1. [Extended version from last assignment] You want to build a corpus of all the (hand-written) assignments you’ve turned in during any portion of your schooling. In approximately two pages:

(a) Discuss potential applications of this corpus. Keep these in mind as you address the next points.

(b) Propose a corpus structure.

(c) Propose a set of meta-data tags to use for your files.

(d) Take a page or two of your writing and scan it, performing OCR in the process. Describe the challenges, referencing specific examples. Are there systematic/automatic solutions for your challenges? What is the potential impact on downstream processing of the corpus?

(e) Look at the documentation for PAULA XML: https://www.sfb632.uni-potsdam.de/en/paula.html and discuss how you would make your corpus fit the PAULA model. Specific examples will again help, perhaps even a whole or partial corpus file.

2. Take a basic text analysis / concordancing tool like AntConc and load two texts of interest into it, from different authors (ideally from the same genre/domain). Using the various options in the tool, propose & explore features which could automatically distinguish the two authors. It might help to search a bit on authorship attribution to see what is often used.

3. We’re going to try to get you comfortable downloading, learning, & running new software. Download the Multilingual Corpus Toolkit (MLCT) at https://sites.google.com/site/scottpiaosite/software/mlct or some other tool of your choosing.

Perform the same task as in question #2, but this time focus less on authorship attribution and more on tool usage. That is, I want a consumer report comparing the tools for your task. Consider ease of installation, ease of use, functionality (be sure to push the tools to their limits), speed, etc.

- If you don’t have Java, you can always use the machines in MM 401.
- If you don’t like MLCT, you can use a different tool, so long as it isn’t one we’ve already discussed. Some tools are here: http://wmtang.org/resources/corpus-resources/corpus-linguistics/tools/
- If you’re feeling adventurous, why not try 3 or 4 tools? :)

Assignment 2

L415 / L615

Due Wednesday, September 23
4. 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 a quick overview of the main papers that you have found. For each paper, include a note on why you’re looking at it and ideally some of its limitations that your work could address.