Name: _____ ID: _____

Exercise 4 Corpus Exploration

Download HW4.zip from the course website and unzip it. You will find three text files. They are of comparable size (between 240 and 260 words). Obtain the following information from each file (NOTE: do not forget to remove punctuations):

1. Number of types
2. Number of tokens
3. Type-token ratio (TTR)
4. Five most frequent words along with their frequencies
5. Five most frequent bigrams along with their frequencies

Step 1: Complete the table below with information from each file.

	Text 1	Text 2	Text 3
1. Types			
2. Tokens			
3. TTRs			
4. 1st			
2nd			
3rd			
4th			
5th			
5. 1st			
2nd			
3rd			
4th			
5th			

Step 2: Answer the following questions:

1. Why does text 1 have the highest lexical diversity? Does this have anything to do with the genre of the text? What is text 1 about? What purpose does it serve? [To answer this

question well, you will need to do some research on the genre of text 1. You may need to speculate the reasons as to why the text is most lexically diverse, based on its genre.]

2. Go over the three bigram lists. Which one is most able to reveal the text's topic? What is the topic about?
3. Go over the bigram lists once again. Which text is most casual? What leads you to this conclusion?

NOTE: I worked with [NAME] _____ to complete this exercise.