Task 2 - Exploratory Data Analysis

The first step in building a predictive model for text is understanding the distribution and relationship between the words, tokens, and phrases in the text. The goal of this task is to understand the basic relationships you observe in the data and prepare to build your first linguistic models.

Tasks to accomplish

1. **Exploratory analysis** - perform a thorough exploratory analysis of the data, understanding the distribution of words and relationship between the words in the corpora.
2. **Understand frequencies of words and word pairs** - build figures and tables to understand variation in the frequencies of words and word pairs in the data.

Questions to consider

1. Some words are more frequent than others - what are the distributions of word frequencies?
2. What are the frequencies of 2-grams and 3-grams in the dataset?
3. How many unique words do you need in a frequency sorted dictionary to cover 50% of all word instances in the language? 90%?
4. How do you evaluate how many of the words come from foreign languages?
5. Can you think of a way to increase the coverage -- identifying words that may not be in the corpora or using a smaller number of words in the dictionary to cover the same number of phrases?