Implemented to Be Shared: the WoPoss Annotation of Semantic Modality in a Latin Diachronic Corpus

Francesca Dell'Oro[1,2*], Helena Bermúdez Sabel[1], Paola Marongiu[1]

1 University of Lausanne

2 Center for Hellenic Studies, Harvard University

*Corresponding author: Francesca Dell'Oro francesca.delloro@unil.ch

Abstract
The SNSF project *A World of Possibilities* [WoPoss] aims at tracking the evolution of modal meanings in the diachrony of the Latin language. Passages expressing modal notions of 'necessity', 'possibility' and 'volition' are annotated following a pipeline that combines both automatic and fine-grained manual annotation. Texts are first gathered from different online open access resources to create the initial dataset. Owing to the heterogeneity of formats and encodings, the files are regularized before the implementation of the automatic annotation of linguistic features. They are then uploaded to the annotation platform [INCEpTION] which has been previously customized for the fine-grained manual annotation [Bermúdez Sabel, in press]. This second type of annotation is carried out following the WoPoss guidelines [Dell'Oro, 2019] and it also involves checking the automatic annotation for curation. In a third stage the files are automatically enriched with metadata. In this paper, we focus on the first two phases of the workflow – i.e. the gathering and automatic annotation of the texts and the fine-grained manual annotation –, which have been the core tasks of the WoPoss team in the first year of the project (2019-2020).

keywords
historical linguistics; Latin; semantic annotation; semantic modality

INTRODUCTION: IMPLEMENTED TO BE SHARED

The SNSF project *A World of Possibilities. Modal Pathways over an Extra-Long Period of Time: the Diachrony of Modality in the Latin Language* (from now on [WoPoss])[1] tackles the question of the semantic diachronic development of modal meanings of 'necessity', 'possibility' and 'volition' in Latin. Complying with the SNSF effort for open science, WoPoss was conceived with the precise goal of sharing not only its analysed and annotated data, but also the design of both its general workflow and its specific pipeline practices. In this paper we will focus on the first two phases of the general workflow, which correspond to our main tasks in the first year of the project (2/2019 – 1/2020): the preparation of the text files for manual annotation and the manual annotation itself. Section I describes the pipeline, while section II outlines the manual annotation process. Section III presents some views on future work and in particular on the third phase of our workflow, i.e. the post-processing stage and the implementation of a user-friendly search interface.

I PIPELINE

1.1 Setting-up of the corpus

Starting from an empirical approach, modal meanings are elicited for a predefined list of markers [Dell'Oro, 2019:8-9] from a diachronic corpus. The WoPoss corpus ranges from the

[1] The SNSF project n° 176778 is led by F. Dell'Oro. This paper was written collaboratively: F. Dell'Oro is mainly responsible for section I and both the introduction and the conclusions; F. Dell'Oro and P. Marongiu are mainly responsible for section II; H. Bermúdez Sabel is mainly responsible for section III.

3rd century BCE to the 7th century CE. In order to see the emergence of variations in the uses of modal markers in the whole Latin diasystem [Coseriu, 1969], texts are carefully selected without overlooking any potential source of knowledge about (socio-)linguistic variation. The corpus therefore includes both literary and documentary texts as they are attested by different types of sources, such as manuscripts, inscriptions and papyri, in the various Latin-speaking regions of the ancient world and in different textual genres.
The setting-up of the corpus as a (future) online open access tool relies heavily on projects which share the same objectives [WoPoss-credits]. With these projects we have both informal – e.g. with the Perseus Digital Library project – and formal (i.e. signed) agreements – e.g. with [EAGLE-IDEA] for the use of its conglomeration of epigraphic corpora.
Likewise, the texts of WoPoss will be freely available for reuse under a creative commons licence (such as [CC BY 4.0]). The team annotates modal passages (see section II) after having checked their philological correctness against reference editions (which are mostly not open access products). This process allows us to offer to future users a philologically correct text for the annotated modal passages.

1.2 Text preparation

After the texts have been selected, text files – whatever their format – are converted to plain text. It must be noted that not only the formats (XML, HTML, plain text), but also the standards can vary widely, as in the case of [TEI] since there are collections following a [TEI-lite] customization, such as [digilibLT], and others that are compliant with the [EpiDoc guidelines], such as the [Perseus Digital Library]. Relevant already tagged information is transformed into unambiguous punctuation marks according to a predefined convention we established for this purpose (pseudo-markup). This allows us to keep important structural information in the plain text files: for instance, in a play, we add pseudo-markup tags for the character name preceding each takeover.
The plain text files are then automatically annotated using the [StanfordNLP] library for Python. The output is a [CONLL-U] file where each line represents a single word with a series of tab-separated fields that contain the linguistic annotation: form, lemma, part of speech, morphological features and dependency relations. These files are then uploaded to the annotation platform [INCEpTION]. The results of the automatic annotation are visible and editable in the platform as layers to which the manual fine-grained semantic annotation can be easily added.

II ANNOTATION

2.1 Customization of the annotation tool

The tool INCEpTION is a multi-modular annotation platform especially suited to carrying out semantic annotation [Klie et al., 2018], such as the one designed in the WoPoss theoretical framework. Our annotation deconstructs a modal expression into its different components: the modal marker, its scope (with the state of affairs) and their relation. These elements are further described by the annotators with reference to different linguistic features.
In order to customize the INCEpTION tool and adapt it to the needs of our project, the principles of the annotation guidelines [Dell'Oro, 2019] have been formalized in terms of layers (the levels of annotation, e.g. 'modal unit'), features (the components of the layer, e.g. 'type of modal unit'), controlled vocabularies (tagsets, that is, the possible values of a feature, e.g. 'modal marker', 'scope'), and constraints (the rules). The objective is that the annotation interface should, as far as possible, correspond to the annotation workflow; thus features and tags appear in a logical order (see [Bermúdez Sabel, in press]).

2.2 Examples from the manual annotation practice

The main task of the whole WoPoss team is the manual annotation. Presented below are three complex examples which allow us to deal with some specific aspects of our annotation (see also [Dell'Oro, 2019]).

(1) QUINTILIAN, Institutio Oratoria 12, 1 § 3, p. 367, lin. 6: [...] *dico, eum qui sit orator, virum bonum esse oportere.*

dic-o,	*e-um,*	*qu-i*	*si-t*	*orator,*	*vir-um*
say.PRS-1SG	he-ACC.M.SG	who-NOM.M.SG	be.PRS.SBJV-3SG	orator.NOM.M.SG	*man*-ACC.M.SG

bon-um	*esse*	*oport-ere.*
honest-ACC.M.SG	be.PRS.INF	be_necessary-PRS.INF[2]

'[…] I say that it is necessary for an orator to be an honest man.'

In example (1), the modal marker is the infinitive *oportere*, and its scope is the subordinate *eum, qui sit orator, virum bonum esse* (see fig. 1)[3]. The scope contains the participant in the state of affairs, *eum, qui sit orator*, that is annotated as an 'animate' entity. After having considered the passage (in which Quintilian highlights the moral qualities of a good orator, that is, to be a decent and honest man), we decided to interpret the relation between marker and scope in terms of deontic acceptability [see Nuyts, 2016:36] and choose the tag 'deontic - acceptability - absolutely necessary'. Quintilian is in fact suggesting that it is morally mandatory for a man who wants to be an orator to have these qualities.

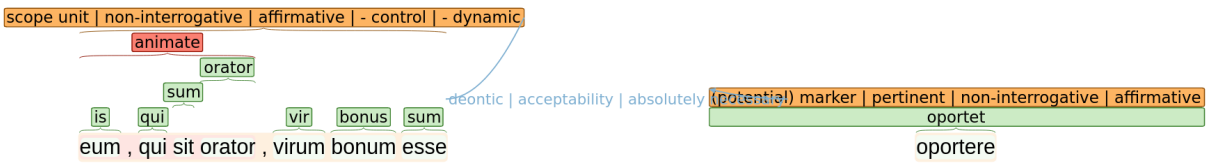

Figure 1. The annotation of example (1) in the INCEpTION platform.

One of the distinctive features of the WoPoss annotation is the possibility for the annotator to describe a passage as potentially conveying different modal readings. This is very important from a theoretical point of view, as the annotation of ambiguous passages will allow the team (and future users) to appreciate ambiguity as one of the potential triggers of semantic shift. The next example (2) illustrates a case of double annotation of the same passage.

(2) PETRONIUS, Satyricon 71, § 11, p. 68, lin. 26 *et unam licet fractam sculpas* [...]

et	*un-am*	*licet*	*fract-am*	*sculp-as*
and	one-ACC.F.SG	may.PRS.IND.3SG	broken-ACC.F.SG	carve.PRS-SBJV.2SG

[2] See Annex 1 for a list of abbreviations according to the Leipzig Glossing Rules [Comrie et al. 2008].

[3] For reasons of space in this figure and in figure 3 we have deleted the layers 'morphological features' and 'parts of speech', but these can still be seen in figure 2.

‘And you may carve a broken one […]’[4]

In the passage from which this example is taken, Trimalchio is giving instructions for his funerary monument. Such instructions can generally be conceived as orders to the sculptor who is present in this scene. As the verb *licet* basically indicates ‘possibility’, the reading here is deontic ‘permission’ (‘you are allowed to carve a broken jar’), where Trimalchio is the source of authority. Nevertheless, in this specific passage, the idea of carving a broken jar can also be seen just as a suggestion to embellish the monument. This reading can then be conceived in terms of the desirability (acceptability) of the imagined state of affairs.
In cases of ambiguity such as this, we perform a double annotation: when connecting the marker *licet* to the scope *unam […] fractam sculpas*, the annotator draws two different modal relations between them, one for each modal reading.

Figure 2. The annotation of example (2) in the INCEpTION platform.

This passage also presents a discontinuous scope, as the marker is positioned between ‘*unam*’ and ‘*fractam sculpas*’. Discontinuous elements are annotated by introducing a new layer of annotation (see layer ‘scope’ in green in fig. 2) and drawing an arrow (see line ‘scope’ in fig. 2) to connect the two elements.

As mentioned above, it can happen that both the scope and the marker are discontinuous. In the next example (3) we illustrate how to tackle this problem.

(3) QUINTILIAN, Institutio Oratoria 12, 1 § 44, p. 376, lin. 7 [...] *istud illi quem formamus viro saepe sit faciendum.*

ist-ud	*ill-i*	*qu-em*	*forma-mus*
this-ACC.N.SG	that-DAT.M.SG	who-ACC.M.SG	shape.PRS-1PL

vir-o	*saepe*	*si-t*	*facie-nd-um.*
man-DAT.M.SG	often	be.SBJV.PRS-3SG	do.PRS-GDV-ACC.N.SG

‘[We do not say that] the orator we are shaping often has to do this.’

Example (3) presents a discontinuous marker (*sit … -ndum*) and it needs to be annotated by introducing a new layer of annotation (see layer ‘Marker’ in red in fig. 3) and drawing an arrow (see line ‘Marker’ in fig. 3) to connect the two elements. The same holds for the scope of this passage which is discontinuous too: *istud illi, quem formamus, viro saepe* and *facie-* are connected by using the layer ‘Scope’ (see fig. 3). Inside the scope there are also two participants: *istud* (annotated as ‘inanimate - patient’) and *illi, quem formamus, viro* (which is the agent, annotated as ‘animate’).

[4] The text of the Perseus Digital Library has *urnam* instead of *unam* of the Teubner reference edition [Petronius, 2013]. In such cases, the annotator enters the correct text in the field ‘Note’ for the post-processing phase.

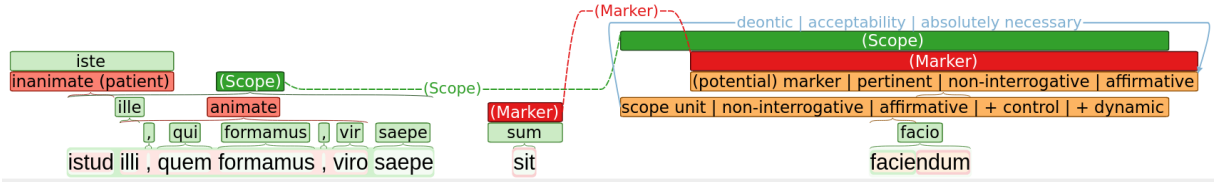

Figure 3. The annotation of example (3) in the INCEpTION platform.

III THE POST-PROCESSING STAGE AND BEYOND

3.1 Planned pipeline after the manual fine-grained annotation

After the manual annotation, we plan to enrich the dataset (see Annex 2 for details). The annotated files will be exported in XMI format and then transformed according to the TEI standards. We will use stand-off annotation methods to preserve all the multiple layers of linguistic annotation. The following layers will be added:

a) Annotation of the most ancient meaning of each modal marker.
b) Transformation of the pseudo mark-up into the correspondent TEI elements.
c) Addition of metadata for each work concerning chronology, genre, type of transmission and authorship information. Metadata will be added automatically thanks to the interaction between WoPoss and the Digital Humanities Toolkit [Picca and Egloff, 2017].

Subsequently, the TEI dataset will be stored in a No-SQL database [eXist-DB] and it will be freely accessible through a user-friendly interface.

3.2 Data sharing and preservation

The WoPoss project is developed through a version control system [Git] stored in an open repository [WoPoss-GitHub]. Thus, every stage of the pipeline is documented there and every written script is reusable. In addition, we plan to publish the dataset in an open repository [Zenodo].

An important step in terms of data sharing is the integration of our results in the [LiLa] Knowledge Base thus profiting from the advantages of the linked open data ecosystem [Breitman et al., 2007]. Through the use of standards and the practice of multi-publishing, we aim to render the results of our research sharable, extensible, and easily re-usable.

Conclusions

In this paper we have shown how the WoPoss project has been conceived to be an open science product in all of its implementation stages.

References

Bermúdez Sabel, H. Digital tools for semantic annotation: the WoPoss use case. *Bulletin de linguistique et des sciences du langage*. In press;30. [Preprint version: https://zenodo.org/record/3572410]

Breitman, K.K., Casanova, M.A. and Truszkowski, W. *Semantic Web: concepts, technologies and applications*. Springer (London), 2007.

CC BY 4.0 https://creativecommons.org/licenses/by/4.0

Comrie, B., Haspelmath, M., and Bickel, B. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig (Leipzig), 2008.

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

Corpus corporum http://www.mlat.uzh.ch/MLS/

Coseriu, E. *Einführung in die strukturelle Linguistik*. Autorisierte Nachschrift besorgt von Gunter Narr und Rudolf Windisch (Tübingen), 1969.

Dell'Oro, F. *WoPoss guidelines for annotation*. Zenodo. 2019. doi:10.5281/zenodo.3560951

digilibLT http://digiliblt.lett.unipmn.it/

EAGLE-IDEA https://www.eagle-network.eu/about/who-we-are/
EpiDoc Guidelines http://www.stoa.org/epidoc/gl/latest/
eXist-DB http://exist-db.org/
Git https://git-scm.com/
INCEpTION https://inception-project.github.io/
Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. and Gurevych, I. The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, Association for Computational Linguistics, 2018:5–9. https://www.aclweb.org/anthology/C18-2002/
LiLa https://lila-erc.eu
Nuyts, J. Analyses of the Modal Meanings. *The Oxford Handbook of Modality and Mood.* Oxford University Press (Oxford), 2016:31–49. https://doi.org/10.1093/oxfordhb/9780199591435.013.1
Open Greek & Latin Project http://www.opengreekandlatin.org/
Perseus Digital Library http://www.perseus.tufts.edu
Petronius [Arbiter]. *Satyrica.* De Gruyter (Berlin, Boston), 2013. Bibliotheca Teubneriana Latina (BTL) Online. https://www.degruyter.com/view/BTL/APETRTSAT/203477
Picca, D. and Egloff, M. DHTK: The Digital Humanities ToolKit. *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II).* 2017:81–86. http://ceur-ws.org/Vol-2014/paper-09.pdf
Roman Law Library https://droitromain.univ-grenoble-alpes.fr/
StanfordNLP https://stanfordnlp.github.io/stanfordnlp/
TEI https://tei-c.org/
TEI-lite https://tei-c.org/guidelines/customization/lite/
WoPoss http://woposs.unil.ch/
WoPoss-credits http://woposs.unil.ch/credits.php
WoPoss-GitHub https://github.com/WoPoss
Zenodo https://zenodo.org/

ANNEX 1

List of abbreviations used according to the Leipzig Glossing Rules:
1 first person
2 second person
3 third person
ACC accusative
DAT dative
F feminine
IND indicative
INF infinitive
M masculine
N neuter
NOM nominative
PL plural
PRS present
SBJV subjunctive
SG singular

ANNEX 2

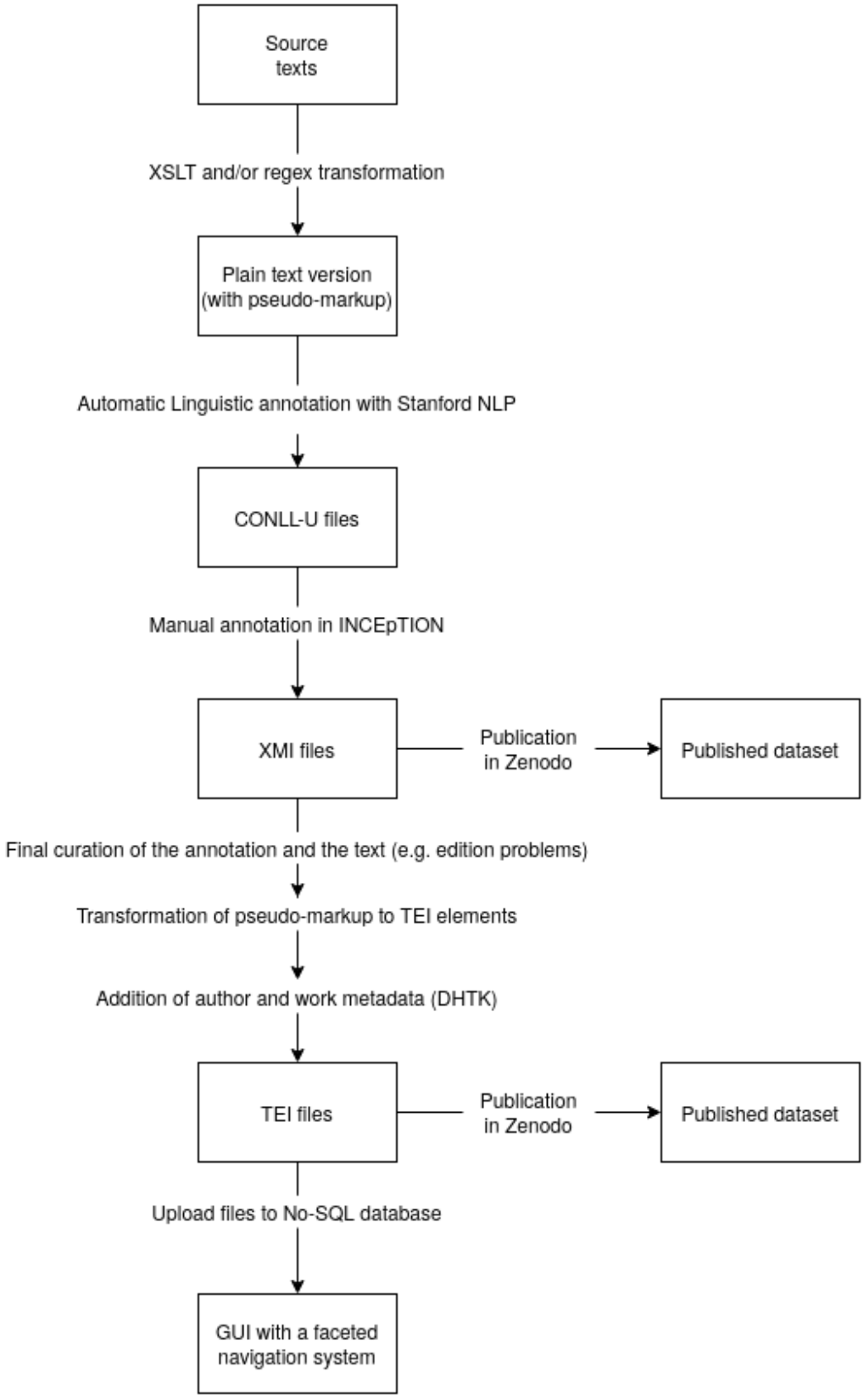

Figure 4. WoPoss workflow