WoPoss: a Workflow for the Semantic Annotation of Modality in a Diachronic Corpus

Helena Bermúdez Sabel* , Francesca Dell'oro*[1], and Paola Marongiu*[1]

[1]University of Lausanne – Switzerland

Abstract

The FNS project *A world of possibilities* (WoPoss) aims at reconstructing the evolution of modal meanings in the Latin language. Passages expressing modal notions such as 'possibility and 'necessity' are annotated in three phases. Phase 1 and 3 being carried out automatically, annotators can concentrate on the fine-grained semantic analysis of modal readings (Phase 2).
1. The initial dataset is gathered from different online open access resources. The formats of the documents necessarily vary and even when some standard is used, e.g. TEI, the encoding schemas are different. To tackle this heterogeneity, we convert our sources into plain text with additional pseudo-markup to preserve the pertinent semantic information previously conveyed in the XML or HTML tags. These files are then automatically annotated using the StanfordNLP library for Python. The resulting CONLL-U files are uploaded to the annotation platform INCEpTION.

2. Through this platform the annotators add the relevant semantic information, following the WoPoss schemas and guidelines. We later examine their analyses opening a revision process.

3. The revised results are then exported in a XMI format. To guarantee the re-usability of our dataset, we perform a post-processing that transforms this format into TEI with linguistic information encoded through stand-off annotation (Bański 2010). The complete dataset will be stored in a no-SQL database and exploited through an user-friendly interface. The dataset together with the software and code that makes it searchable will be provided in a open access repository.

References (accessed on 12/09/2019):

WoPoss. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language : http://woposs.unil.ch

Bański, Piotr. 'Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless.' Presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3 - 6, 2010. In *Proceedings of Balisage: The Markup Conference 2010.* Balisage Series on Markup Technologies, vol. 5 (2010). https://doi.org/10.4242/BalisageVol5.Banski01

CONLL-U: https://universaldependencies.org/format.html

*Speaker

INCEpTION: https://inception-project.github.io
StanfordNLP library: https://stanfordnlp.github.io/stanfordnlp