Technical Documentation

The language corpus of spoken performances by non-native speakers of Czech, focused on the A2 language level (according to the CEFR), required for obtaining permanent residency in the Czech Republic, is the result of a project implemented at the Institute of Formal and Applied Linguistics of the Faculty of Mathematics and Physics, Charles University. The corpus contains recordings capturing the oral part of the Czech Language Certificate Exam at the A2 level. The recordings include dialogues between the examiner (a native speaker) and the candidate (a non-native speaker). We have provided transcriptions of the recordings, enriched with extensive linguistic annotations. Some recordings are accompanied by multiple transcriptions from different annotators, allowing for comparisons of various transcriptions of the same recording and the assessment of the degree of agreement when converting spoken language into written text.

The corpus is published as a specialized public database and is freely accessible to the general public, the scientific community, educators, and students. The database is integrated into the TEITOK system, managed on the LINDAT/CLARIAH-CZ platform.

TEITOK

TEITOK is a framework for creating, managing, and publishing annotated corpora. Its web interface is implemented using a combination of PHP and JavaScript. For our project, which combines recordings of spoken speech and their transcriptions, the key functionality of the TEITOK environment allows us to create, display, and edit recordings' transcriptions. To work with the recordings themselves, TEITOK utilizes the JavaScript library wavesurfer.

Data Storage

The corpus data is primarily stored in the TEITOK environment in the form of files. In this case, the recordings are in MP3 format, while the main components are TEITOK format files, which contain all transcriptions and annotations, including metadata. These files are interconnected with the corresponding recordings.

Structure of TEITOK Files

The TEITOK format is an XML format that fully complies with the Text Encoding Initiative (TEI) standards, but with a slightly different approach to tokenization. The structure of TEITOK files in our database is as follows:

Header with Metadata `<teiHeader>`

1. `<fileDesc>` – File description
   - `<titleStmt>`: Contains the title of the file and information about authors and annotators.
   - `<editionStmt>`: Contains version number.
   - `<publicationStmt>`: Publication details, such as publisher, release date, and license.
   - `<sourceDesc>`: Description of the source recording and a link to it.
2. `<encodingDesc>` – Description of encoding
   - `<projectDesc>`: A brief description of the project under which the data was created.
   - `<annotationDecl>`: Details of the individual annotation steps (primary, revision, linguistic annotation).
3. `<profileDesc>` – Profile of the text
   - `<langUsage>`: Language used (Czech).
   - `<textClass>`: Document metadata:
     - `database`: Database name.
     - `exam-id`: Exam identifier.
     - `cefr-level`: CEFR level. This database contains recordings exclusively from A2 level exams.
     - `task-number`: Task number.
     - `preannot-source`: Source of preliminary annotation.
     - `annotator`: Annotator code.
     - `canonical`: A value of `1` indicates a canonical transcription.

Main Content `<text>` The `<text>` section contains individual segments of spoken speech structured using `<u>` elements: - `<u>`: Each `<u>` element represents a segment of speech and has attributes: - `start` and `end`: Start and end time in seconds. - `who`: Speaker (e.g., "EXAM_1" for the examiner and "CAND_1" for the candidate). - `<s>`: Each sentence is marked with the `<s>` element. - `<tok>`: Token elements whose attributes describe lemma, part of speech, morphological features, and syntactic relations. - `<anon/>`: Anonymized segment of the recording. - `<gap reason="unintelligible"/>`: Unintelligible segment of the recording.

Preparation of TEITOK Files

The preparation of TEITOK files took place in several phases:

1. **Preliminary Annotation**. In the research associated with the creation of the database, we compared direct manual annotation with manual post-editing of outputs from automatic speech recognition systems. Thus, manual annotation may be based on automatically prepared preliminary annotation. The source of the preliminary annotation is distinguished using

the `preannot-source` attribute, which can have the following values:

- `from_scratch`: Completely manual annotation, i.e., the preliminary annotation is empty.
- `from_whisperX`: Preliminary annotation obtained using the WhisperX system.
- `from_mixed`: Preliminary annotation obtained by randomly combining outputs from four systems at the level of utterances. When the preliminary annotation was not empty, we converted it into the basic version of the TEITOK format. At the end of this phase, the transcriptions contained segments divided into utterances (the `<u>` elements), assignment of speakers to utterances (the `who` attribute), and time alignment with the recording (the `start` and `end` attributes).

2. **Manual Annotation**. After uploading the files, trained annotators performed manual annotation in the TEITOK environment, during which they created or corrected transcriptions, assigned speakers to utterances, and aligned utterances with the recording using timestamps. The recordings were anonymized in accordance with the requirements of the Institute for Language and Preparatory Studies of Charles University (ILPS CU), which provided the audio recordings for the corpus. Some annotators, out of caution, anonymized even data that did not need to be anonymized (e.g., fictitious names).
3. **Revision**. Manual review of the manual annotations by a co-author of the database.
4. **Normalization**. Automatic adjustment of transcriptions that removes discrepancies in speaker names, orders utterances according to start time, and assigns new sequential IDs to utterances.
5. **Segmentation by Tasks and Selection**. The provider of the recordings (ILPS CU) permitted the publication of only selected tasks. We had to cut these from the recordings and adjust timestamps in the transcriptions to preserve the alignment of utterances in the transcription with the recording. We used the FFmpeg tool for cutting the recordings.
6. **Linguistic Annotation**. Until this phase, the utterances in the transcriptions had not been further structured. In this phase, we divided the text into sentences (the `<s>` element) and then into tokens (the `<tok>` elements). At the token level, the transcriptions are automatically linguistically annotated. Each token is assigned a lemma (the `lemma` attribute), language-specific morphological tag (the `xpos` attribute), part of speech, and morphological properties according to the categorization of the Universal Dependencies project (the `upos` and `feats` attributes). Additionally, each token is assigned a reference to the parent ID according to dependency syntax rules (the `head` attribute) and the type of dependency of the token in relation to its parent (the `deprel` attribute). For linguistic annotation, including tokenization, we used the UDPipe 2 tool, specifically the model `czech-pdt-ud-2.12-230717` for Czech. Although it is possible to perform tokenization and automatic linguistic annotation directly in the TEITOK environment, we carried out this process separately. The

reason is that the tokenization method in the TEITOK environment differs from the one optimized for UDPipe, which could lead to errors when combining these two steps.

7. **Completion of the TEI Header**. Finally, we supplemented the header according to all available metadata to comply with TEI standards.

All tools and scripts (primarily in Python 3 and BASH) are available in the public repository of the project in the `data_preparation` directory.

Querying, Searching, and Filtering

Rapid querying, searching, and filtering are enabled by the integrated CQP Query Processor, a key component of the IMS Open Corpus Workbench (CWB) toolkit. CQP converts XML-formatted corpora into binary format and efficiently indexes them. Querying in indexed corpora is conducted using the CQL language, which is a standard in corpus linguistics. TEITOK also offers a Query Builder, in which users can specify a query by filling out a form. The results of the query returned from CQP are subsequently processed using TEITOK and presented to the user in a clear format. Query results can be downloaded in XML format.