Frequency analysis [ongoing]

* Brown corpus

Issues with English vs American spelling

* British National Corpus:

Very large corpus / not automated within nltk.

Have to access each file 🡪 time consuming

Larger data-set in BNC than brown corpus 🡪 had to raise the “cut-off” frequency

For example, several words in brown corpus were 0-frequency

Almost none are zero-frequency in BNC