Computing unigrams, bigrams and trigrams of the corpus from CHILDES database.

- Select an age group you are interested in.
- Calculate unigrams, bigrams and trigrams within sentences.
- Write the first 150 most frequent items (and the corresponding frequencies) in each method into individual text files.
- Re-calculate three n-grams across sentences and write the first 150 most frequent items into files.
- Identify common and uncommon elements by comparing the results of these two methods. Write the results into text files.
- (bonus) Many bigrams and trigrams you will find simply represent common syntactic constructions involving individually extremely common words (e.g. “in the” and “of the”). Develop and implement a method to filter and remove those items to make your lists contain more phrases that we would like to list in a dictionary.

Hint: the text files are quite clean but you will still need to remove non-letter symbols.

Acknowledgement: Brian Riordan, a graduate student in linguistics and cognitive science, provides the clean corpus for this class.