Reflection part 3: Initial corpus Analysis

This final part of the assignment focused on turning my manually transcribed conversation into structured data for computational analysis. Using Python and NLTK, I explored basic quantitative features of my mini corpus, including word frequency, type-token ratio, sentence length and the distribution of function and content words. The process was new to me, but I found it fascinating to see how spoken language can be visualized and measured with code.

The first step was to export my ELAN transcript to plain text and cleaning it by removing symbols and time markers such as ((laughs)) and (1.0). After tokenizing the text, I created frequency plots and discovered that the most common words were short function words, such as “ja”, “en” and “toen”. This was expected since

Working with ELAN to transcribe my sisters’ dialogue was both challenging and rewarding. Since this was my second recording, I noticed that the conversation occasionally sounded slightly less spontaneous than the first version, which may have influenced the rhythm of the speech. Nevertheless, the transcription still contained rich conversational features such as overlaps and laughter, which made it engaging to analyze.

At the beginning, it took me some time to understand how to create tiers for each speaker and align their speech with the correct audio segments. Once I became familiar with the interface, the process started to feel more systematic and even satisfying. I selected a three to four-minute section that included moments of laughter, interruptions and short disagreements, as these elements best represented natural interaction.

One of the main challenges was accurately representing overlapping speech and timing. The Jefferson Transcription Conventions were very helpful, especially the use of symbols ((laughs)) for laughing and [ ] for overlapping talk. Listening closely also made me notice how many small Dutch discourse markers appeared, words such as “ja”, “oh” and “weet je nog”, which subtly structured the flow of the conversation.

Overall, transcribing the dialogue helped me appreciate how even informal family talk follows systematic linguistic patterns. The process gave me a deeper understanding of how speakers co-construct meaning through timing, alignment and shared emotional cues.