Presentations

Assignment: Give a presentation (15-20 minutes for M.DH.01, 25-30 minutes for B.DH.32 or M.DH.12) introducing or discussing some approach to natural language processing (NLP) and proposing how it may be brought to bear on a specific premodern corpus. Submit a write-up or project report (8-10 pp. for M.DH.01, 12-15 pp. for B.DH.32 or M.DH.12) by 31 March 2023 that builds on the content of your presentation. Run the topic by your instructor for approval in the early weeks of term.

Finding a Corpus

As a nonmedievalist, you'll want to start by getting a feel for what corpora are available. Ideally you'll choose a language that is not too alien to you: Latin if you've taken Latin, Old Church Slavonic if you read Cyrillic, Middle English or Middle High German if you have no other language expertise. But if you plan to take on a practical project, you'll also want to let yourself be guided by what corpora are available as clean plaintext, unless you want to set yourselves the task of cleaning up HTML or extracting text nodes from XML. To this end, the Oxford Text Archive may be helpful. If you don't know where to start, *The Canterbury Tales* and *Troilus and Criseyde* are good single-author options, while ECHOE and The Anglo-Saxon Poetic Records are accessible larger corpora. If you'd like to look farther afield, here is a larger set of recommendations:

- Old English:
  - ECHOE covers all non-Ælfrician Old English homiletic prose (537,000 words of Old English, once the Latin is filtered out). Although it is not yet publicly available, you have access to a clean, single-language copy in the course repository.
  - Ælfric's *Lives of Saints* are available to you in Old English *and* in translation via `aels.py` in the course repository, which pulls them in from Wikisource and cleans up the text for you.
  - The Anglo-Saxon Poetic Records cover all Old English verse.
  - YCOE is a sizeable corpus of Old English prose, marked up for parts of speech and syntax, and NLTK has a reader module providing direct access to the text content as well as to these metadata. See my manual of the reader module here. Try setting it up by downloading the corpus from OTA.
  - DOEC is, by one definition, a complete corpus of Old English prose and verse. It too is available through OTA.
- Individual Middle English texts:
  - *The Canterbury Tales*
  - *Troilus and Criseyde*

  - *Confessio Amantis*
  - *Sir Gawain and the Green Knight*
  - Layamon's *Brut* (plaintext retains some XML formatting)
- Middle High German:
  - Referenzkorpus Mittelhochdeutsch (1050–1350) (XML)
  - *Visio Tnugdali*
- Old Low German (Old Saxon):
  - *Heliand* (`psd` parsed format)
- Old Norse:
  - MeNoTa (XML)
- Old French:
  - Old French Corpus
- Latin:
  - A wealth of classical, medieval, and Neolatin texts is available through Corpus corporum (XML; you'll need to learn the basics of `lxml.etree` or Beautiful Soup to extract plaintext).
  - For classical texts, look into the corpora available through CLTK: see Appendix 2 to my manual, which describes CLTK's `FetchCorpus()` function.

Finding a Topic

The easiest solution is to pick an approach demonstrated in an online tutorial or practical textbook, such as the NLTK textbook, or one of the approaches used in class, and propose its application to your corpus of choice. A slightly more advanced approach is to locate a methodology in the ACL Anthology and propose its application to your corpus of choice. Although ideally you will carry out the proposed work, report on your findings, and submit a Jupyter notebook as an appendix to your written report, you may alternatively design your paper as a detailed project proposal discussing the nature of the corpus, your research goals, the nature of the approach, the steps and challenges on your way, and the anticipated outcome without submitting working code.

Structuring your Presentation

As your work will involve a corpus, a set of aims, and a method, you'll want to make each of these very clear in your in-class presentation before you get lost in the details of your approach and results (if applicable). So (1) introduce your corpus briefly but clearly, including any relevant limitations or biases inherent in the corpus; (2) tell us what information you'd like to extract and how that information can help advance our

knowledge; (3) introduce the chosen method in such a way that your audience understands what it does without feeling patronized; and then proceed with the details of your approach and/or findings. Above all, keep in mind that your presentation has to be both relevant and interesting to your audience!

On the Use of Code Listings

If you are carrying out a practical project, you will be expected to submit a Jupyter notebook along with your paper. But whether or not you submit a notebook, there is a good chance you'll want to include one or more code snippets directly in your report. (NB neither the notebook nor the code snippets count towards your page count.) While there are other, technically preferable solutions for the printing of listings especially in technical environments like LaTeX, a quick hack for traditional word processors is to paste your code into an online tool like planetb.troye.io and from there into your word processor.

Marking Grid

The following marking grid indicates the considerations used to mark papers for this course (but not their weighting):

Aspect	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Methodological grasp										
Programming skill										
Case for relevance										
Structure										
Use of secondary sources										
Command of academic English										
Mechanics (style, referencing, formatting)										
Length										

To clarify: "methodological grasp" means you've understood what the chosen method (e.g. TF-IDF) does, while "programming skill" means you've understood how to go about tackling the proposed project on an applied, technical level. The latter can be shown through the submission of a Jupyter notebook, but it can also be demonstrated by a clear explanation in prose of how you plan to take on your project, and ideally you'll do both. The "case for relevance" is where you explain of what use your project is: what will we have learned about the corpus, or what information can be more easily extracted, once you're done? The remainder of criteria are general essay-writing skills: your paper should be well structured, make sufficient and proficient use of scholarship, cite its sources correctly (corpora and software resources as well as scholarship!), and be formatted correctly. If you are accustomed to writing in German, be aware of English conventions around punctuation, and use an English style sheet for citations. These are available directly in Zotero, or if you use LaTeX you can rely on `biblatex` to format your citations. If you compose in the Pandoc flavour of Markdown, you can combine the two: download a Zotero style and source a `biblatex` bibliography file, then compile using Pandoc's `--citeproc` option.

Presentation Slots

Presentations will be evenly distributed over the weeks towards the end of the term. Inform your instructor about your preferred slot when you approach them about your topic.