Computational Linguistics 1 - Fall 2011
CMSC/LING723, LBSC 744
Homework 0 - Due 8 Sept

Submit your responses and solutions to: `compling723.fall2011@gmail.com`
Homework write-ups may be in either plain text or .pdf format. Be sure include your full name, e-mail, *and any code* you write for the programming questions as part of your submission, and clearly indicate each problem number in your solution set.

Name:
E-mail:

Survey questions:

1. What is your general area of research interest?
2. Why did you choose the Computational Linguistics course? Do you plan to take Computational Linguistics 2 this spring?
3. How would you rate your programming skills? What language(s) are you most comfortable programming in?

Note: The purpose of the programming problems below is to allow us to calibrate everyone's programming skills and design appropriate assignments for subsequent classes. If doing this assignment requires more than a couple of hours, then you might want to reconsider whether you have the programming background necessary for the class. To assist you in completing these tasks, feel free to reference the Unix scripting tutorial or the Python tutorial posted on the class webpage.

Programming problems:
Download a copy of our training corpus, f2-21.txt, which was distributed to the course mailing list (`umd-cmsc723-fall-2011@googlegroups.com`).

1. How many sentences are in this corpus? What is the definition of "sentence" that you're using here?
2. How many words (separate by type and token) are there in this corpus? Let's define a "word" as being whitespace-delineated.
3. What are the top-20 most frequent words in this corpus? [Hint: in Python or another scripting language, try using a dictionary data structure to track occurrences.] Do they seem reasonable?
4. What are the top-20 most frequent *bigrams* in this corpus? Do they seem reasonable?
5. Now lowercase all of the text in the corpus, and repeat problems #3 and #4. Did anything change? Does that tell you anything interesting about the corpus?
6. The term bigrams in #4 refers to two words that co-occur in sequence, or next to each other. These are typically referred to as *collocations*. A similar term is *co-occurrences*, words that co-occur in the same sentence but not necessarily next to each other. Repeat problem #4 for bigram co-occurrences. Is there a difference from the bigram collocations you saw in #4?