

Authors: Aleš Pražák, Luboš Šmídl

Title: Czech Parliament Meetings

Identifiers: ZCU_CZ_Parliament

Subject keywords: speech corpus; acoustic model; speaker identification and verification

Sponsors:

Description:

The corpus consists of recordings from the Chamber of Deputies of the Parliament of the Czech Republic. It currently consists of 88 hours of speech data, which corresponds roughly to 0.5 million tokens. The annotation process is semi-automatic, as we are able to perform the speech recognition on the data with high accuracy (over 90%) and consequently align the resulting automatic transcripts with the speech. The annotator's task is then to check the transcripts, correct errors, add proper punctuation and label speech sections with information about the speaker. The resulting corpus is therefore suitable for both acoustic model training for ASR purposes and training of speaker identification and/or verification systems.

File description:

Speech data and corresponding annotations

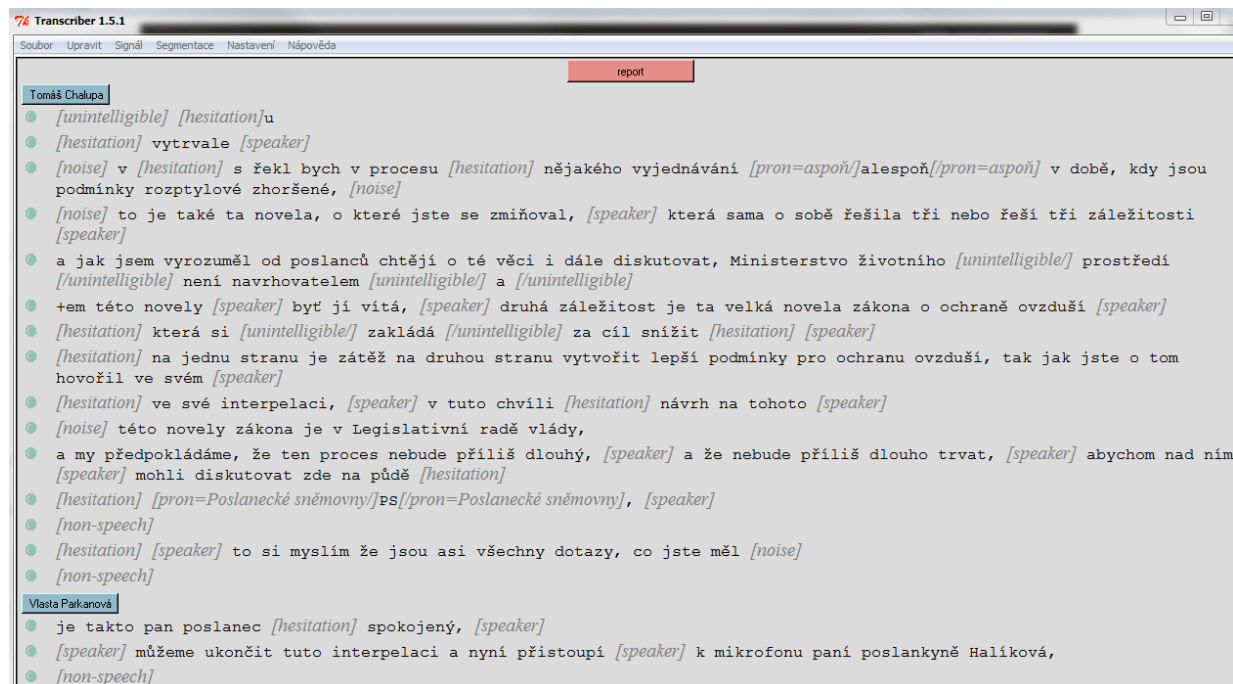
The archive contains 18 sound files (WAV PCM, 16-bit, 44.1 kHz, mono) and corresponding transcriptions in XML-based standard Transcriber format (<http://trans.sourceforge.net>)

Annotation conventions description:

Besides word-by-word transcription of the speech, the following non-speech events are also captured in the annotation:

- *[speaker]* – noises produced by a speaker (loud breath, coughing, lip smacking, sneezing, etc.)
- *[hesitation]* – fillers stemming from speaker's hesitation ("Er" and "Uhm", etc.)
- *[noise]* – other environmental noises that significantly stand out from the background noise level (knocking, rustling, car horn sounds, etc.)
- *[non-speech]* – whole segment containing only silence, noise, music and/or other non-speech events
- *[background-speech]* – overlapping speech from additional speaker(s)
- *[unintelligible]* – unintelligible segment of speech
- *[pron=pronunciation]word[/pron=pronunciation]* – for recording non-standard (unexpected) pronunciations of words enclosed within the tag. Also used for writing down the full word forms of abbreviations, acronyms and various special symbols (like section sign §)

Example of the annotation in the Transcriber tool



The screenshot shows the Transcriber 1.5.1 application window. The title bar reads "74 Transcriber 1.5.1". The menu bar includes "Soubor", "Upravit", "Signál", "Segmentace", "Nastavení", and "Nápověda". A "report" button is visible in the top right. The main area displays a transcript with phonetic annotations. The speaker is identified as "Tomáš Chalupa".

report

Tomáš Chalupa

- [unintelligible] [hesitation]u
- [hesitation] vytrvale [speaker]
- [noise] v [hesitation] s řekl bych v procesu [hesitation] nějakého vyjednávání [pron=aspoň]alespoň[pron=aspoň] v době, kdy jsou podmínky rozptylové zhoršené, [noise]
- [noise] to je také ta novela, o které jste se zmiňoval, [speaker] která sama o sobě řešila tři nebo řeší tři záležitosti [speaker]
- a jak jsem vyrozuměl od poslanců chtějí o té věci i dále diskutovat, Ministerstvo Životního [unintelligible] prostředí [unintelligible] není navrhovatelem [unintelligible] a [unintelligible]
- +em této novely [speaker] byť jí vítá, [speaker] druhá záležitost je ta velká novela zákona o ochraně ovzduší [speaker]
- [hesitation] která si [unintelligible] zakládá [unintelligible] za cíl snížit [hesitation] [speaker]
- [hesitation] na jednu stranu je zátěž na druhou stranu vytvořit lepší podmínky pro ochranu ovzduší, tak jak jste o tom hovořil ve svém [speaker]
- [hesitation] ve své interpelaci, [speaker] v tuto chvíli [hesitation] návrh na tohoto [speaker]
- [noise] této novely zákona je v Legislativní radě vlády,
- a my předpokládáme, že ten proces nebude příliš dlouhý, [speaker] a že nebude příliš dlouho trvat, [speaker] abychom nad ním [speaker] mohli diskutovat zde na půdě [hesitation]
- [hesitation] [pron=Poslanecké sněmovny]PS[pron=Poslanecké sněmovny], [speaker]
- [non-speech]
- [hesitation] [speaker] to si myslím že jsou asi všechny dotazy, co jste měl [noise]
- [non-speech]

Vlasta Parkanová

- je takto pan poslanec [hesitation] spokojený, [speaker]
- [speaker] můžeme ukončit tuto interpelaci a nyní přistoupí [speaker] k mikrofonu paní poslankyně Halíková,
- [non-speech]