Technická dokumentace

Jazykový korpus mluvených projevů nerodilých mluvčích češtiny zaměřený na jazykovou úroveň A2 (podle SERR), požadovanou pro udělení trvalého pobytu v České republice, je výsledkem projektu realizovaného v Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Univerzity Karlovy. Korpus obsahuje nahrávky zaznamenávající ústní část Certifikované zkoušky z češtiny pro cizince na úrovni A2. Nahrávky zahrnují dialogy mezi zkoušejícím (rodilým mluvčím) a kandidátem zkoušky (nerodilým mluvčím). Náhravky jsme opatřili jejich přepisy a 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 a je volně dostupný široké veřejnosti, vědecké komunitě, pedagogům a studentům. Databáze je integrována do systému TEITOK, který je spravován na platformě LINDAT/CLARIAH-CZ.

TEITOK

TEITOK je framework pro vytváření, správu a zveřejňování anotovaných korpusů. Jeho webové rozhraní je implementováno v kombinaci jazyků PHP a JavaScript. Pro náš projekt, který kombinuje nahrávky mluveného projevu a jejich přepisy, je stěžejní funkcionalita prostředí TEITOK, která umožňuje vytvářet, zobrazovat a upravovat přepisy nahrávek. K práci se samotnou nahrávkou TEITOK využívá Javascript knihovnu wavesurfer.

Uložení dat

Data korpusu jsou v prostředí TEITOK primárně uložena ve formě souborů. V tomto případě se jedná o nahrávky ve formátu MP3, hlavní části jsou však soubory ve formátu TEITOK, které obsahují všechny přepisy a anotace včetně metadat. Tyto soubory jsou navzájem provázány s odpovídajícími nahrávkami.

Struktura souborů TEITOK

Formát TEITOK je formát XML, který plně odpovídá standardu Text Encoding Initiative (TEI), avšak s mírně odlišným přístupem k tokenizaci. Struktura TEITOK souborů v naší databázi je následující:

Hlavička s metadaty `<teiHeader>`

1. `<fileDesc>` – Popis souboru
   - `<titleStmt>`: Obsahuje název souboru a informace o autorech a anotátorech.
   - `<editionStmt>`: Obsahuje číslo verze.
   - `<publicationStmt>`: Publikační detaily, jako je vydavatel, datum vydání a licence.
   - `<sourceDesc>`: Popis zdrojové nahrávky a odkaz na ni.
2. `<encodingDesc>` – Popis kódování
   - `<projectDesc>`: Stručný popis projektu, v rámci něhož data vznikla.
   - `<annotationDecl>`: Detaily o jednotlivých krocích anotace (primární, revize, lingvistická anotace).
3. `<profileDesc>` – Profil textu
   - `<langUsage>`: Použitý jazyk (čeština).
   - `<textClass>`: Metadata dokumentu:
     - `database`: Název databáze.
     - `exam-id`: Identifikátor zkoušky.
     - `cefr-level`: Úroveň podle SERR. Tato databáze obsahuje výhradně nahrávky zkoušek úrovně A2.
     - `task-number`: Číslo úlohy.
     - `preannot-source`: Zdroj předběžné anotace.
     - `annotator`: Kód anotátora.
     - `canonical`: Hodnota `1` značí kanonický přepis.

Hlavní obsah `<text>` Sekce `<text>` obsahuje jednotlivé úseky mluveného projevu strukturované pomocí elementů `<u>`: - **`<u>`**: Každý element `<u>` reprezentuje úsek projevu a má atributy: - `start` a `end`: Počáteční a koncový čas v sekundách. - `who`: Mluvčí (např. “EXAM_1” pro zkoušejícího a “CAND_1” pro kandidáta). - **`<s>`**: Každá věta je označena elementem `<s>`. - **`<tok>`**: Elementy tokenů, jejichž atributy popisují lemma, slovní druh, morfologické rysy a syntaktický vztah. - **`<anon/>`**: Anonymizovaný úsek nahrávky. - **`<gap reason="unintelligible"/>`**: Nesrozumitelný úsek nahrávky.

Příprava souborů TEITOK

Příprava souborů TEITOK probíhala v několika fázích:

1. **Předběžná anotace**. V rámci výzkumu spojeného s vytvářením databáze jsme porovnávali přímou ruční anotaci s manuální post-editací výstupů systémů pro automatické rozpoznávání řeči. Manuální anotace tak může vycházet z automaticky připravené předběžné anotace. Zdroj předběžné anotace rozlišujeme pomocí atributu `preannot-source`, jehož hodnota může být:
   - `from_scratch`: Kompletně manuální anotace, t.j. předběžná anotace je prázdná.

- `from_whisperX`: Předběžná anotace získaná pomocí systému WhisperX.
- `from_mixed`: Předběžná anotace získaná náhodným kombinovaním výstupů čtyř systémů na úrovni replik.

Když předběžná anotace nebyla prázdná, převedli jsme ji do základní verze formátu TEITOK. Na konci této fáze tak obsahovala přepisy rozdělené do replik (elementy `<u>`), přiřazení mluvčích k replikám (atribut `who`) a časové zarovnání s nahrávkou (atributy `start` a `end`).

2. **Manuální anotace**. Po nahrání souborů provedly zaškolené anotátorky manuální anotaci v prostředí TEITOK, během níž vytvářely nebo opravovaly přepisy, přiřazovaly mluvčí k replikám a pomocí časových značek zarovnávaly repliky s nahrávkou. Nahrávky byly anonymizovány v souladu s požadavky Ústavu jazykové a odborné přípravy Univerzity Karlovy (ÚJOP UK), který audionahrávky pro korpus poskytl. Některé anotátorky z opatrnosti anonymizovaly i údaje, které anonymizovány být nemusely (např. smyšlená jména osob).

3. **Revize**. Ruční kontrola manuálních anotací spoluautorkou databáze.

4. **Normalizace**. Automatická úprava přepisů, která odstraní odchylky ve jménech mluvčích, seřadí repliky podle počátečního času a přidělí replikám nové sekvenční ID.

5. **Rozdělení na úlohy a selekce**. Poskytovatel nahrávek (ÚJOP UK) povolil ke zveřejnění pouze vybrané úlohy. Ty jsme museli z nahrávek vystřihnout a upravit časové značky v přepisech, aby se zachovalo zarovnání replik v přepisu s nahrávkou. Pro stříhání nahrávky jsme použili nástroj FFmpeg.

6. **Lingvistická anotace**. Až do této fáze nebyly repliky v přepisech dále strukturovány. V této fázi jsme text rozdělili na věty (element `<s>`) a následně věty na tokeny (elemety `<tok>`). Na úrovni tokenů jsou přepisy automaticky lingvisticky anotovány. Každému tokenu je přiděleno lemma (atribut `lemma`), jazykově specifická morfologická značka (atribut `xpos`), slovní druh a morfologické vlastnosti dle kategorizace projektu Universal Dependencies (atributy `upos` a `feats`). Dále je každému tokenu přiřazen odkaz na ID rodiče podle pravidel závislostní syntaxe (atribut `head`) a typ závislosti tokenu ve vztahu k jeho rodiči (atribut `deprel`). Pro lingvistickou anotaci, včetně tokenizace, jsme použili nástroj UDPipe 2, konkrétně model `czech-pdt-ud-2.12-230717` pro češtinu. Ačkoli je možné provádět tokenizaci a automatickou lingvistickou anotaci přímo v prostředí TEITOK, my jsme tento proces realizovali samostatně. Důvodem je, že metoda tokenizace v prostředí TEITOK se liší od té, která je optimalizována pro UDPipe, což by mohlo způsobovat chyby při spojování těchto dvou kroků.

7. **Doplnění hlavičky TEI**. Na závěr jsme doplnili hlavičku podle všech dostupných metadat, aby odpovídala standardům TEI.

Všechy nástroje a skripty (převážně v jazycích Python 3 a BASH) jsou k dispozici ve veřejném repozitáři projektu v adresáři `data_preparation`.

Dotazování, vyhledávání a filtrování

Rychlé dotazování, vyhledávání a filtrace jsou umožněny integrovaným procesorem dotazů CQP, klíčovou komponentou sady nástrojů IMS Open Corpus Workbench (CWB). CQP převádí korpusy ve formátu XML do binární podoby a efektivně je indexuje. Dotazování v indexovaných korpusech probíhá pomocí jazyka CQL, který je standardem v korpusové lingvistice. TEITOK také nabízí Query builder, v němž může uživatel specifikovat dotaz vyplněním formuláře. Výsledek dotazu vrácený z CQP je následně zpracován pomocí TEITOKu a zobrazen uživateli v přehledné formě. Výsledky dotazů je možné stáhnout ve formátu XML.