**NPL Project Report**

**Preprocessing**

First, I stripped out the speech marks from the text replacing them with spaces.

Then I expanded out standard contractions that I have compiled in a txt file. In cases of ambiguity of contractions I selected one of the pair. This still leaves some contractions where they are less widely used and more likely to be character specific.

Then I striped apostrophise to avoid names splitting. This is to avoid my regex tokenizer splitting john’s into john s rather than johns when it is removing punctuation.

Then I formed bigram phrases out of commonly occurring word pair

**Training.**