Database of Spoken Czech as a Foreign Language (Permanent Residency in the Czech Republic)

Database of Spoken Czech as a Foreign Language (Permanent Residency in the Czech Republic) is the language corpus of spoken performances by non-native speakers of Czech focused on A2 level (according to the CEFR), which is required for the granting of permanent residency in the Czech Republic. It includes recordings capturing the oral part of the Czech Language Certificate Exam. The recordings consist of dialogues between the examiner (a native speaker) and the candidate (a non-native speaker). In addition to the recordings, the corpus also contains their transcriptions, which are richly linguistically annotated. Some recordings are accompanied by multiple transcriptions from different annotators, allowing for comparisons of various transcripts of the same recording and evaluations of the degree of consistency in converting spoken language into written text.

The corpus is published as a specialized public database aimed at providing a structured and easily accessible source of authentic spoken data for linguists, educators, students, the scientific community, and the general public.

The corpus was created at the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics, Charles University to support teaching, research, and assessment of language competence among non-native speakers of Czech as part of the project *Automated Speech Scoring in Czech*. Audio recordings were provided by the Institute for Language and Preparatory Studies, Charles University (ujop.cuni.cz).

Statistics

The database contains 63 recordings. It captures 41 exams and the same number of non-native speakers. The total length of all recordings is 3h 18min 40s. The table below shows the transcription statistics, with one canonical transcription selected for each recording.

	All	Canonical
Files	106	63
Utterances	4,773	2,888
Tokens	33,267	20,035

Documentation

- User Manual
- Technical Documentation

License

Acknowledgment

The database was funded by the Programme to Support Applied Research in the Area of the National and Cultural Identity for the Years 2023 to 2030 (NAKI III) of the Ministry of Culture of the Czech Republic within the project *Automated Speech Scoring in Czech* (DH23P03OVV037).

Special Thanks

The authors of the database sincerely thank PhDr. Pavel Pečený, Ph.D., from the Institute for Language and Preparatory Studies, Charles University for providing audio data.

How to Cite

Rysová Kateřina, Novák Michal, Rysová Magdaléna, Polák Peter, Bojar Ondřej: *Database of Spoken Czech as a Foreign Language (Permanent Residency in the Czech Republic)*. Institute of Formal and Applied Linguistics MFF UK, Prague 2024. Available from WWW https://lindat.mff.cuni.cz/services/teitok-live/evaldio/en/index.php?action=db_residency.