Term Paper: Specifications and Suggested Topics

Assignment: Write a paper either on an aspect or implementation of natural language processing, or on a linguistic or literary topic in which you involve the methodology of natural language processing.
Length: Within ten percent of 3,000 words for M.EP.02b; within ten percent of 7,500 words for B.EP.11b, M.EP.05b, and M.EP.05d. Please note that the paper length for your module has substantial implications for the appropriate choice of topic and scope.
Approval: Run at least the topic, and ideally (at a later stage) a detailed sentence outline, by your instructor for approval. (See these slides for an explanation of the appropriate sentence outline format.)
Deadline: Final term paper due on 26 August; try to have a topic by late May.

Some Possible Topics

The following lists are intended to give you an impression of the range of appropriate topics, but please do not infer that you are limited to the options listed. While you may copy out a topic directly from this list, you may prefer to come up with your own, and it is at any rate your own responsibility to formulate a research question. Remember that originality of argument (and by implication originality of research question) is a factor in the assessment of your paper (see marking grid below). All the approaches listed below presuppose a primary reliance on new work, though you'll have to cite scholarship nonetheless at least as a way of embedding your project in the wider field. Given that only a modest amount of digital humanities scholarship is available on medieval languages and literatures, you may find it hard to locate relevant scholarship to cite. In such cases you would be especially well advised to open with a solid section on more general work done in the field, be it from a literary-linguistic angle (authorship analysis) or digital methods.

Papers for this course may investigate a medieval text or corpus in Latin or any Germanic language, but Old and Middle English are especially encouraged given their relevance to the hosting department. Whereas humanities scholarship is typically expected to advance an argument, essays on technical topics may alternatively take the form of review papers, particularly if they include proposals for technical improvements or the further development of existing projects.

Please refer to the following lists of topics for inspiration.

Practical projects on stylometry and authorship

Projects like these require you to spend some time preparing your corpus, after which you'll either run NLTK routines on them as described in Bird et al. or you'll use off-the-shelf tools like Lexos (or both). Please be aware that the results of stylometrical analysis are rarely clear-cut and that they typically require a great deal of discussion.

- How many authors were involved in the production of the "Alfredian" corpus?
  (For secondary sources start with Godden, "Did King Alfred Write Anything?")

- Is the Old English *Ecclesiastical History of the English People* the work of a single translator?
  (This is DOEC `T06860.txt` to `T06910.txt`.)

- Can stylometry help define the Wulfstanian corpus?
  (Best use Napier, i.e. `T04020.txt` to `T04280.txt` in DOEC, as your corpus, along with some other sermons from DOEC or ECHOE as control material.)

- Can we quantify the differences in style between Ælfric, Wulfstan, and the anonymous collections of homilies contained in the Blickling and Vercelli manuscripts?
  (Ælfric is in DOEC `T02040.txt`–`T03000.txt`; Blickling is ECHOE `382.*`; Vercelli is ECHOE `394.*`; see previous item for a first, imprecise Wulfstan corpus.)

- Does *Beowulf* have stylistically distinct parts?
  (See Appendix 2 to the guide to CLTK for access to the Old English poetic corpus.)

- Can stylometry confirm the poetic corpus of Cynewulf?
  (His signed poems are *The Fates of the Apostles*, *Juliana*, *Elene*, and *Christ II*; see previous item for access to the Old English poetic corpus.)

- A quantitative perspective on the Latin lexicon of Saxo Grammaticus
  (The *Gesta Danorum* is available in HTML format at the Royal Library of Denmark and may be downloaded from the command line using `wget -m -k -l 7 -t 6 -w 5` followed by the URL (takes several hours, after which the corpus still requires substantial cleanup); a glossary exists at http://www.rostra.dk/latin/saxo1.html. Control corpora may be downloaded from Corpus Corporum.)

Practical projects involving artificial intelligence

With AI the technical work involved after corpus prep is usually modest, and you won't have much to say about the learning process. This means you'll have to write at greater length about something else instead, such as edge cases; issues with the results and how you might improve them; and/or a more thorough introduction to the known differences between the categories you are attempting to separate out and what a rule-based approach might look like.

- Training a computer to distinguish between Old and Middle English
  (The Corpus of Middle English Prose and Verse is at https://quod.lib.umich.edu/c/cme/.)

- Discerning Old English dialects through stylistic analysis
  (Campbell's *Old English Grammar* is a good place to learn about dialects.)

- Quantifying bias in medieval (English/Norse/Latin/German) texts using word embeddings

- Quantifying the associations and evaluation of named individuals in medieval (English/Norse/Latin/German) texts

- Generating a Latin–Old English dictionary on the basis of the psalter glosses

Practical projects with rule-based approaches

Rule-based NLP work is more labour-intensive, but it leaves you more opportunity to discuss your methods.

- Stemming Old English
- Lemmatizing Old English (prose/verse)
- Automating the sentence division of medieval manuscript transcriptions in the absence of modern punctuation
- Harnessing digital text corpora in the development of word games
- Inferring phonotactics from digital corpora

Project reviews and proposals for further development

Topics like these require no programming, relying instead on your understanding of the potential of NLP. Even so, a good way to demonstrate that understanding would be to offer technical detail on how the proposed goals would be achieved.

- Digital methods at the *Dictionary of Old English* Project: assessment and recommendations
- Digital methods at the Skaldic Project: assessment and recommendations
- NLP use cases for the data of the *Prosopography of Anglo-Saxon England*
- NLP use cases for the data of the *Thesaurus of Old English*
- NLP use cases for ECHOE data
- NLP use cases for the data of the Corpus of Middle English Prose and Verse

For prior work on digital methods, you'll want to refer to existing projects (see, e.g., §§2–3 of the bibliography on the syllabus), but in addition you had best browse such journals as *Digital Medievalist*, *Digital Humanities Quarterly*, and the journals listed here. To identify relevant scholarship in medieval language and literature, you would be well advised to make use of the International Medieval Bibliography in addition to resources like Google Scholar. Not all Old English prose texts are available in translation, so if you need to know what your text says you'll have to hunt down translations in editions, readers, and translation series, but translations of the poetry may be found in Craig Williamson's *Complete Old English Poems*. If you are pursuing a linguistic topic, Campbell's *Old English Grammar* and Noreen's *Altnordische Grammatik* are some of the most thorough reference works.

On the Use of Code Listings

While there are other, technically preferable solutions for the printing of code snippets 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.

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**
Argument *or* survey (quality, originality, thoroughness)										
Structure										
Methodology										
Technical analysis										
Use of secondary sources										
Grasp of source language *and/or* digital environments										
Command of academic English										
Mechanics (style, referencing, formatting)										
Length										