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

Základní funkce databáze zahrnuje prohlížení záznamů s různými způsoby jejich zobrazení, filtrování záznamů podle různých kategorií a komplexní vyhledávání v obsahu databáze. Databáze rovněž umožňuje stáhnout korpus jako celek nebo stáhnout vybrané záznamy.

Prohlížení záznamů

Po vstupu do korpusu se v přehledné tabulce zobrazí všechny záznamy (tj. soubory transkriptů) uložené v databázi. Pro každý soubor s transkriptem tabulka kromě názvu souboru zobrazuje v dalších sloupcích úroveň a identifikátor zkoušky, číslo úlohy, zdroj předběžné anotace, kód anotátora a informaci o tom, zda je přepis pro danou nahrávku kanonický. Soubory v tabulce je možné třídit podle hodnot vybraného sloupce. Záznamy lze také filtrovat na základě libovolného podřetězce v názvu souboru zadáním tohoto podřetězce do textového pole "Search" umístěného vpravo nad tabulkou. Kliknutím na konkrétní soubor se tento soubor zobrazí.

Zobrazení souboru

Databáze umožňuje prohlížet přepisy jednotlivých replik spolu s anotacemi a metadaty a také poslouchat příslušné zvukové nahrávky. Charakter zobrazených informací se liší podle zvoleného režimu zobrazení, mezi kterými lze přepínat v dolní části stránky pod samotným přepisem.

Režim Text View

Text View je základní režim zobrazení, který se objeví po otevření souboru. V horní části obrazovky se nachází hlavička s názvem přepisu a vybranými metadaty. V dolní části je zobrazen samotný přepis, rozdělený na repliky. Každá replika je označena identifikátorem mluvčího (EXAM_1 pro zkoušejícího a CAND_1 pro kandidáta).

Tento režim rovněž umožňuje zobrazit automatickou morfologickou anotaci a lemmatizaci. Po najetí kurzorem na konkrétní token se zobrazí příslušná anotace v kontextu. Pro zobrazení vybraného atributu pro všechny tokeny v přepisu lze využít ovládací prvky umístěné pod hlavičkou, které obsahují následující tlačítka: - PoS: Zobrazí slovní druhy. - Tag: Ukáže morfologické tagy. - Features: Poskytne podrobné morfologické informace. - Lemma: Zobrazí základní tvary slov.

Režim Waveform View

V horní části obrazovky se nachází rozšířený ovládací prvek pro přehrávání nahrávky, který zobrazuje graf signálu (tzv. waveform). Pod ním jsou zobrazeny přepisy jednotlivých replik. Kliknutím na konkrétní repliku se tato replika přehraje.

Režim Dependencies

Tento režim zobrazuje syntaktickou anotaci. Po kliknutí na konkrétní repliku se zobrazí automaticky vygenerovaný závislostní strom, u nějž je možné zobrazit detaily pomocí myši. Vpravo nahoře od stromu se nachází tlačítko ≡ pro další možnosti zobrazení stromu. Je tak možné uspořádat uzly podle slovosledu, zobrazit interpunkci nebo uložit obrázek stromu ve formátu SVG.

Filtrování záznamů přes kategorie

Po kliknutí na tlačítko *Kategorie* v levém hlavním menu je možné filtrovat přepisy na základě hodnot jednotlivých kategorií. Například je tak možné zobrazit si pouze seznam kanonických přepisů nebo přepisů od konkrétního anotátora.

Vyhledávání

Vyhledávání v korpusu lze provádět na stránce, která se zobrazí po klinutí na tlačítko *Hledat* v levém hlavním menu. Stránka umožňuje zadávat dotazy ve formátu CQL (Corpus Query Language). Např.

```
[upos = "NUM.*"] [lemma = "otázka"]
```

pro nalezení tvarů slova *otázka*, jimž předchází číslovka.

Pro usnadnění vyhledávání nabízí rozhraní TEITOK nástroj pro sestavování dotazů. Tento nástroj umožňuje snadno definovat jednoduché dotazy v CQL prostřednictvím formuláře. Stačí kliknout na ikonu *Query builder*, definovat svůj dotaz a poté stisknout tlačítko *Create query*, čímž se dotaz vloží do textového pole CQL, kde jej můžete případně upravit.

V základním nastavení TEITOK provádí vyhledávání v celém korpusu, který může obsahovat k jedné nahrávce více přepisů. Pokud chcete vyhledávat pouze v té části korpusu, v níž je ke každé nahrávce přiřazený jen jediný přepis, je nutné omezit hledání na tzv. kanonické přepisy. Např.

```
[lemma = "situace"] :: match.text_canonical = "1"
```

vyhledává lemma *situace* jenom v kanonických přepisech.

Stahování

Celý korpus včetně nahrávek a dokumentace je možné stáhnout z hlavního menu vlevo.

Konkrétní přepis lze stáhnout v režimu *Text view* kliknutím na tlačítko *Download XML* umístěné v dolní části stránky.

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.

Dedikace

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).