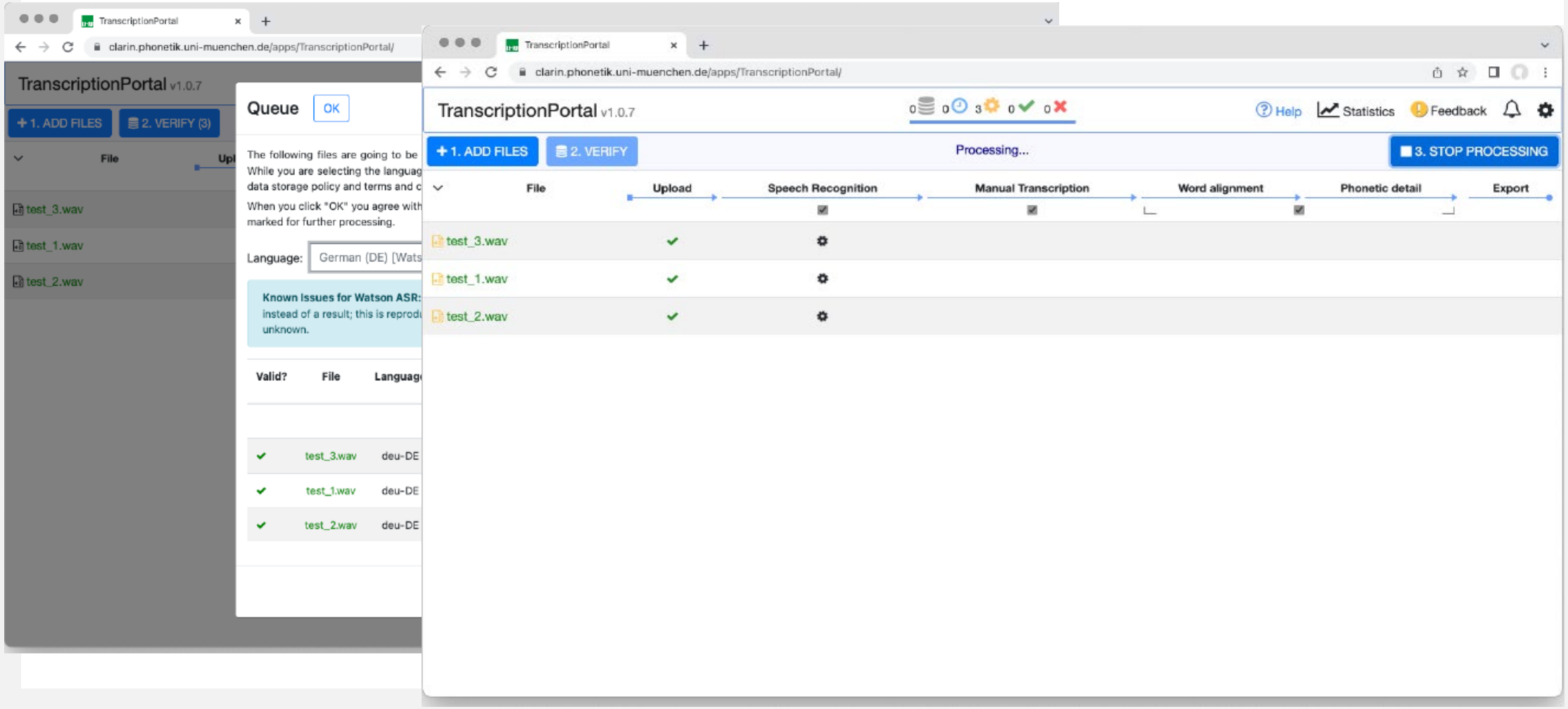
Search:

Output File	Template	Format	Viewers
error.log	Log file with (standard) error output	Plain Text Format	Download | *More...*
KNMP2013najr.ctm	Transcription with full word segmentation/alignment	Conversation Time Marked File	Download | *More...*
KNMP2013najr.json	Transcription with full word segmentation/alignment and speaker attribution	JSON Format (generic, not further specified)	Download | *More...*
KNMP2013najr.srt	Timed transcriptions with speaker attribution (srt)	SubRip Text	Download | *More...*
KNMP2013najr.tsv	Timed transcriptions with speaker attribution (TSV)	Tab Separated Values	Download | *More...*
KNMP2013najr.txt	Plain text transcriptions without time stamps and speaker attribution	Plain Text Format	Download | *More...*
KNMP2013najr.vtt	Timed transcriptions with speaker attribution	WebVTT	Download | *More...*

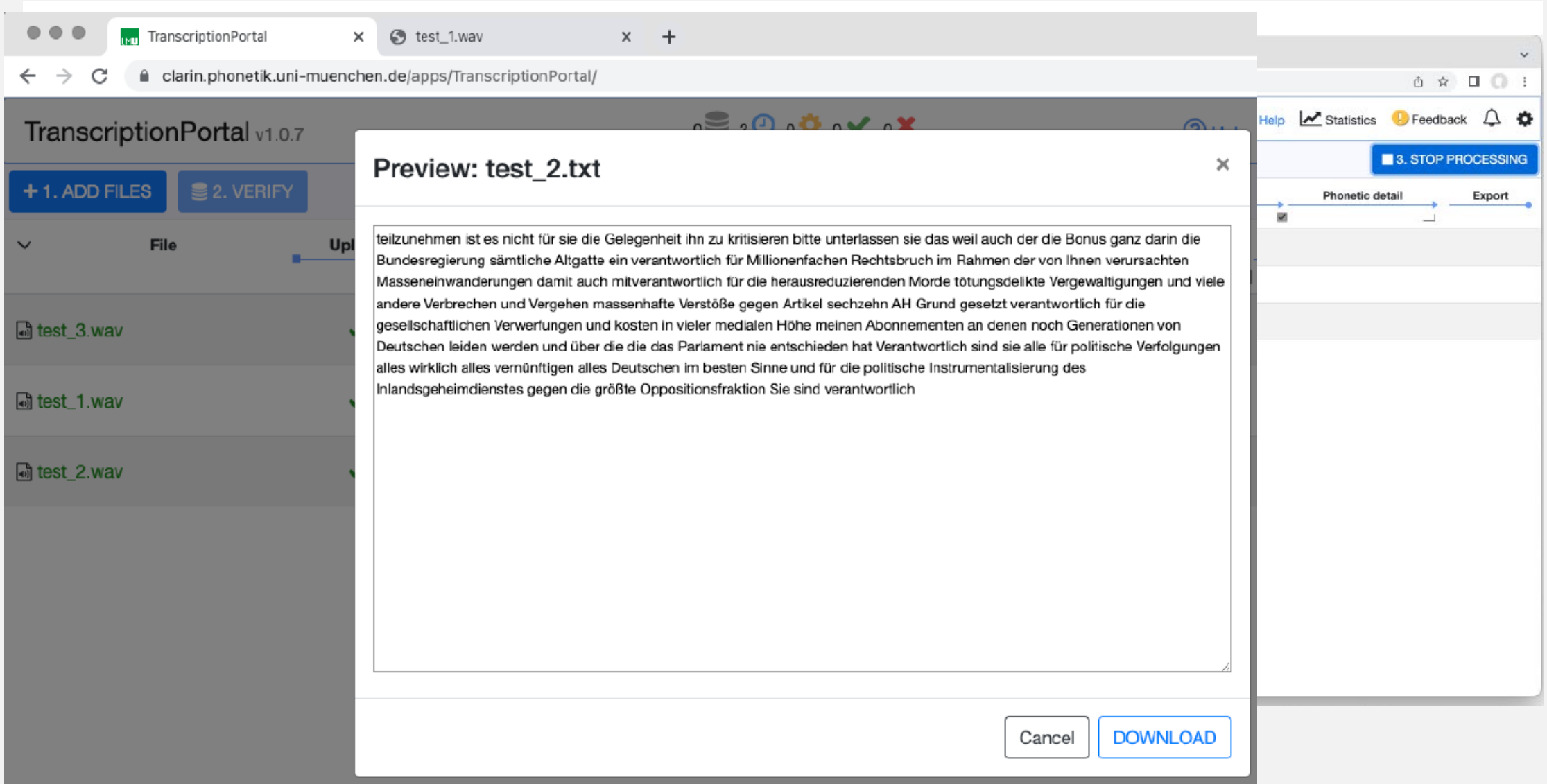
3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL



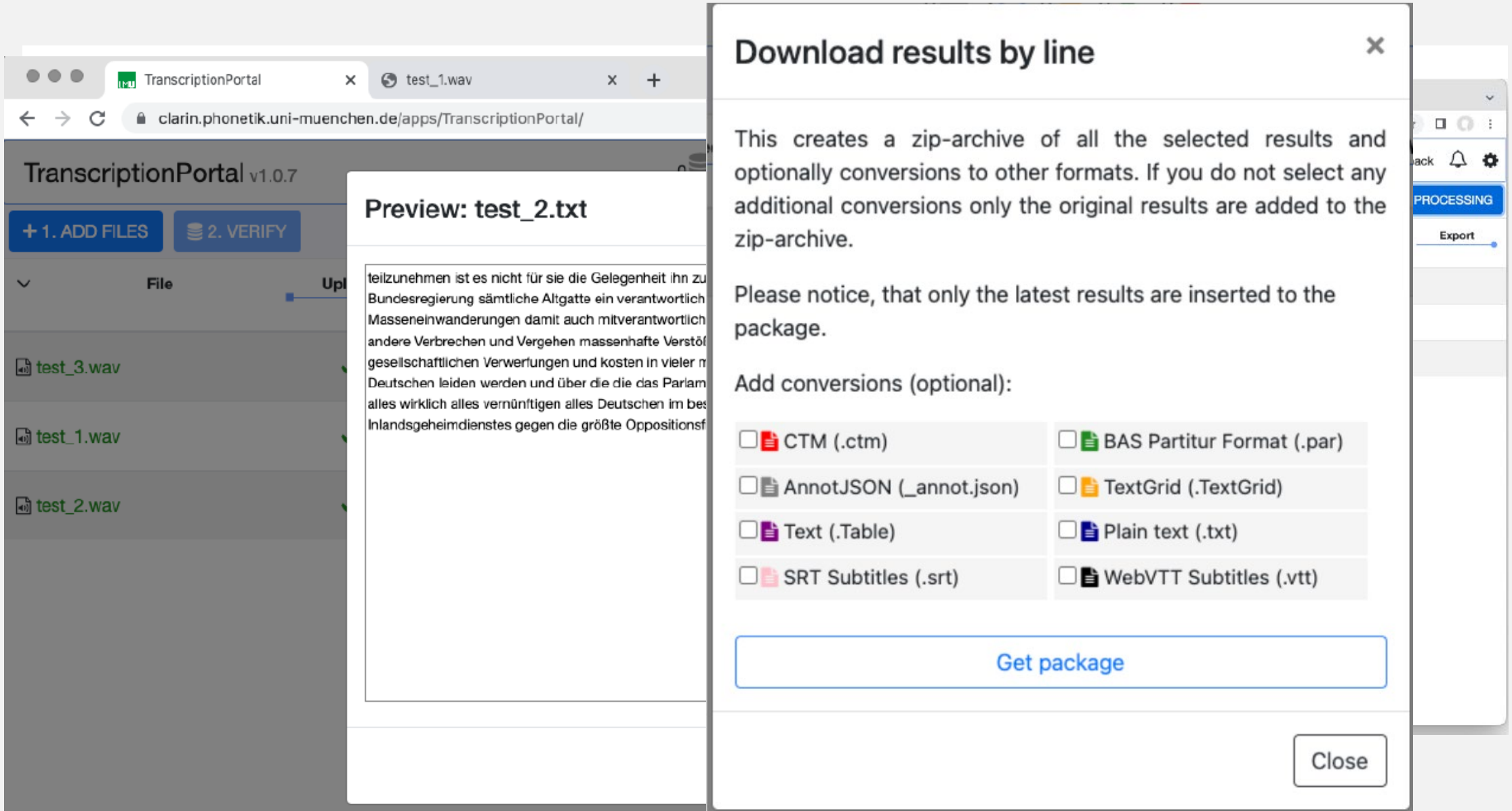
3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL



3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL

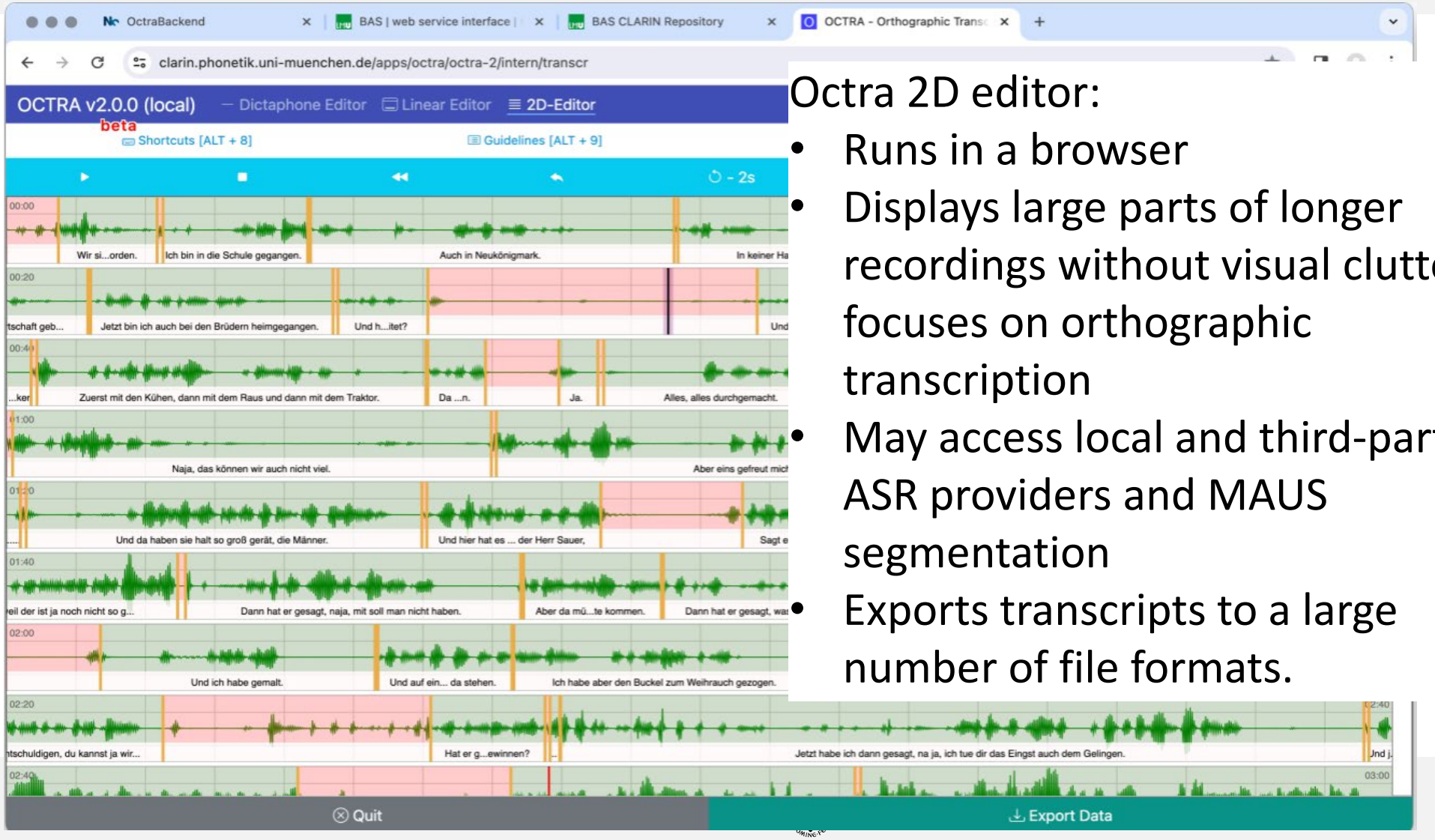


3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL



3. WEBSERVICES AT BAS: OCTRA EDITOR

3. WEBSERVICES AT BAS: OCTRA EDITOR

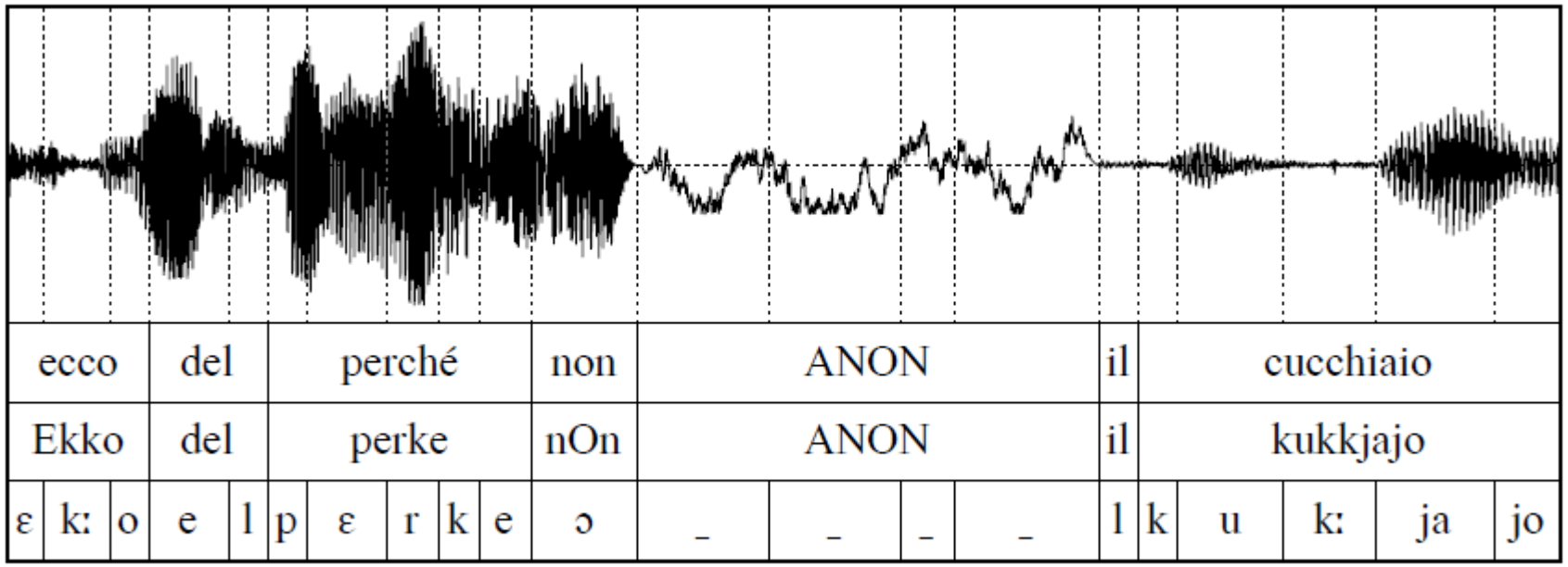


Octra 2D editor:

- Runs in a browser
- Displays large parts of longer recordings without visual clutter focuses on orthographic transcription
- May access local and third-party ASR providers and MAUS segmentation
- Exports transcripts to a large number of file formats.

3. WEBSERVICES AT BAS: ANONYMIZER

web service Anonymizer: removes words from a stop list from the transcript (on all levels) and replaces the corresponding fragment in the audio with a noise



ecco			del		perché					non	ANON				il	cucchiaio				
Ekko			del		perke					nOn	ANON				il	kukkjajo				
ɛ	kː	o	e	l	p	ɛ	r	k	e	ɔ	_	_	_	_	l	k	u	kː	ja	jo

3. WEBSERVICES AT BAS: OTHER SERVICES

- ChannelSeparator separates the individual channels of a stereo recording and retains in each channel only the voice of the dominant speaker
- G2P (Grapheme to phoneme) converts an orthographic text to its phonemic representation. The service allows a customized specification of pronunciation rules for vernacular language, dialects, and common coarticulation phenomena (e. g. ‘haben wir’ (we have) → /hamva/ or /hama/ in German). These rules improve the performance of automatic word alignment
- WebMAUS (**M**unich **au**tomatic **s**egmentation) aligns an orthographic transcript in one of the available languages and regional variants) with the audio signal

4. DO IT YOURSELF WITH WHISPER: GENERAL

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. Whisper can recognise more than 99 different languages.

To download Whisper, go to their Github-page:
https://github.com/openai/whisper/blob/main/README.md

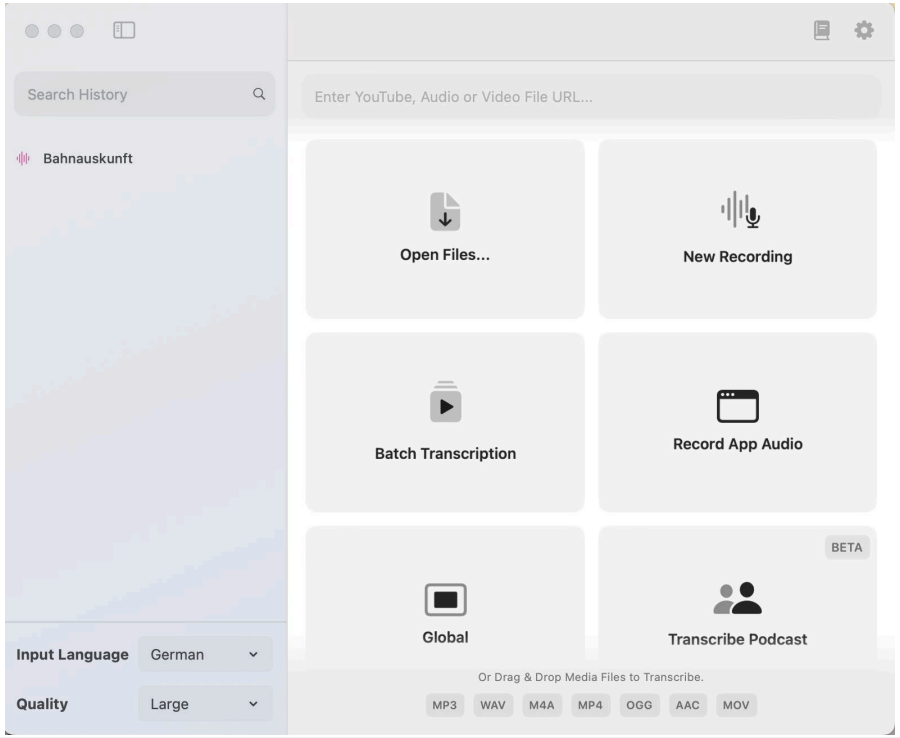
Because it is Open Source, other people can modify and add their own “additions” like Diarization & VAD (WhisperX) and can decrease the working time (Faster-Whisper).

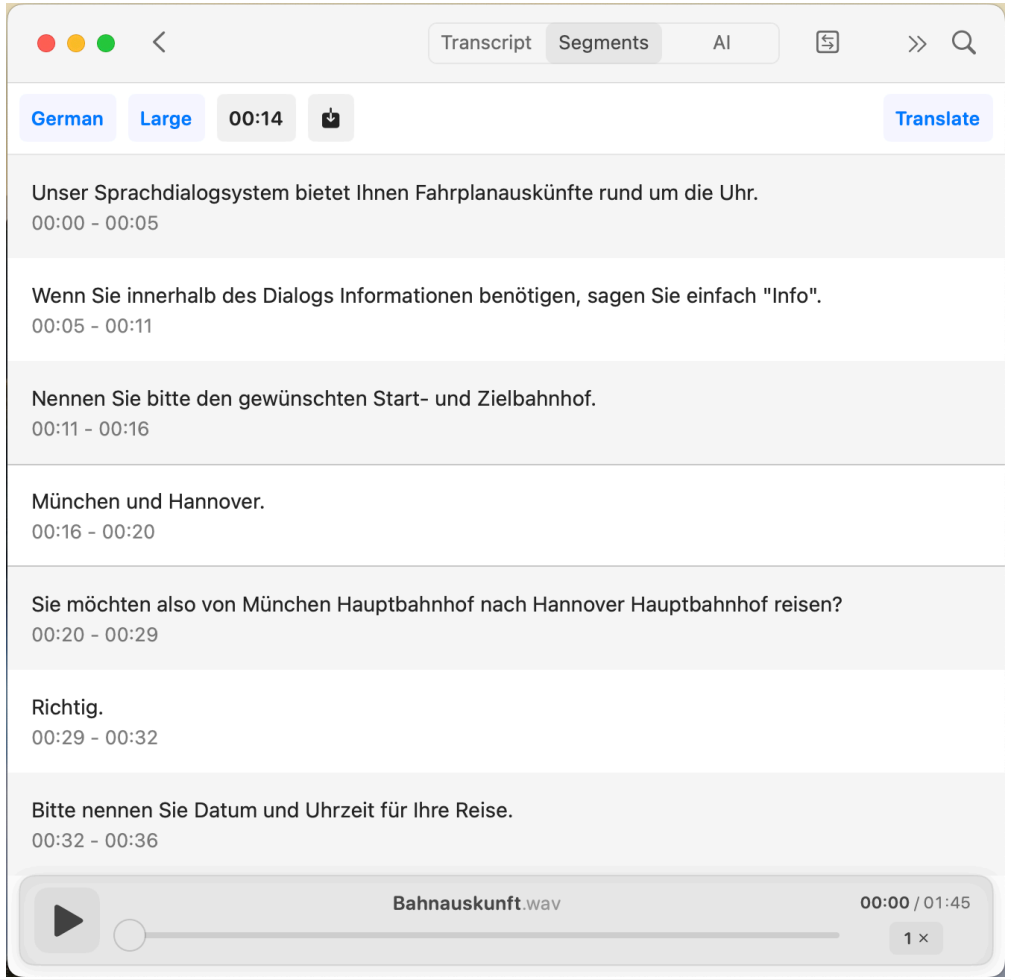
Moreover, Georgi Gerganov developped a C++ version that can be used in stand-alone software packages like **MacWhisper** (Apple), and **aTrain** (Windows).

A more detailed description of Whisper and its derivatives, can be found at
https://speechandtech.eu/news

4. DO IT YOURSELF WITH WHISPER: MACWHISPER

MacWhisper is a simple but complete ASR package that enables the recognition and correction of AV-files.



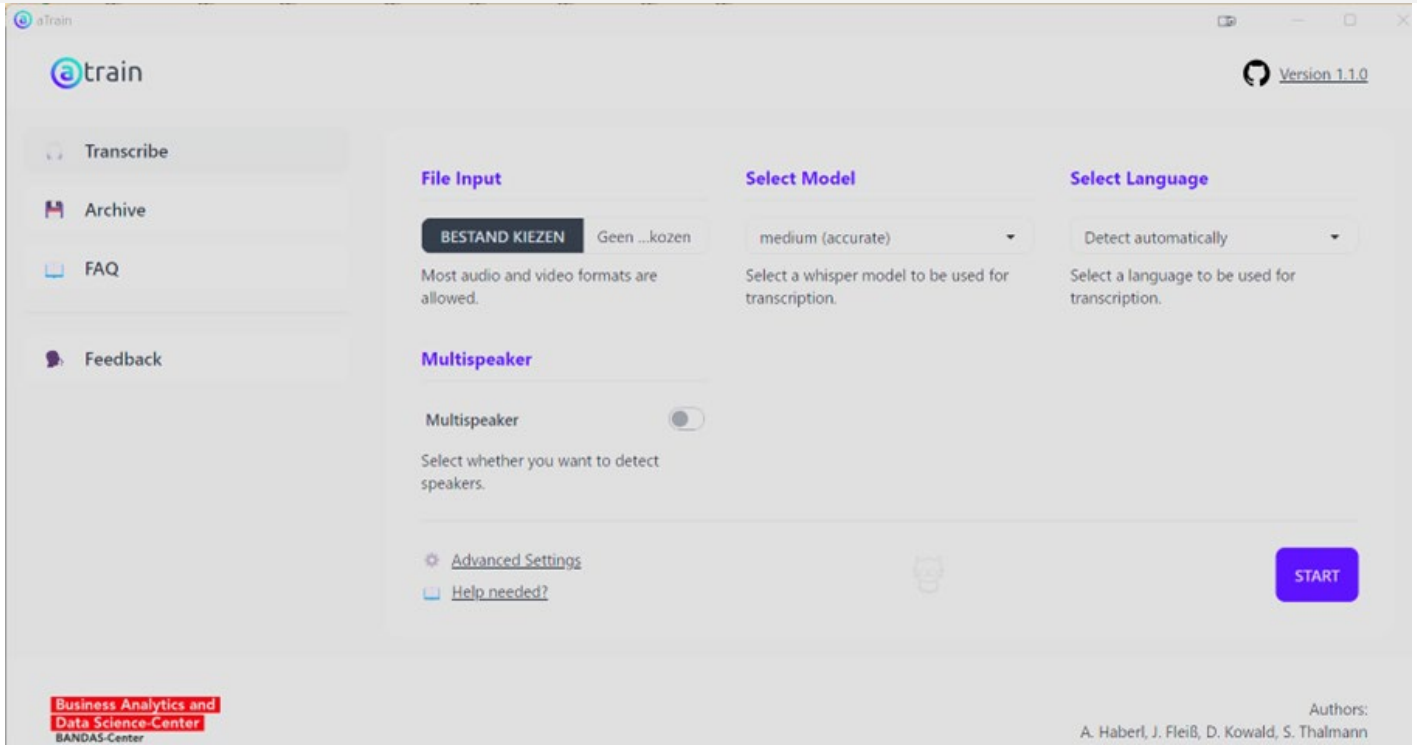


4. DO IT YOURSELF WITH WHISPER: ATRAIN

ICT people from the Universität of Graz developed aTrain: an ASR for Windows and Linux.

Whereas MacWhisper only does what Whisper can do, aTrain is more comprehensive and incorporates several new developments used today.

For example, Diarization can be done, and the programme is very fast by using Fast-Whisper.



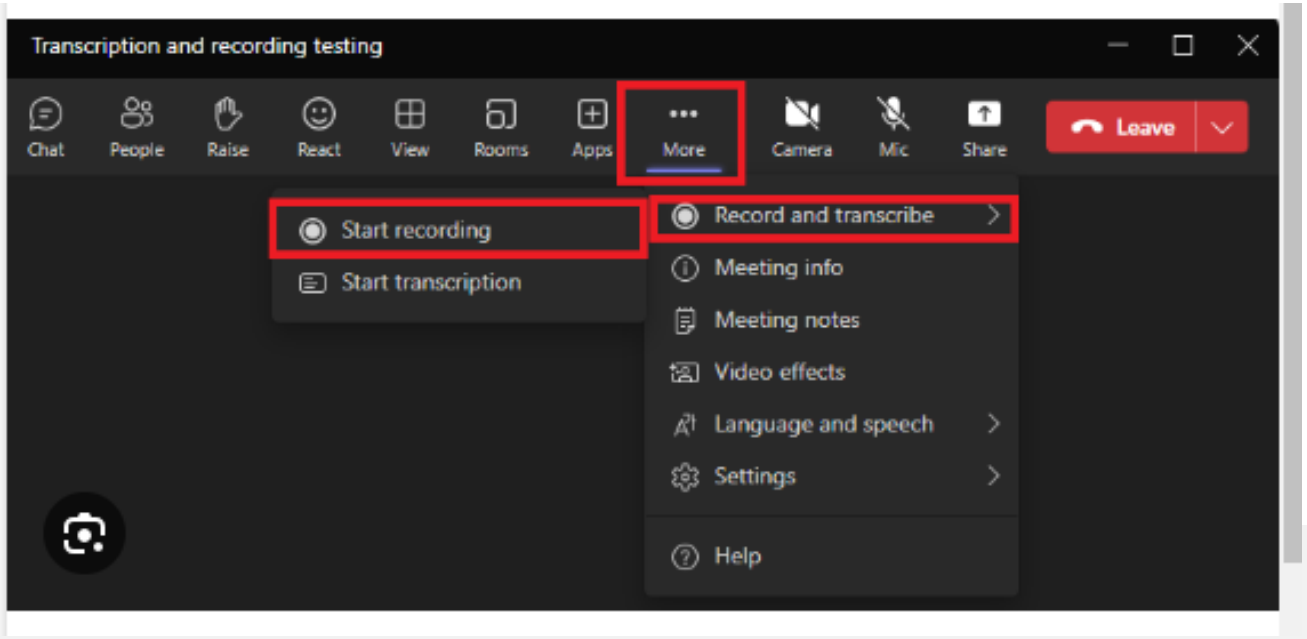
aTrain can be downloaded (>10GB) in the Microsoft store: https://apps.microsoft.com/detail/9n15q44szns2

5. COMMERCIAL SOLUTIONS

Companies offering transcription services, e.g.

- Amberscript, Sonix, Scriptoman, Happy Scribe, Autoscriber
- Transcription solutions in Zoom and MS Teams for online interviews / conversations / focus groups

Special arrangements / licenses needed for sensitive recordings



6. REMAINING CHALLENGES

What remains under the ASR radar:

- Filled pauses and back channels
- Repetitions with word truncations
- Self-repairs
- Pronunciation peculiarities

However, these phenomena are very relevant for specific research domains

6. REMAINING CHALLENGES

Problematic conditions:

- Background noise
- Speaker overlap
- Dialect speech
- Children & older people
- Speakers with language impairments / speech disorders