Databáze mluvených projevů v češtině jako cizím jazyce (trvalý pobyt v ČR)

Databáze mluvených projevů v češtině jako cizím jazyce (trvalý pobyt v ČR) je jazykový korpus mluvených projevů nerodilých mluvčích češtiny zaměřený na jazykovou úroveň A2 (podle SERR), požadovanou pro udělení trvalého pobytu v České republice. Obsahuje nahrávky zaznamenávající ústní část Certifikované zkoušky z češtiny pro cizince. Nahrávky zahrnují dialogy mezi zkoušejícím (rodilým mluvčím) a kandidátem zkoušky (nerodilým mluvčím). Kromě nahrávek korpus obsahuje také jejich přepisy, které jsou opatřeny bohatou lingvistickou anotací. K některým nahrávkám je připojeno více přepisů od různých anotátorů, což umožňuje srovnání různých přepisů téže nahrávky a vyhodnocení míry shody při převodu mluvené řeči do psaného textu.

Korpus je zveřejněn jako specializovaná veřejná databáze s cílem poskytnout strukturovaný a snadno přístupný zdroj autentických mluvených dat pro lingvisty, pedagogy, studenty, vědeckou komunitu a širokou veřejnost.

Jazykový korpus byl vytvořen v Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy za účelem podpory výuky, výzkumu a hodnocení jazykové kompetence nerodilých mluvčích češtiny v rámci projektu *Automatické hodnocení mluveného projevu v češtině*. Audionahrávky poskytl Ústav jazykové a odborné přípravy Univerzity Karlovy (ujop.cuni.cz).

Statistiky

Databáze obsahuje 63 nahrávek. Zachycuje 41 zkoušek a stejný počet nerodilých mluvčích. Celková délka všech nahrávek je 3h 18min 40s. Tabulka níže ukazuje statistiky přepisů, přičemž pro každou nahrávku byl vybrán právě jeden kanonický přepis.

	Všechny	Kanonické
Soubory	106	63
Repliky	4 773	2 888
Tokeny	33 267	20 035

Dokumentace

- Uživatelská příručka
- Technická dokumentace

Licence

Financování

Vznik databáze byl financován z prostředků Programu na podporu aplikovaného výzkumu v oblasti národní a kulturní identity na léta 2023 až 2030 (NAKI III) Ministerstva kultury ČR v rámci projektu *Automatické hodnocení mluveného projevu v češtině* (DH23P03OVV037).

Poděkování

Autoři databáze srdečně děkují PhDr. Pavlovi Pečenému, Ph.D., z Ústavu jazykové a odborné přípravy Univerzity Karlovy za poskytnutí audiodat.

Jak citovat

Rysová Kateřina, Novák Michal, Rysová Magdaléna, Polák Peter, Bojar Ondřej: *Databáze mluvených projevů v češtině jako cizím jazyce (trvalý pobyt v ČR)*. Ústav formální a aplikované lingvistiky MFF UK, Praha 2024. Dostupná z WWW https://lindat.mff.cuni.cz/services/teitok-live/evaldio/cs/index.php?action=db_residency.