

Assignment 2

L615

Due Tuesday, February 24

1. We're going to try to increase the range of tools you can use, while at the same time getting you comfortable downloading, learning, & running new software. Download the Multilingual Corpus Toolkit (MLCT) at <http://personalpages.manchester.ac.uk/staff/scott.piao/research/Download/download.htm>, or use a concordancer of your choice, perhaps one listed at <http://devoted.to/corpora>

If you don't have Java, you can always use the machines in MM 401.

- (a) Load one of the `.txt` files from `/Volumes/Corpora/en/susanne/`
 - These are reformatted corpus files from the SUSANNE Corpus (available at <http://www.grsampson.net/Resources.html>) containing sequences of word/POS, one sentence per line.
 - A = press reportage, G = belles lettres, biography, memoirs, J = learned (mainly scientific and technical) writing, N = adventure and Western fiction
 - (b) Hypothesize a stylistic difference between the genres, preferably one involving the use of some part-of-speech tag(s).
 - (c) Search for qualitative and quantitative differences. Write up a result of your findings, including what you expected to find (and why); what your results are and what they show; and critical discussion of the MLCT (or other) tool, its ease of use, and range of features.
2. Write a Perl program that does the following:
 - takes a corpus file as input
 - counts the number of lines
 - stores every 4th line into an array
 - loops over the array and outputs each element to an output file
 3. Starting with the research question that you described in the first assignment, or switching to another one that interests you a bit more ...
 - (a) Describe the question on one paragraph, as it stands now.

- (b) Start searching for relevant literature on the same topic or on similar topics. (Note that the library offers electronic access to the *International Journal of Corpus Linguistics*, among other resources.) In the selection of relevant literature, you should prefer literature on corpus-based approaches. If you cannot find any, describe that in your paper and concentrate on non-corpus-based research instead.
- (c) Write an overview of the main papers that you have found. The overview should describe the approaches in a way that it is possible for someone who has not read the original papers to understand the main points of the papers. The overview should also include your own critical opinion about what was good about the papers and what the difficulties with the paper are (including methodological problems).