Project Specification

Medieval Words, Modern Methods

revision of October 21, 2025

Overview

This seminar is centred around your own personal project. Over the course of the term, you will select a short handwritten text or excerpt of circa 300 words, medieval or early Modern at the latest, transcribe and encode it in TEI XML, write a brief scholarly introduction, mark up your text with a range of metadata, style it in CSS, access the document tree using Python, generate a basic statistical profile of the text, and document your methods in a Jupyter notebook. Students in DH modules are expected to submit the weekly homework following the specifications given in the syllabus. Students of B.Eng.611, on the other hand, are additionally expected to submit a fair copy of their portfolio. The present document specifies the requirements of the latter portfolio and should be read in conjunction with four further documents: the course syllabus, which specifies what part of the preliminary work is due when; the project suggestions; the presentation specification; and the resources document. DH students should consult this document for guidance, but may ignore the submission requirement.

Scholarly Introduction

As part of your project, you will write a brief scholarly introduction of 300 words or more, encoded in the relevant parts of the TEI header. It should at least have a palaeographical component describing the text's transmission (the surviving manuscript witnesses, their date and origin, your choice of manuscript) and especially that of the chosen witness. The introduction should address at least one further scholarly question of your choosing, such as authorship, date of composition, dialect, sources, metrics, or vocabulary; this information may be gleaned from existing scholarship (which should be duly cited; see TEI Guidelines § 3.12, "Bibliographic Citations and References") along with your own research. The introduction is best written after you have gained a familiarity with the text and its scholarship, and in close conjunction with your class presentation.

Encoding Specification

Mandatory Markup

Your text or excerpt should be encoded following the TEI P5 standard on the basis of the template supplied in the course repository, with at least the following features:

- A codicological account of your manuscript in the TEI header (see TEI P5 ch. 11, "Manuscript Description");
- Such basic units of markup as `<w>`, as well as `<l>` (for verse) or `<p>` (for prose; see TEI P5 ch. 3, "Elements Available in All TEI Documents"), and `<caesura/>` and/or `<seg>` if applicable;
- Abbreviations (if applicable), using at least `<ex>` to indicate supplied expansions, or `<choice>` nodes with children `<abbr>` and `<expan>`, and/or `<am>` and `<ex>`, where required;
- Scribal interventions and your own editorial emendations (`<add>`, `<del>`, `<sic>`, `<corr>`, wrapped in `<choice>` as required, alongside `<surplus>` and `<supplied>` as needed; likewise documented in ch. 3 of the TEI Guidelines).

Advanced Markup

You are expected to mark up your text with at least two of the following three additional types of metadata, or others like it:

- lemmatization (`<w>` with `@lemma`; see TEI P5 ch. 18, "Simple Analytic Mechanisms");
- part of speech (`<w>` with `@pos`; likewise documented in ch. 18 of the TEI Guidelines);
- metre (`<l>` with `@met`, and/or `<seg>` with `@met` for halflines; see TEI P5 ch. 6, "Verse").

The format of your lemma and POS references is up to you; we will discuss some of the options in class.

Additional Markup

You are encouraged, but not required, to encode any further type of information you encounter, such as named entities (`<persName>`, `<placeName>`, `<name>`).

Styling

Your project should include a customization of the CSS stylesheet template supplied in the course repository. Your document should render appropriately.

Processing and Evaluation

Your portfolio should include one or more Jupyter notebooks demonstrating at least the following features, each amply documented in a markdown cell and/or code comments:

- Retrieval of your XML text nodes into lists of tokens, preserving such structures as verse lines or paragraphs;
- A metadata structure encoding such features as part of speech, lemmata, and/or metrical data alongside the text of each token, stored under `tokens[0]['lemma']` etc. alongside `tokens[0]['text']`;
- Some basic statistics on tokens and metadata (e.g. token count, lexical diversity, words/syllables per verse line, metrical counts);
- Comparison of at least two features with a modest comparable corpus sourced elsewhere (e.g. TF-IDF vectors, lexical diversity, words/syllables per verse line, metrical counts);
- At least four graphs visualizing these data, at least two of which should involve the broader corpus.

Submission

Whereas the regular homework you upload to Stud.IP serves only as a progress check, students of B.Eng.611 will additionally submit (by the FlexNow deadline) the definitive version of their portfolios by email as a ZIP, RAR, or `tar.gz` archive containing the following files:

- One XML transcription;
- One CSS stylesheet, correctly associated with the XML document by stylesheet declaration;
- One or more Jupyter notebooks demonstrating the rest of the work required, saved after running (i.e. with the output cells showing the desired output).

If you rely on a large, conventional external corpus (e.g. ASPR, DOEC, YCOE, YCOEP, CME, Corpus corporum) for your comparative data, and haven't had to modify that corpus to meet your needs except as demonstrated in the notebooks you submit, don't include the corpus in your submission. If, on the other hand, you have established a small custom corpus e.g. by manually copying text from your browser, include that corpus in a subfolder. In either scenario, carefully document in your notebooks how you obtained and modified the corpus.