# Data – pre-processing

## Data sets

As mentioned, we used a data set for native and non-native English speakers.

We converted both datasets to lower case and we only took samples that have 45 tokens in them.

Since “Tofel” was the smaller dataset, we started with it and then took the same size of samples from “Reddit”.

For the non-native English speakers, we took 11044 samples. For the native English speakers, we had 6 different files from different countries. So, we took 1840 from each and had 11040 samples on hand. To avoid taking similar samples (sentences from the same discussion) we shuffled the files before picking the 1840 samples. We took all the 22084 samples and wrote them to a file shuffled after prepending the class label.

## Function words

We used several sources to gather a dictionary of function words. We achieved a dictionary of size 311. Every sample was represented as a vector in the size of 311 and each entry i was the count the ith function word. Below we show the top function words used by native and non-native English speakers:

|  |  |  |  |
| --- | --- | --- | --- |
| Native | | Non-native | |
| Function word | Count | Function word | Count |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |