1) For students familiar with Unix command line tools:

- choose an English text of at least (approximately) 10K words

- process the text with the TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/), a tool that: a) splits the input into words (and punctuation marks) b) tries to guess the part-of-speech of each word c) maps each word to its lemma ("dictionary form", e.g., "dogs" is mapped to "dog")

- what is the proportion of nouns in the text?

- what is the proportion of verbs in the text?

- what are the 10 most frequent verbs?

- find an error in the part-of-speech assignment of the tool, and present a hypothesis for why the error was made

Some tips:

- Explanations for the part-of-speech codes are found here: http://sslmit.unibo.it/~baroni/collocazioni/english.tt.tagset

- You will need to install the tagger package and at least the parameter file for English

- The appropriate program to launch the English version of the system is tree-tagger-english

2) For students that are not familiar with Unix command line tools:

- Install some form of Unix on your computer, or gain access to a machine with Unix-like functionalities (Mac Os X and Linux are flavours of Linux; if you use Windows and do not want to install another OS, you can try: http://www.cygwin.com/)

- Follow the first four tutorials from: http://www.ee.surrey.ac.uk/Teaching/Unix/

- Now, install the TreeTagger and follow the instructions under 1) above