The Values of French Language and Literature in the European Middle Ages

The Histoire ancienne jusqu'à César : *Lemmatisation and Search* (2.0)

The lemmatisation of *The* Histoire ancienne jusqu'à César*: A Digital Edition* is a key output of *The Values of French*. This dataset offers significant opportunities for linguistic and lexical analyses of medieval French.

The edition is of the two most important manuscript witnesses of the *Histoire ancienne jusqu'à César*, which were created in regions that demonstrate the supralocal scope of French in this period. The first, Paris, Bibliothèque nationale de France, f. fr. 20125 (2/2 13th c., possibly Acre), is one of the earliest and the most extensive witnesses of the first redaction of the *Histoire ancienne*. The second, London, British Library, Royal 20 D I (1330s, Naples), is the earliest extant copy of the so-called second redaction, distinguished primarily from the first by its more exhaustive account of the Fall of Troy known as 'Prose 5' (see Jung 1996: 505-62).

Users are able to access the data from lemmatisation on the digital edition's faceted search page (see below).

Digital Workflow
Custom-designed scripts automatically tokenise the words in the edition (whilst maintaining the XML structure), and generate the KWIC ('Key Word In Context') index required for lemmatisation.

Lemming
Lemming is a custom-made digital tool for lemmatisation, designed by our colleagues at the *Dictionnaire étymologique de l'ancien français* (*DEAF*). The lemmata in Lemming are based on the *Tobler-Lommatzsch, Altfranzösisches Wörterbuch* and/or the *DEAF*. *Lemming* operates on the fly allowing multiple users to work at once. A KWIC index of the words in the (tokenised) interpretive text is uploaded to *Lemming*. Users are able to attribute the same lemma to numerous keywords and annotate ambiguous cases. When users identify inaccuracies in the editorial work (e.g. missing accents, incorrect word division or letter forms, etc.), corrections are made to the TEI-XML source files and the content is then re-uploaded to *Lemming*. A verification page ensures that any changes to the keywords or contexts are correctly integrated into the database.

XML Files

- **Lemmatised Contexts**

The output KWIC index comprises the data from lemmatisation. Each <item> contains the following attributes associated with each keyword (encased in <string>):
@type = seg_item (prose); verse_item (verse); rubric_item (rubric)
@location = xml:id in the edition
@n = token number in the <seg> of the edition
@preceding = 8 words/marks of punctuation preceding the keyword
@following = 8 words/marks of punctuation following the keyword
@lemma = lemma attributed to the keyword

@lemmaPos = part of speech associated with the lemma
@pos = part of speech attributed to the keyword
@sp = keyword occurs in direct/indirect speech

- **Tokenised Files**

The tokenised files number every word in every segment of the text in each manuscript. The KWIC uploaded to *Lemming* is based on this file and only extracts data from the interpretive text.

The data in the output KWIC from *Lemming* are combined with the original tokenised files on the digital edition's search page. This enables us to use information from the encoding in the TEI-XML source files to broaden the search possibilities (e.g. the encoding of direct/indirect speech).

Search **(**https://tvof.ac.uk/search/)
The search page enables users to search by lemma or form. There are three types of results available:

1. Words in Context (concordance)
2. Names (index of names)
3. Lemmata (all lemmata and names)

It is possible to move directly from Names and Lemmata to Words in Context when a name or lemma is selected. The Words in Context results list is also linked to the Text Viewer.

The facets available for filtering the Words in Context results list are:

- Manuscript
- Lemma
- Form
- Section
- Grammatical Attribute
- Text Body/Rubrics
- Textual Form (verse/prose)
- Narration/Speech (narration/direct speech/indirect speech)

Credits
Marcus Husar designed the tool for lemmatisation (*Lemming*), in collaboration with Stephen Doerr of the *Dictionnaire étymologique de l'ancien français* (*DEAF*). Marcus Husar also supported the integration of the lemmatised data into the faceted search page.

All members of the *TVOF* team were involved in lemmatising the text: Simon Gaunt, Hannah Morcos, Maria Teresa Rachetta, Henry Ravenhall, Natasha Romanova, and Simone Ventura.

Paul Caton, Ginestra Ferraro, and Geoffroy Noël collaborated on the design, development, and digital workflow of the edition and search page.