**Please answer the questions below:**

1. How many word types (unique words) are there in the training corpus? Please include the padding symbols and the unknown token.

Total Unique Words in the Training corpus 31245.

1. How many word tokens are there in the training corpus?

Total Word Tokens in the Training corpus – 446475

1. What percentage of word tokens and word types in each of the test corpora did not occur in training (before you mapped the unknown words to <unk> in training and test data)?

Percentage of word types in test corpus(brown-test) that do not occur in training corpus is - 6.84%

Percentage of word types in test corpus(learner-test) that do not occur in training corpus is - 6.61%

1. What percentage of bigrams (bigram types and bigram tokens) in each of the test corpora that did not occur in training (treat <unk> as a token that has been observed).

Percentage of bigrams in brown-test corpus that do not occur in training is- 28.61%

Percentage of bigrams in learner-test corpus that do not occur in training is- 25.40%

1. Compute the log probabilities of the following sentences under the three models (ignore capitalization and pad each sentence as described above). Please list all of the parameters required to compute the probabilities and show the complete calculation.

Which of the parameters have zero values under each model?

• He was laughed off the screen .

• There was no compulsion behind them .

• I look forward to hearing your reply .

6. Compute the perplexities of each of the sentences above under each of the models.

7. Compute the perplexities of the entire test corpora, separately for the *brown-test.txt*

and *learner-test.txt* under each of the models. Discuss the differences in the results you

obtained.