COSI 136 Project 1

Timothy Obiso, Anastasiia Tatlubaeva

November 14, 2023

1 Data

For this project, we used the Bible audio and text from the Book of Genesis in Czech. The audio files were downloaded from https://www.wordproject.org/. The txt files were obtained by scraping the same website using `BeautifulSoup`. In total, there were 50 mp3 files and 50 txt files, corresponding to 50 distinct chapters. The length of each mp3 file varied, ranging from 3 to 13 minutes.

2 Preprocessing

All mp3 files were first converted to wav format. Then, these wav files were resampled to 16 kHz and rechanneled to one channel.

3 Alignment

For this assignment, we use WebMAUS Basic aligner. We have experimented with using Russian and Polish as the language setting and discovered that using Polish produces reliable textgrids. This fact is not particularly surpising since both Czech and Polish are West Slavic languages. We have also tried using Montreal Forced Aligner but ran into unresolvable technical difficulties. For the Speechbrain aligner, we could not find a pre-trained model that would work with Czech or any Slavic language in general.

4 Segmenting

For segmenting, we used the python packages `pydub` and `textgrid`. The first 25 chapters produced 13,189 segments; the second 25 chapters produced 16,970 segments.

5 Metadata

All 50 audio files were recorded by the same male speaker who seems to be 30-65 years old. Unfortunately, we did not find more information about this speaker.

The total length of all 50 audio files is 7 hours 9 minutes. The total number of segments is 30,159.

6 Strength and weaknesses

One of the major strengths of our corpus is the lack of background noise. In addition, the speaker reads with clear and deliberate pronunciation which is very important for training an ASR model. Furthermore, the language used in religious texts like the Bible is often formal and standardized. As a result, it would be easier for us to maintain consistency in the dataset.

At the same time, the Bible is not able to cover the full spectrum of language used in everyday conversations since it includes too many archaic or uncommon words. This aspect of our dataset can become a challenge if we try to apply the model trained on our dataset to some data collected in more informal environment. Our model could also struggle with more casual or colloquial pronunciation and styles found in other, non-religious contexts. Finally, our model might be biased towards male speakers, leading to potential challenges when faced with female speakers.

7 Repository description

Our repository contains the following components:

- folder `text` with all txt files.
- folder `textgrids` with all textgrids generated by the MAUS aligner.
- script `resample_files` used for resampling and rechanelling the wav files.
- this write-up.

We experienced some difficulties uploading the audio files to Github. For this reason, we will share them via Google drive instead. There will be two zip files: one with the resampled wav files and one with the segments.