WHEN EPIGRAPHY MEETS LINGUISTICS

The semantic annotation of modality in the Latin inscriptions of the Republican age

Francesca Dell'Oro
Helena Bermúdez Sabel
Pauline Jacsont
contact: delloro.fr@gmail.com

A WORLD OF POSSIBILITIES

unine — Université de Neuchâtel

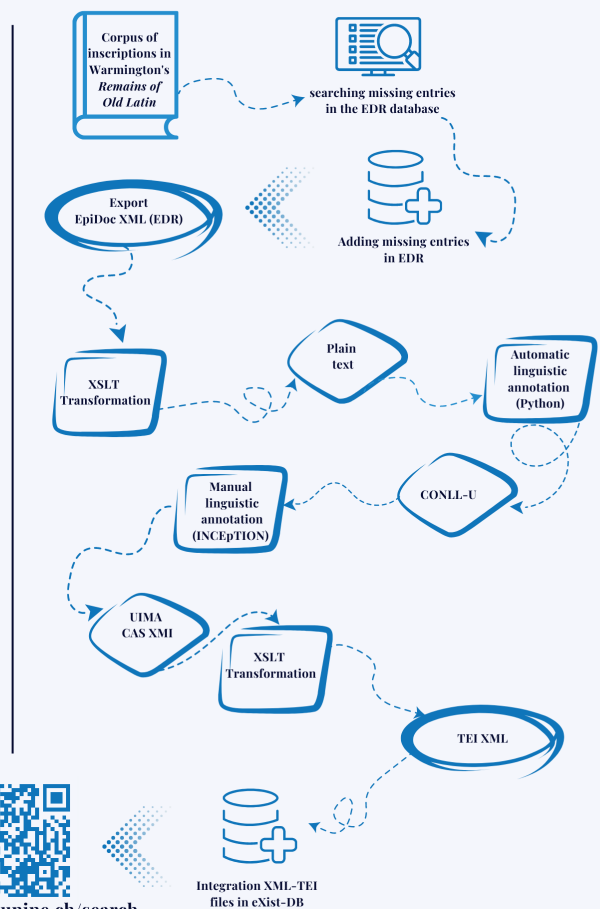Swiss National Science Foundation

Modality is the linguistic expression of the notions of necessity, possibility and volition. The team of the project *A World of Possibilities. Modal pathways over the extra-long period of time: the diachrony of modality in the Latin language* (WoPoss) aims at studying modality in the history of the Latin language by setting up a representative corpus.

In order to be representative of (socio)linguistic variation, it is mandatory that the WoPoss corpus includes both literary and documentary texts. This is where the WoPoss project has met the Epigraphic Database Roma (EDR) project.

The WoPoss corpus includes now 74 inscriptions, mainly from the Republican age. This recent achievement crucially enriches the WoPoss corpus and enhances epigraphic research, by providing a fine-grained semantic analysis. In the next future, we plan to add a selection of inscriptions of the Imperial age.

THE WORKFLOW

- Identification of ca. 90 missing inscriptions in the EDR database taking Warmington's *Remains of Old Latin* as reference for the selection of inscriptions of the Republican age.

- Creation of a detailed EDR record for each missing inscription, based on the most recent epigraphic publications.

- Further selection of pertinent inscriptions: only the inscriptions potentially containing a modal marker were selected to become part of the WoPoss corpus.

- Retrieval of the pertinent inscriptions from the EDR EpiDoc Dataset (Panciera *et al.* 2020).

- Conversion of the XML files to plain text adopting the Leiden conventions.

- Automatic linguistic annotation of the inscriptions based on the Stanza tools for Natural Language Processing (Qi *et al.* 2020) resulting in CONLL-U files. The automatic annotation includes lemmatisation, tagging of parts of speech, morphosyntactic description and syntactic dependencies parsing.

- Manual linguistic annotation of the modal passages in the inscriptions, after upload of the CONLL-U files in the annotation platform INCEpTION (Klie *et al.* 2018).
  ○ In this phase errors due to the automatic annotation were manually corrected.
  ○ The manual linguistic annotation was carried out based on the WoPoss guidelines for the annotation of modality (Dell'Oro 2022). For more information, cf. Dell'Oro *et al.* 2020.

- Conversion of the UIMA CAS XMI files exported from INCEpTION to the TEI format.

- Validation of the XML files against the WoPoss project's XML schemas (Bermúdez Sabel 2020) and adding of the validated files to the WoPoss eXist database.

- **The corpus containing the annotated inscription can be accessed through a dedicated interface.**

Corpus of inscriptions in Warmington's *Remains of Old Latin*

searching missing entries in the EDR database

Adding missing entries in EDR

Export EpiDoc XML (EDR)

Plain text

Automatic linguistic annotation (Python)

XSLT Transformation

CONLL-U

Manual linguistic annotation (INCEpTION)

UIMA CAS XMI

XSLT Transformation

TEI XML

Integration XML-TEI files in eXist-DB

https://woposs.unine.ch/search

Selected references:

- Bermúdez Sabel, Helena. "XML Annotation Schemes of the WoPoss Project." 2020, WoPoss project. https://github.com/WoPoss-project/annotation-schemes.
- Dell'Oro, Francesca. WoPoss Guidelines for the Annotation of Modality. Revised Version. Zenodo, 2022. https://doi.org/10.5281/zenodo.6417878.
- Dell'Oro, Francesca, Helena Bermúdez Sabel and Paola Marongiu. 2020. "Implemented to Be Shared: the WoPoss Annotation of Semantic Modality in a Latin Diachronic Corpus". Sharing the Experience: Workflows for the Digital Humanities. Proceedings of the DARIAH-CH Workshop 2019. December 5-6, 2019. Neuchâtel, Switzerland.
- Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation." In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics, 2018. http://tubiblio.ulb.tu-darmstadt.de/106270/.
- Panciera, Silvio, Giuseppe Camodeca, Giovanni A. Cecconi, Silvia Orlandi, Lanfranco Fabriani, and Silvia Evangelisti. "EDR - Epigraphic Database Roma EpiDoc Files." Zenodo, July 6, 2020. https://doi.org/10.5281/zenodo.7561922.
- Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf