

Linguistic annotation of Ancient Greek and Latin inscriptions

Francesca Dell'Oro – University of Neuchâtel

francesca.delloro@unine.ch

Outline

1. Linguistic annotation: what is it?
2. Treebank of Euboean Greek inscriptions of Sicily through manual annotation (Arethusa)
3. Annotation of (Latin) inscriptions in the framework of the WoPoss project (focus: modality) through both types of annotation, manual and automatic
4. Some ideas to explore a linguistically annotated corpus
5. Exercise

Preliminary remarks

1. linguistic annotation of Ancient Greek and Latin inscriptions is developing right now
2. there is no standard (yet) for the linguistic annotation (treebank) of Ancient Greek and Latin inscriptions

Linguistic annotation in general: what is it?

- core linguistic annotation
 - 1) parts of speech (POS): N, V, A, etc.
 - 2) lemma: *possum*, *potes*, *possim*, *potuerunt*, etc. → 'possum'
 - 3) morphology: person, gender, number, case, tense, mood, aspect...
 - 4) syntactic relations (dependencies)
 - → treebank
- less frequently annotated levels of linguistic analysis:
 - phonetics/phonology: how sounds are pronounced (vowels: short/long, open/close and so on)?
 - spelling (ancient languages)
 - semantics: e.g. modality, Smyth Grammar layer (genitive of possession, genitive of separation, etc.), animacy
 - pragmatics
- onomastics: named entities

Some already extant examples
of linguistic annotation of
Ancient Greek and Latin
inscriptions

Digital Marmor Parium

- a complete analysis and study of this Hellenistic inscription entails also a linguistic analysis
- <http://www.digitalmarmorparium.org/linguistics.html>
- → pioneering annotation (treebank)
 - POS
 - Lemma
 - Morphology
 - Syntactic dependencies
 - Semantics (Smyth Grammar layer)

CLaSSES- Corpus for Latin Sociolinguistic Studies on Epigraphic texts

- spelling variants as “clues for phonetic-phonological (and morpho-phonological) variation” (De Felice, Marotta & Donati 2015: 127)
 - the user can search for a predefined set of linguistic phenomena
- <http://classes-latin-linguistics.fileli.unipi.it/en>
- manual annotation
 - lemma
 - spelling: annotation of non-classical variants
 - vowels
 - consonants
 - morpho-phonology

Table 2

Classification of the 1869 words constituting CLaSSES I according to their graphic form.

<i>Graphic form</i>							
<i>complete</i>	<i>abbreviat.</i>	<i>incomplete</i>	<i>integrated</i>	<i>misspelling</i>	<i>uncertain</i>	<i>number</i>	<i>(lacunae)</i>
1017	560	153	28	12	56	9	34
54.4%	30%	8.2%	1.5%	0.6%	3%	0.5%	1.8%

De Felice, Marotta & Donati (2015: 127)

A treebank of Euboean inscriptions from Sicily

For an introduction to Arethusa:

<https://github.com/SunoikisisDC/SunoikisisDC-2019-2020/wiki/DC-Session-6-Treebanking-1>

The guidelines of the Ancient Greek Dependency Treebank 2.0 have been written to annotate Ancient Greek texts. The epigraphic texts, however, pose a challenge for those carrying out morphosyntactic annotation: **should we remain as close as possible to the actual epigraphic text, or represent it in an interpreted and normalized version? How should all epigraphic peculiarities which do not have standard editorial representation, such as, for example, punctuation marks, be treated?** A small corpus such as that of the inscriptions of the Euboean colonies of Sicily of the archaic and classical period has allowed us to test different options and evaluate the annotation challenges. This contribution is the result of a discussion about the advantages and disadvantages of often opposed annotation possibilities.

Dell'Oro - Celano (2019: 1)

Why to work with short inscriptions as a test-case?

- with short inscriptions, you are more often constrained to take into account the extra-linguistic context
 - inscribed object to be read in its extra-linguistic context
- with short inscriptions, hopefully you will have to deal with a problem at a time

Greek local alphabets and dialectal variants

- If you normalize, you can use the MATE tagger and Morpheus morphological analyzer (Arethusa)
- If you decide not to normalize the text, you will have to annotate PoS and morphology by yourself
- “Manuscripts attest the form Ζάγκλη for the ancient name of Messene, but inscriptions and coin legends attest the same name in the form Δάνκλη (e.g., IGASMG III 39). Morpheus cannot recognize Δάνκλη as the same lemma as Ζάγκλη. In this case, it is not necessary to introduce a second lemma: the annotator can specify that the lemma is the same, in that Ζάγκλη is taken to be a dialectal variant of Δάνκλη. This is in accordance with standard practice in non-dialectal Ancient Greek dictionaries (e.g. Liddell and Scott, 1996).”

Ellipsis

- IGDS I 14b: Ἡρακλείδα

1. look for a template in other inscriptions → speaking objects: IGDS I 14a: [H]εκαταίο ἐ[μί] 'I am of Hekataios'
2. determine what is missing: ἐ[μί] 'I am...'
3. Add it as an elliptical node

IGDS I 14a: I am of [PROPER NAME (gen)]

[H]εκαταίο **ἐ[μῖ]**

selection **none** 0 unused highlight unused

[ROOT]
|
PRED
|
ἐ[μῖ]
|
PNOM
|
[H]εκαταίο

morph relation SG aT selector history

comments ⚙

ἐ[μῖ] ¹⁻²

- verb
 - finite verb
 - independent
 - statement

Select Smyth Categories▾

IGDS I 14b: ? of PROPER NAME (gen)

Ἡρακλείδα εἰμί

selection none 0 unused highlight unused

[ROOT]
|
PRED
|
εἰμί
|
PNOM
|
Ἡρακλείδα

morph relation SG aT selector history comments



εἰμί^{1-1e}

- verb
 - finite verb
 - independent
 - statement

Select Smyth Categories▼

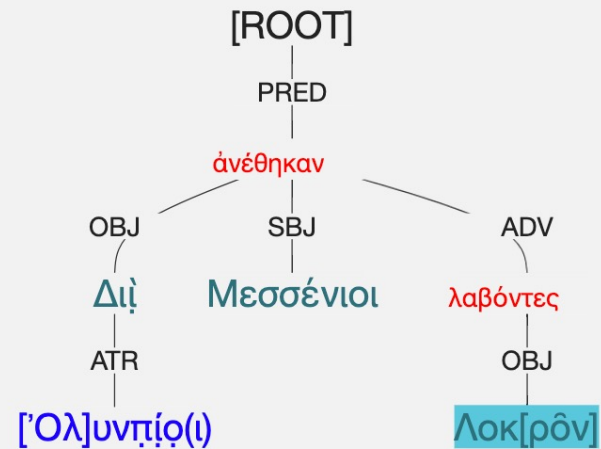
IGDS I 4a

- dedication on leg armor: Διὶ [Ὀλ]υνπῖο(ι) Μεσσένιοι Λοκ[ρῶν]
- Lazzarini (1976: 317): nominative (victorious people who made the dedication) + genitive (the defeated) + λαβόντες or ἐλόντες
- comparison with
 - IG I3 1467 (Olympia): Διὶ Ἀθηναῖοι Μέδον λαβόντες “The Athenians (dedicated this taken) from the Medes”
 - but see also: IGASMG V 13a (Olympia): σκῦλα ἀπὸ Θουρίων Ταραντῖνοι ἀνέθεκαν Διὶ Ὀλυμπίοι δεκάταν “The Tarantines dedicated the spoils (taken) from the Thurians to Olympian Zeus as a tithe”

IGDS I 4a

Διὶ [Ὀλ]υντιῶ(ι) Μεσσήνιοι Λοκ[ρὸν] ἀνέθηκαν λαβόντες

selection **none** 0 unused highlight unused



morph relation **SG** aT selector history



Λοκ[ρὸν] ¹⁻⁵

- noun
 - genitive
 - dependent
 - ablative
 - separation
 - concrete

Select Smyth Categories▼

Punctuation

- Morpurgo Davies (1987)
 - accentual and/or prosodic units: enclitics and proclitics are not separated from 'their' word by punctuation

IGASMG I, 75 (= IGDS I, 1)



graffito on the foot of an Attic kylix

end of the 5th century BCE

omicron → omega in the common alphabet

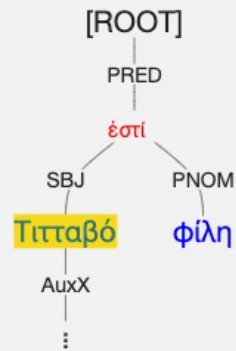
but there is already heta

three points = punctuation → how to interpret it?
How to annotate it according to AGDT 2.1?

One possible interpretation: *Tittabó, (she) is dear (to me)*

Τιτταβό : φίλη ἐστὶ

selection none 0 unused highlight unused



morph relation SG aT selector history comments



Τιτταβό ¹⁻¹

- noun
 - nominative
 - dependent

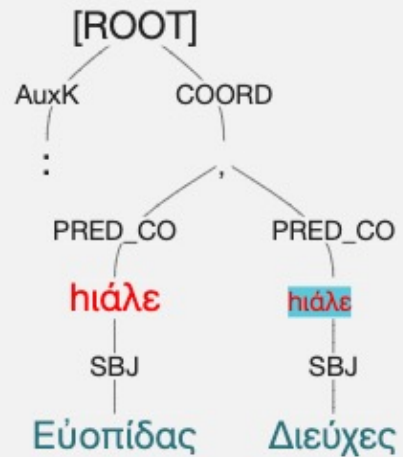
Select Smyth Categories▾

- other possible interpretations: *Tittabó! (She) is dear (to me) / (you) are dear (to me)*
- introduction of a special tag for inscriptional punctuation?

IGDS | 11, 1

Εὐοπίδας **ηιάλε** Διεύχες : , **ηιάλε**

selection **none** 0 unused highlight unused



morph relation **SG** aT selector history



ηιάλε ^{1-5e}

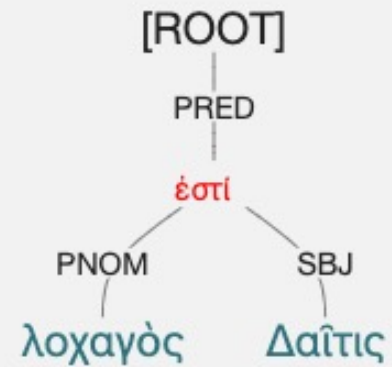
- verb
 - finite verb
 - independent
 - statement

Select Smyth Categories▼

IGDS I 11, 2

λοχαγός Δαίτις ἐστί

selection **none** 0 unused highlight unused



The WoPoss project: automatic and manual linguistic annotation

**Annotation of Latin inscriptions from lemmatisation to
semantics/pragmatics**

<https://woposs.unine.ch/>

WoPoss. A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language

The SNSF-funded project *A world of possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language* (WoPoss) aims at reconstructing **the evolution of modal meanings from the prehistory of the Latin language up to the 7th century CE**. The WoPoss team is working on **the linguistic annotation of a selection of modal markers in a diachronic corpus of Latin literary and documentary texts**.

What is modality?

Expression of the notional domains of necessity, possibility and volition.

- lexical markers: *possum, debeo, volo, forsitan...*
 - morphological markers: *-bilis, -ndus, -turus*
 - (NB. mood: sentence types, realis/irrealis, indicative/subjunctive)
-
- see I. Nuyts & J. van der Auwera 2016 on the distinction ‘modality’ and ‘mood’

Automatic annotation

We use models trained on the treebank data in order to carry out an automatic annotation of:

- PoS

- lemmas

- morphological features

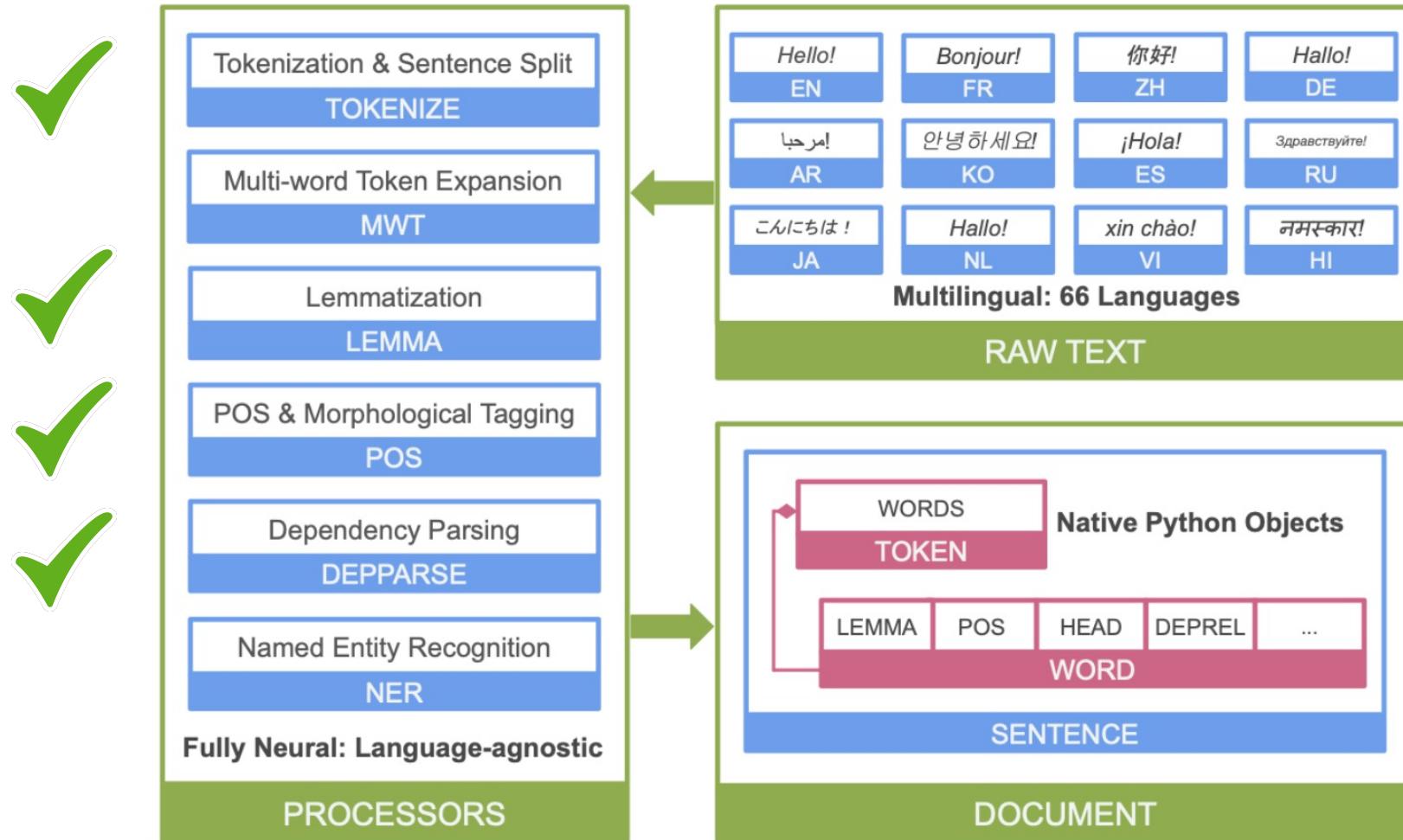
- dependencies

Stanza – A Python NLP Package for Many Human Languages

“Stanza is a Python natural language analysis package. It contains tools, which can be used in a pipeline, to convert a string containing human language text into lists of sentences and words, to generate base forms of those words, their parts of speech and morphological features, to give a syntactic structure dependency parse, and to recognize named entities. The toolkit is designed to be parallel among more than 60 languages, using the Universal Dependencies formalism.”

- <https://universaldependencies.org>

How Stanza works...



Processors work based on pretrained models

Models were trained on the Universal Dependencies treebanks

Latin: ITTB, PROIEL and Perseus

Ancient Greek: PROIEL and Perseus

For all models https://stanfordnlp.github.io/stanza/available_models.html

You can try it yourself:

https://github.com/WoPoss/automatic_annotation

From the already extant treebanks to WoPoss

Treebanks of Ancient Greek and Latin



Stanford NLP used them to create models for Stanza



The WoPoss team uses Stanza to automatically annotate other texts which are not part of those treebanks

Senatus consultum de Bacchanalibus, 4-5: ...*utei ad praitorem urbanum Romam venirent* 'they are to come to Rome to the praetor urbanus'

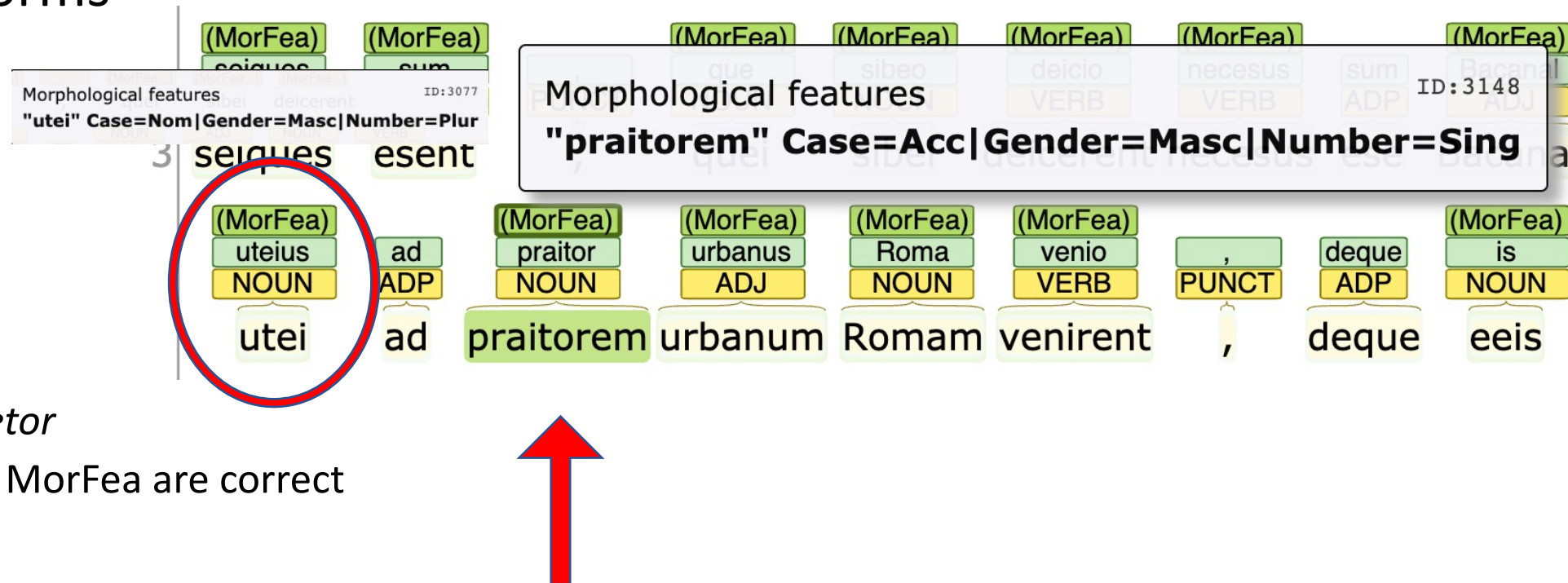
- many archaic forms

- *utei*

- PoS SCONJ
 - Lemma *ut*
 - No MorFea

- *praitor*

- Lemma *praetor*
 - But PoS and MorFea are correct



Manual annotation

- semantics
 - annotation of modality: by using the *WoPoss guidelines for annotation*
 - <https://zenodo.org/badge/DOI/10.5281/zenodo.3560951.svg>
- Inception platform

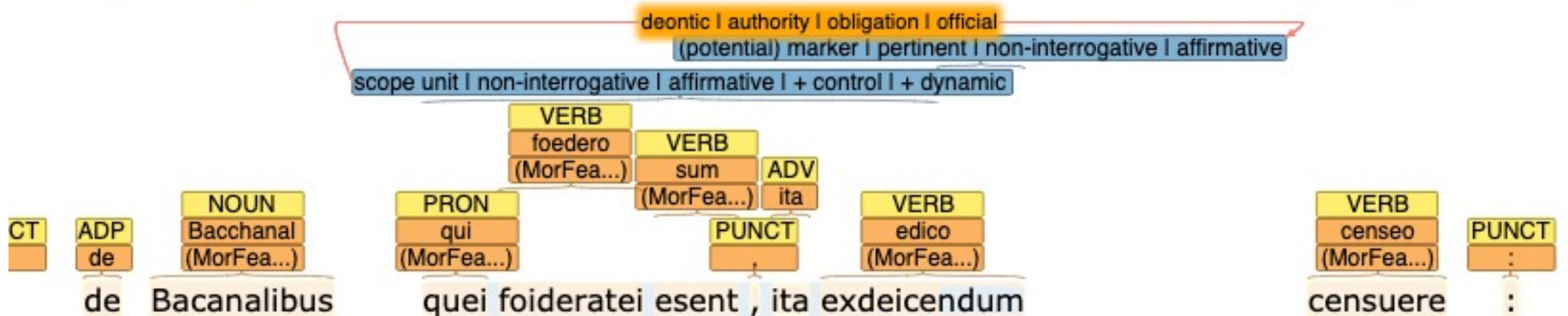
Types of modality

- **dynamic** modality: (1) abilities / needs; (2) external circumstances; (3) generic possibility / necessity
 - (1) *He **can** stand on his head without using his hands.*
 - (2) *The garage is free so you **can** park your car there.*
 - (3) *It **can** rain here every day in winter.*
- **deontic** modality: (1) obligation / permission; (2) moral desirability; (3) volition and intention
 - (2) *This initiative of the federal government is highly **deplorable**.*
 - (3) *I **want** to hear the whole story.*
- **epistemic** modality: *That's **probably** the postman bringing today's newspaper.*

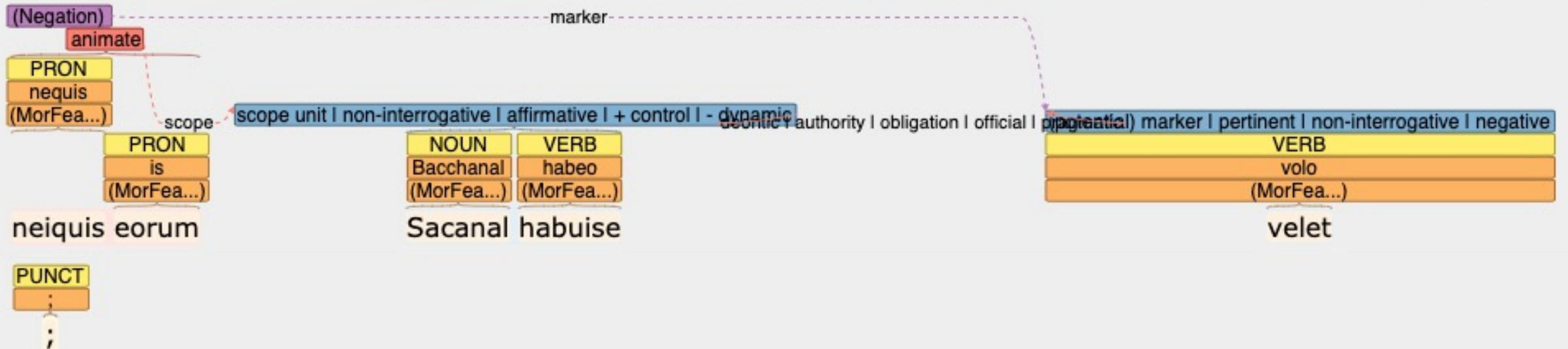
archaic Latin: *De Bacanalibus, quei foideratei esent, ita exdeicendum censuere...*

classical Latin: *De Bacchanalibus (iis) qui foederati essent, ita exdicendum (esse) censuerunt...*

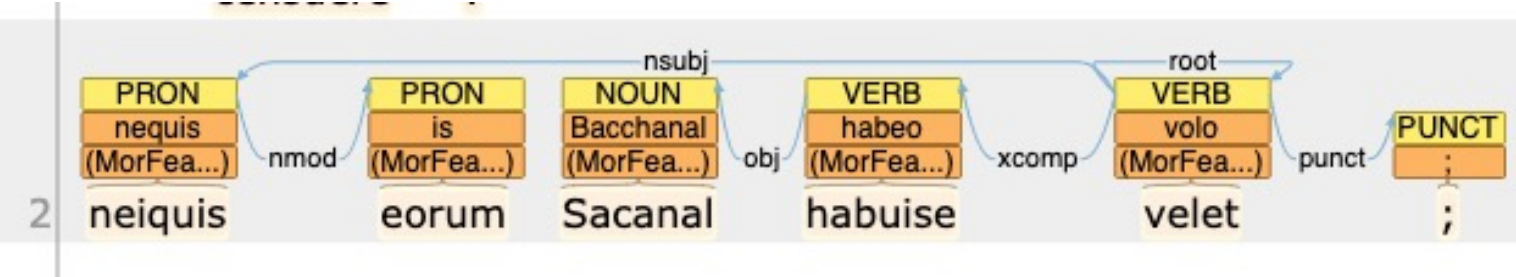
‘Regarding the Bacchanalia, they advised that **it was necessary** to issue an edict as follows to those who are in alliance (with us)...’



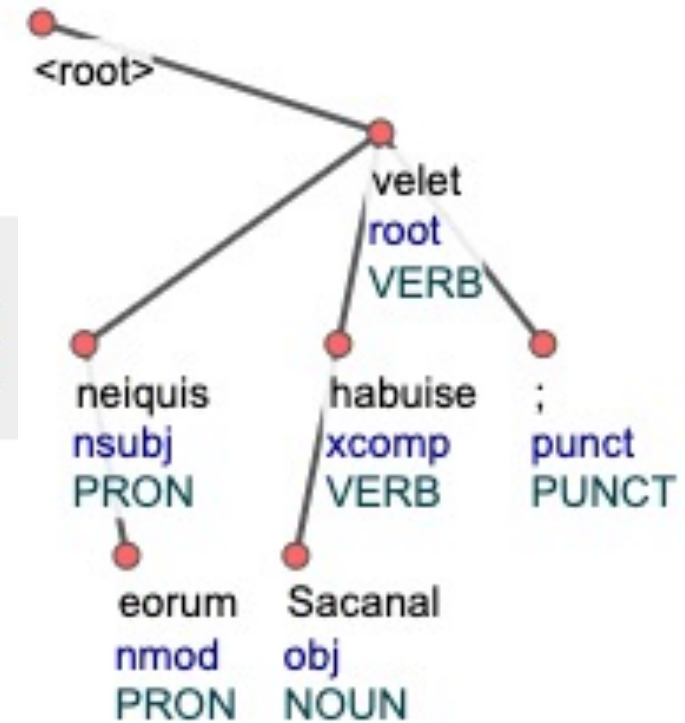
archaic Latin: ...*neiquis eorum Sacanal habuisse velet*;
 classical Latin: ...*nequis eorum Bacchanal habuisse vellet*;
 ‘No one of them should have a sanctuary’



Syntactic dependencies



nequis eorum Sacanal habuisse velet ;



Towards corpus research
with inscriptions: TXM

- TXM textométrie
 - big corpora
 - <http://textometrie.ens-lyon.fr/>
- research questions: which are the most common co-occurents of modal markers? Are there personal pronouns among these co-occurents of impersonal modal markers (e.g. *oportet*)? Is co-occurrence meaningful from a statistical point of view?
 - specificity score (≥ 3) <http://txm.sourceforge.net/doc/manual/0.7.9/fr/manua159.xhtml> (French) or <http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf> (English)
 - left context: between 0 and 4 words
 - right context: between 0 and 4 words

Corpus (EDH)

- automatic annotation (Treetagger)
 - tokenisation
 - lemmatisation
 - POS tagging (with basic morphological analysis)
- the annotation needs to be tested and corrected
- the corpus is not homogeneous from the chronological point of view
- reference corpus: Cicero

Requête



[latinlemma="oporteo"]

Paramètres

Propriétés des cooccurents : latinlemma

Éditer

Seuils : Fmin ≥

0

Cmin ≥

0

Indice ≥

0,0

Contexte : ☒ forme ☐ structure

s

☒ Utiliser le contexte gauche☐ Utiliser le contexte droit☐ inclure la structure contenant le pivot dans le décompte

de -

4

à -

0

et de

0

à

0

Cooccurrent	Fréquence	CoFréquence	Indice v	Distance moyenne
non	17697	164	28	.8
puto	2298	54	26	1.2
fio	1586	35	16	1.1
neao	774	22	13	.5
specto	211	13	12	.4
lex	1585	28	11	1.4
oporteo	930	21	10	2.7
sumo	383	13	8	.7
dico	8552	65	8	1.5
restituo	218	9	7	.7
puteolouto	406	11	6	.7
uti	306	9	5	1.1
credo	778	12	4	.2
quoque	802	12	4	1.3
pulsio	3	2	4	1.0
censeo	462	9	4	1.2
depulsio	22	3	4	2.0
accusator	282	7	4	1.6
intentio	26	3	4	2.0
prius	72	4	3	.8
valeo	804	11	3	.2
scio	1260	14	3	1.1
iudicationem	8	2	3	.0
discento	12	3	3	0

t pivot 930, v cooc 1026, t cooc 3275, T corpus 1370127

Results for the specificity score (≥ 3)

Inscriptions (EDH)

- EGO OPORTET 1
- OPORTET EGO 0
- TU OPORTET –
- OPORTET TU –
- NOS OPORTET –
- OPORTET NOS –
- VOS OPORTET –
- OPORTET VOS –
- IS OPORTET 7
- OPORTET IS 13

Cicero

- EGO OPORTET 1
- OPORTET EGO 0
- TU OPORTET 2
- OPORTET TU 0
- NOS OPORTET 2
- OPORTET NOS –
- VOS OPORTET 1
- OPORTET VOS –
- IS OPORTET 3
- OPORTET IS 4

Exercise

- annotation of a modal passage of an inscription (EDR 150769) in Arethusa
 - .txt file
 - .xml file
- annotation of a modal passage in the Inception Sandbox
 - .zip file

Acknowledgements

- Swiss National Science Foundation: project n° 176778 (<http://p3.snf.ch/project-176778>)
- Center for Hellenic Studies (<https://chs.harvard.edu/>)
- EAGLE-IDEA (<https://www.eagle-network.eu/category/idea-association/>)

The WoPoss annotators



References

- Bermúdez Sabel, Helena (in press), “Digital tools for semantic annotation: the WoPoss use case. Bulletin de linguistique et des sciences du langage 30. [Preprint version: <https://zenodo.org/record/3572410>]
- Bermúdez Sabel, Helena; Dell’Oro, Francesca (2020). “[Automatic annotation of Latin and Greek texts](#)”. GitHub.
- Celano, Giuseppe G.A., Gregory Crane, and Saeed Majidi, 2016, « Part of Speech Tagging for Ancient Greek », Open Linguistics 2: 393–399.
- Berti, Monica (dir.), *Digital Marmor Parium*: <https://www.dh.uni-leipzig.de/wo/dmp/>
- Celano, Giuseppe G.A., 2014, *Guidelines for the Annotation of the Ancient Greek Dependency Treebank 2.0*, https://github.com/PerseusDL/treebank_data/tree/master/AGDT2/guidelines
- Dell’Oro, Francesca, “What Role for Inscriptions in the Study of Syntax and Syntactic Change in the Old Indo-European Languages? The Pros and Cons of an Integration of Epigraphic Corpora”, in C. Viti (dir.), *Perspectives on Historical Syntax*, Amsterdam, Benjamins, 271–290.
- Dell’Oro, Francesca & Celano, Giuseppe G.A., , « Epigraphic treebanks : some considerations from a work in progress », *FirstDrafts@Classics* – Harvard’s CHS, 2019 : <https://chs.harvard.edu/CHS/article/display/1304>
- De Felice, Irene; Marotta, Giovanna & Donati, Margherita, “CLaSSES: a New Digital Resource for Latin Epigraphy”, IJCoL [Online], 1-1 | 2015, Online since 01 December 2015, connection on 28 January 2021. URL: <http://journals.openedition.org/ijcol/331> ; DOI: <https://doi.org/10.4000/ijcol.331>
- EAGLE: <https://www.eagle-network.eu/basic-search/>

IGDS I = Laurent DUBOIS (dir.), 1989, *Inscriptions grecques dialectales de Sicile: contribution à l'étude du vocabulaire grec colonial*, Rome, École Française de Rome.

INCEPTION: <https://inception-project.github.io/>

Lazzarini, Maria Letizia, 1976, *Le formule delle dediche votive nella Grecia arcaica*, Roma : Accademia nazionale dei Lincei.

Morpurgo Davies, Anna, 1987, "Folk-linguistics and the Greek word", in G. Cardona & N. H. Zide (dir.), *Festschrift für Henry Hoenigswald. on the occasion of his seventieth birthday*, *Ars linguistica* 15, Tübingen, Narr, 1987, 263–280.

Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason & Manning, Christopher D.. 2020. [Stanza: A Python Natural Language Processing Toolkit for Many Human Languages](#). In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. [[pdf](#)][[bib](#)]

Universal Dependencies: <https://universaldependencies.org/>

UD Greek Perseus: https://universaldependencies.org/treebanks/grc_perseus/index.html

UD Greek PROIEL: https://universaldependencies.org/treebanks/grc_proiel/index.html

UD Latin ITTB: https://universaldependencies.org/treebanks/la_ittb/index.html

UD Latin Perseus: https://universaldependencies.org/treebanks/la_perseus/index.html

UD Latin PROIEL: https://universaldependencies.org/treebanks/la_proiel/index.html