questions for eric:

Although it has 44,764 tokens, this book has only 2,789 distinct words, or "word types." A **word type** is the form or spelling of the word independently of its specific occurrences in a text — that is, the word considered as a unique item of vocabulary. Our count of 2,789 items will include punctuation symbols, so we will generally call these unique items **types** instead of word types.

word types

so we have already done some of this. or eric has. these are modules to do some of the same.

hapaxes

collocations

bigrams

>>> list(bigrams(['more', 'is', 'said', 'than', 'done']))

[('more', 'is'), ('is', 'said'), ('said', 'than'), ('than', 'done')]

>>>

**3.3 and 3.4 – bigrams and collocations**

**bigrams – list of word pairs**

**collations – bigrams that usually occur together**