SunoikisisDC Autumn 2021
Session 5

Research with Treebanks

Nicole Iu (U of London)
Francesco Mambrini (Milan)
Marja Vierros (Helsinki)

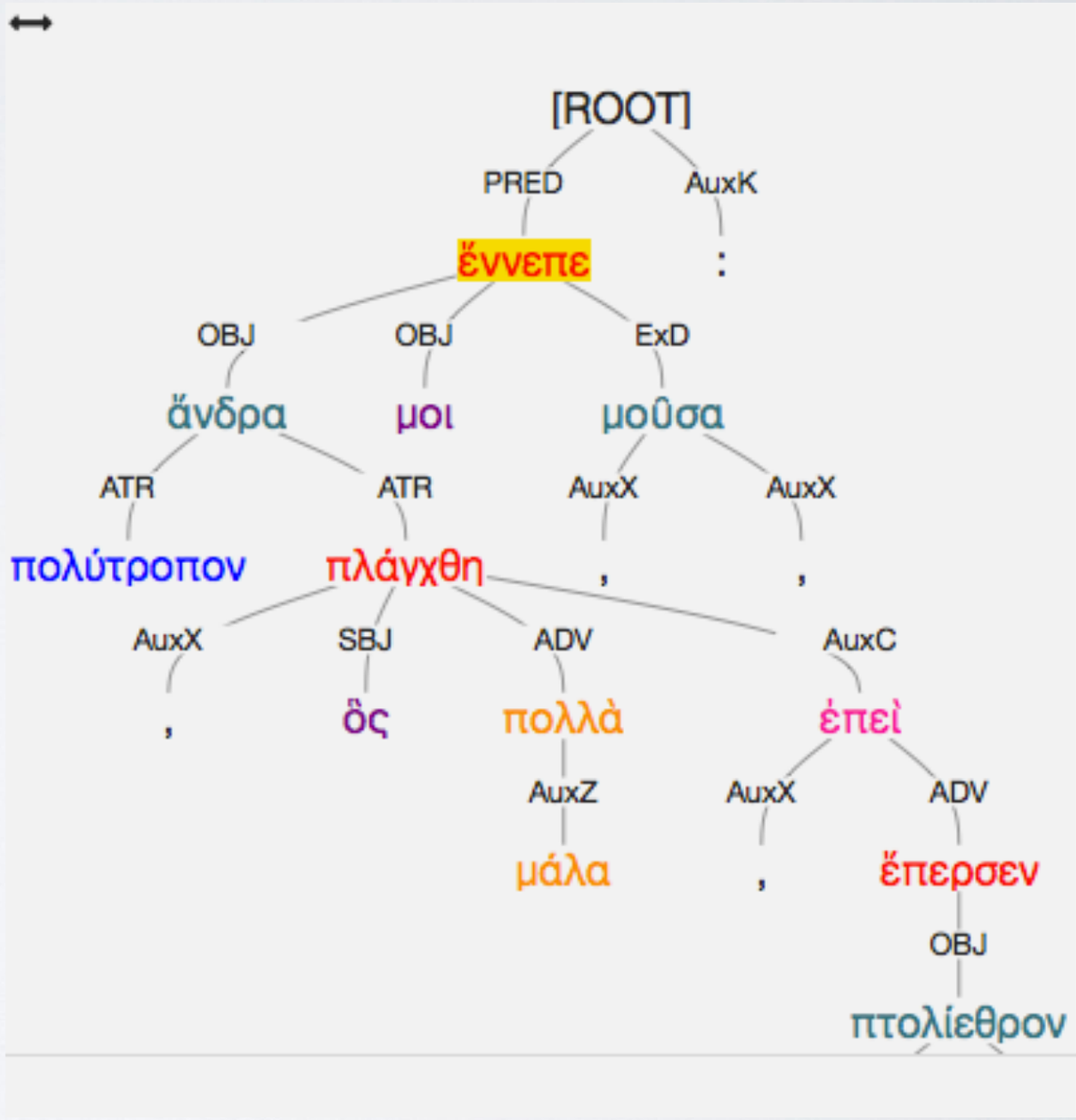
RESEARCH WITH TREEBANKS

Francesco Mambrini

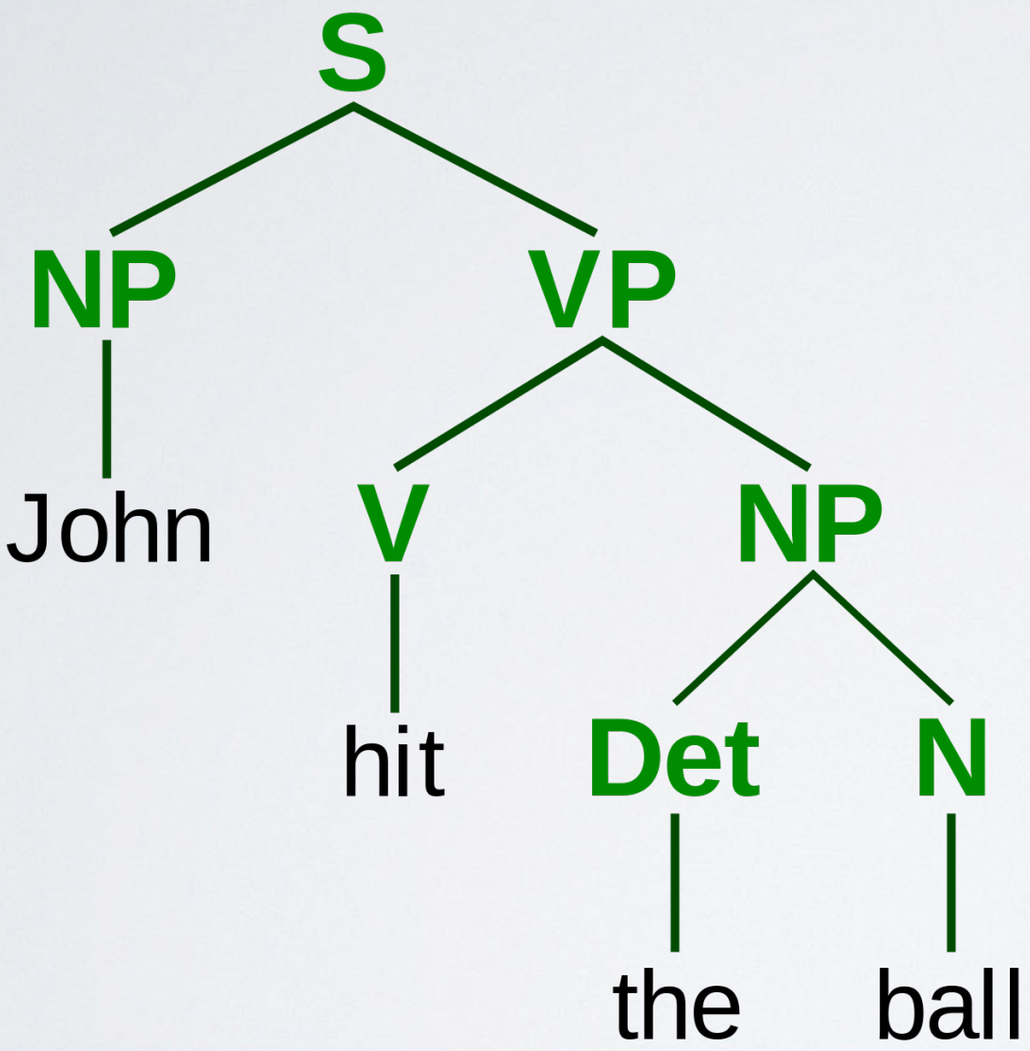
Sunokisis-DC-2021-2022

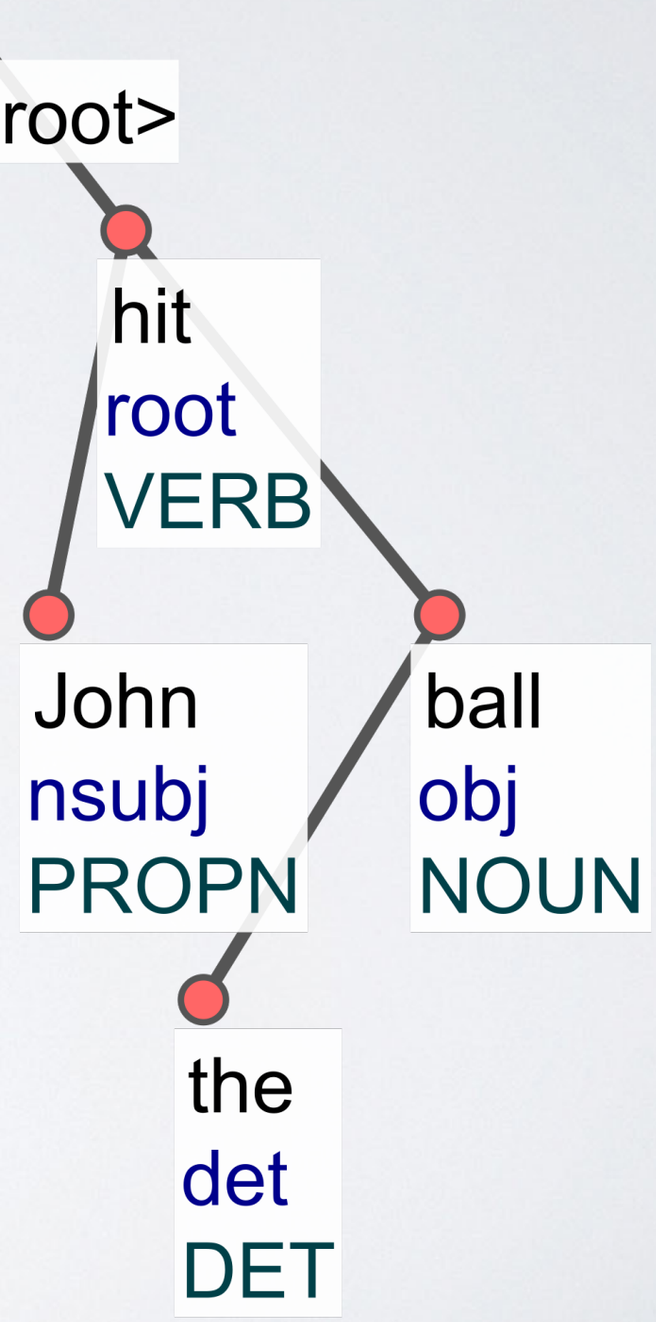
TREEBANKS

- **Annotated corpora**
- with at least two layers of information
  - PoS (morphology)
  - A description of the syntactic structure of the sentence
- Syntax: full description based on some form of theory



DIFFERENT VIEWS





Morphology

```
<sentence subdoc="4-5" id="2386547" document_id="urn:cts:greekLit:tlg0011.tlg004.pe
  <annotator>FrancescoM</annotator>
  <word id="1" form="πόλις" lemma="πόλις" postag="n-s---fn-" head="6" relation="SBJ
  <word id="2" form="θ'" lemma="θέ" postag="g--------" head="6" relation="AuxY" cit
  <word id="3" form="ὁμοῦ" lemma="ὁμοῦ" postag="d--------" head="4" relation="AuxZ"
  <word id="4" form="μὲν" lemma="μέν" postag="g--------" head="9" relation="AuxY" c
  <word id="5" form="θυμιαμάτων" lemma="θυμίαμα" postag="n-p---ng-" head="9" relati
  <word id="6" form="γέμει" lemma="γέμω" postag="v3spia---" head="0" relation="PRED
  <word id="7" form="," lemma="," postag="u--------" head="9" relation="AuxX" cite=
  <word id="8" form="ὁμοῦ" lemma="ὁμοῦ" postag="d--------" head="9" relation="AuxZ"
  <word id="9" form="δὲ" lemma="δέ" postag="g--------" head="6" relation="COORD" ci
  <word id="10" form="παιάνων" lemma="Παιάν" postag="n-p---mg-" head="12" relation=
  <word id="11" form="τε" lemma="τε" postag="g--------" head="12" relation="AuxY" c
  <word id="12" form="καὶ" lemma="καί" postag="c--------" head="9" relation="COORD"
  <word id="13" form="στεναγμάτων" lemma="στέναγμα" postag="n-p---ng-" head="12" re
  <word id="14" form="." lemma="." postag="u--------" head="0" relation="AuxK" cite
</sentence>
```

TEXT

Syntax

DEPENDENCY TREEBANKS

Mainly for Greek and Latin

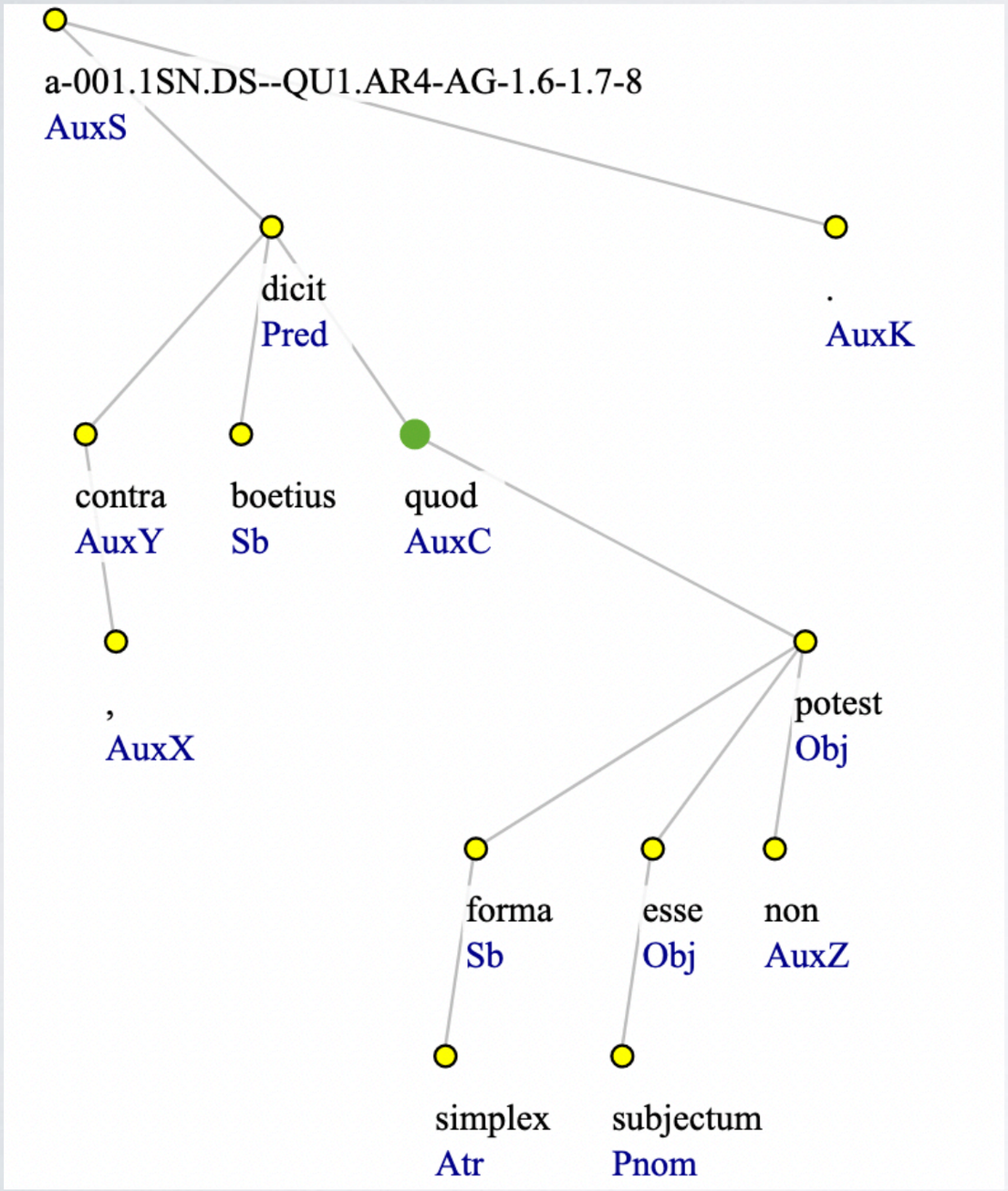
- The "Index Thomisticus Treebank" (inspired by the PDT)
- The "Perseus family" (AGLDT, Pedalion, Gorman Trees, Sematia…)
- PROIEL
- **Universal Dependencies (UD)**

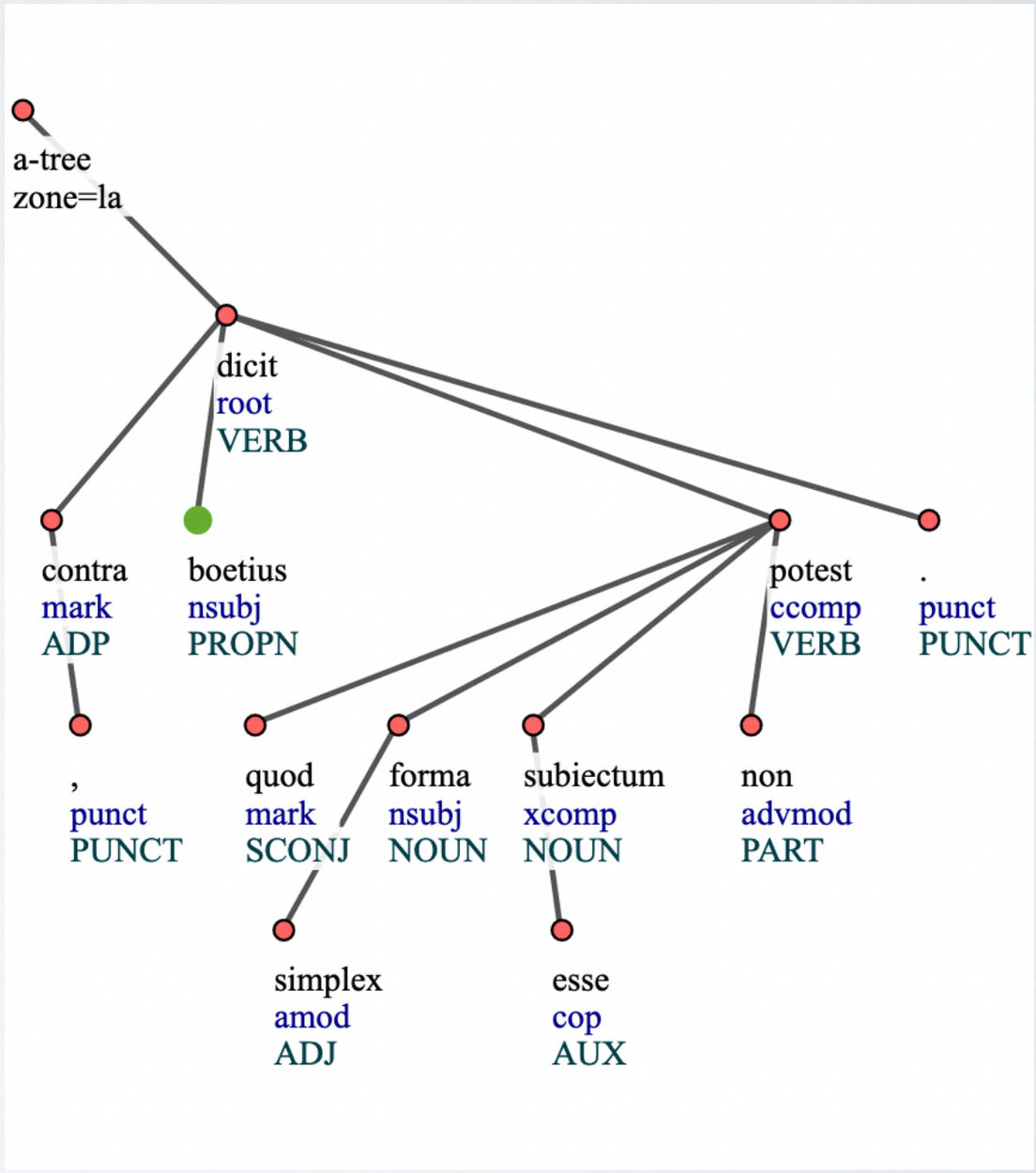
UD

- A large project with shared guidelines for consistent annotation
- More than 200 treebanks in more than 100 languages!
- Both universal and language-specific tags
- But shared framework

universaldependencies.org

PDT-BASED vs UD





Treebanking the Twelve Tables: magical or not?

Nicole Iu

Sunoikisis Digital Classics

November 4th, 2021

General Info on the Twelve Tables

- *Lex XII Tabularum*
- 'Foundation of Roman Law'
- Divided into twelve different sections
- Originally written in the 5th century BC
- Original tablets were destroyed in 390 BC by the Gauls
- Were said to have been transmitted orally and students were taught to recite them aloud
- Cited in the works of Cicero, Horace, and Pliny
- Sometimes quoted verbatim, but often paraphrased
- We know their general content, but original parts are fragmentary
- Potential addenda added to the original laws in 198 BC by jurist, Sextus Aelius Paetus to further clarify some of the laws

Research question: do the Twelve Tables refer to magic or not?

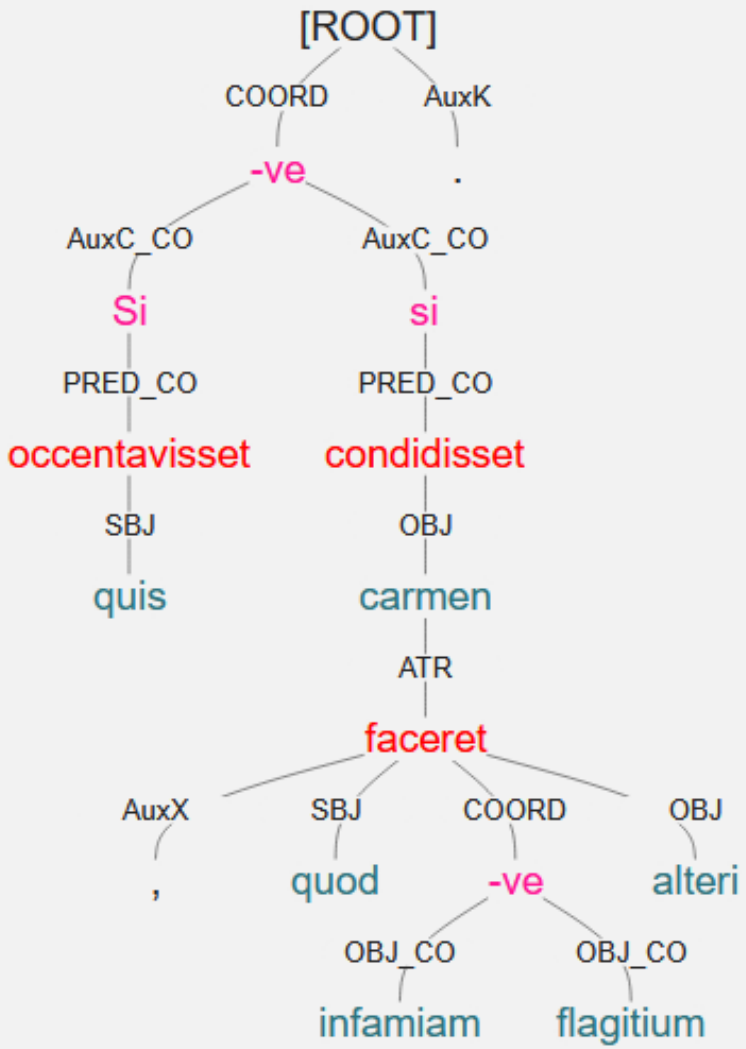
- Two of the laws within the Twelve Tables, VIII 1a–b and 8a–b have been traditionally interpreted as banning magical practices
- 1a: *Si quis occentavisset sive carmen condidisset, quod infamiam faceret flagitiumve alteri.*
- 'Whoever sings a slanderous (?) song or whoever composes a song that causes dishonour or disgrace to another [shall suffer capital punishment]'
- 1b: *Qui malum carmen incantassit*
- 'Whoever sings an evil song'
- Does this law refer to practicing magic (aka spells and curses) or to slandering someone?
- Cicero and Horace refer to it as slander, while Pliny refers to it as magical
- Possible addendum of *quod….alteri*
- This relative clause seems to further clarify that the law was referring to slander rather than magic

A few grammatical notes

- Twelve Tables, VIII 1a–b: *Si quis occentavisset sive carmen condidisset, quod infamiam faceret flagitiumve alteri. Qui malum carmen incantassit…*
- Cicero states A, B, C; Horace states B, D; Pliny states A, D
- Rives 2002: argues that this relative clause was likely added in 198 BC and not part of the original laws
- The *quod…alteri* relative clause is in subjunctive, likely to further qualify the purpose of its antecedent, *carmen*
- Rives further states that such a relative clause that is introduced by *quod* and is subjunctive while dependent on another subjunctive clause (one of the conditional protases) does not appear elsewhere in the Twelve Tables
- In my own search through the Twelve Tables, I have found two other surviving relative clauses introduced by *quod* (Twelve Tables VIII 6)

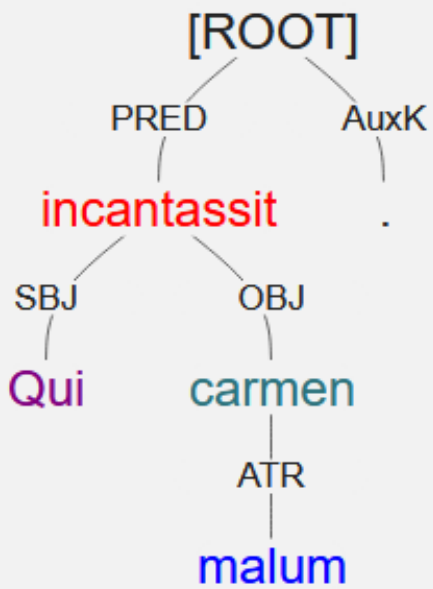
Si quis occentavisset si -ve carmen condidisset , quod infamiam faceret flagitium -ve alteri . (VIII 1a)

selection none 0 unused highlight unused



Qui malum carmen incantassit . (VIII 1b)

selection none 0 unused highlight unused



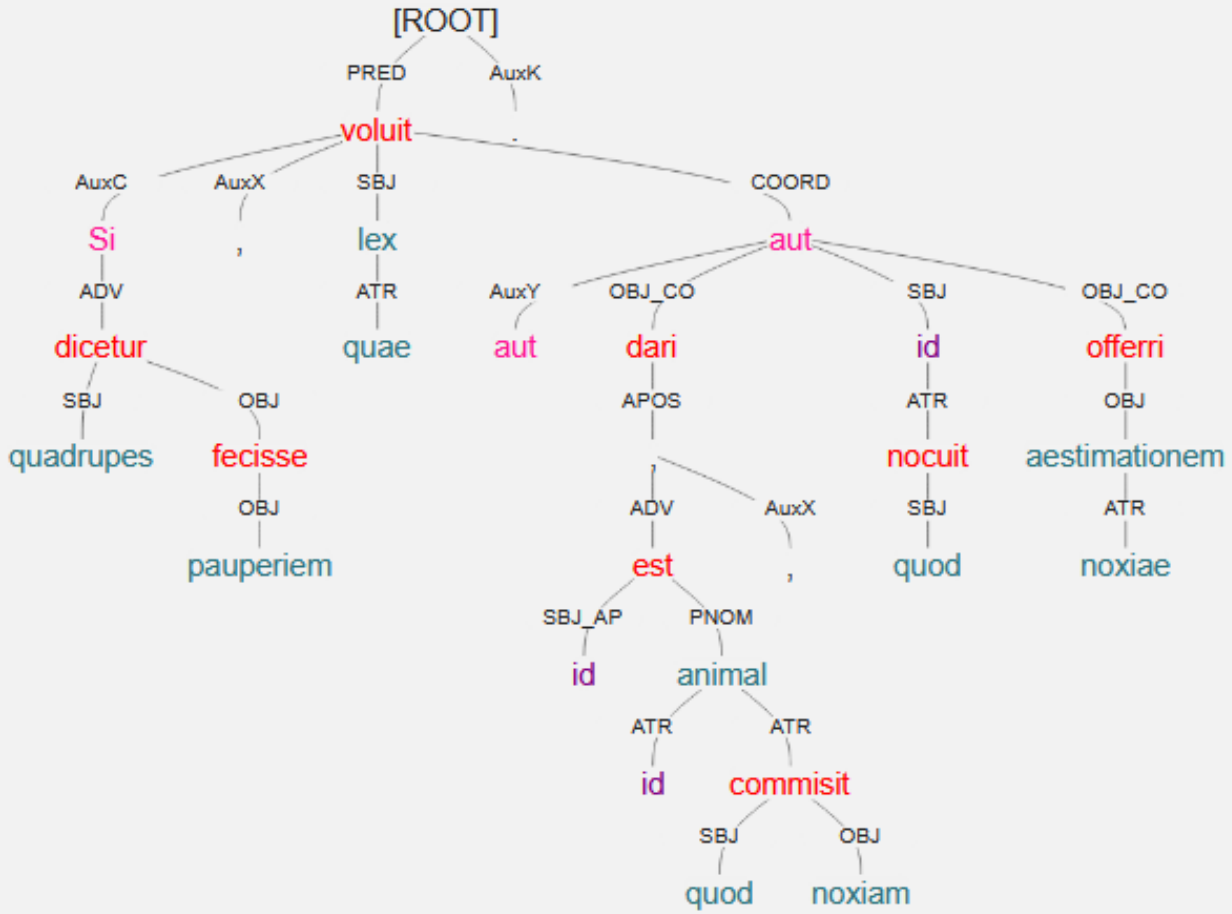
Treebanking VIII 6

Si quadrupes pauperiem fecisse dicetur, …quae lex voluit aut dari id quod nocuit, id est id animal quod noxiam commisit, aut aestimationem noxiae offerri.

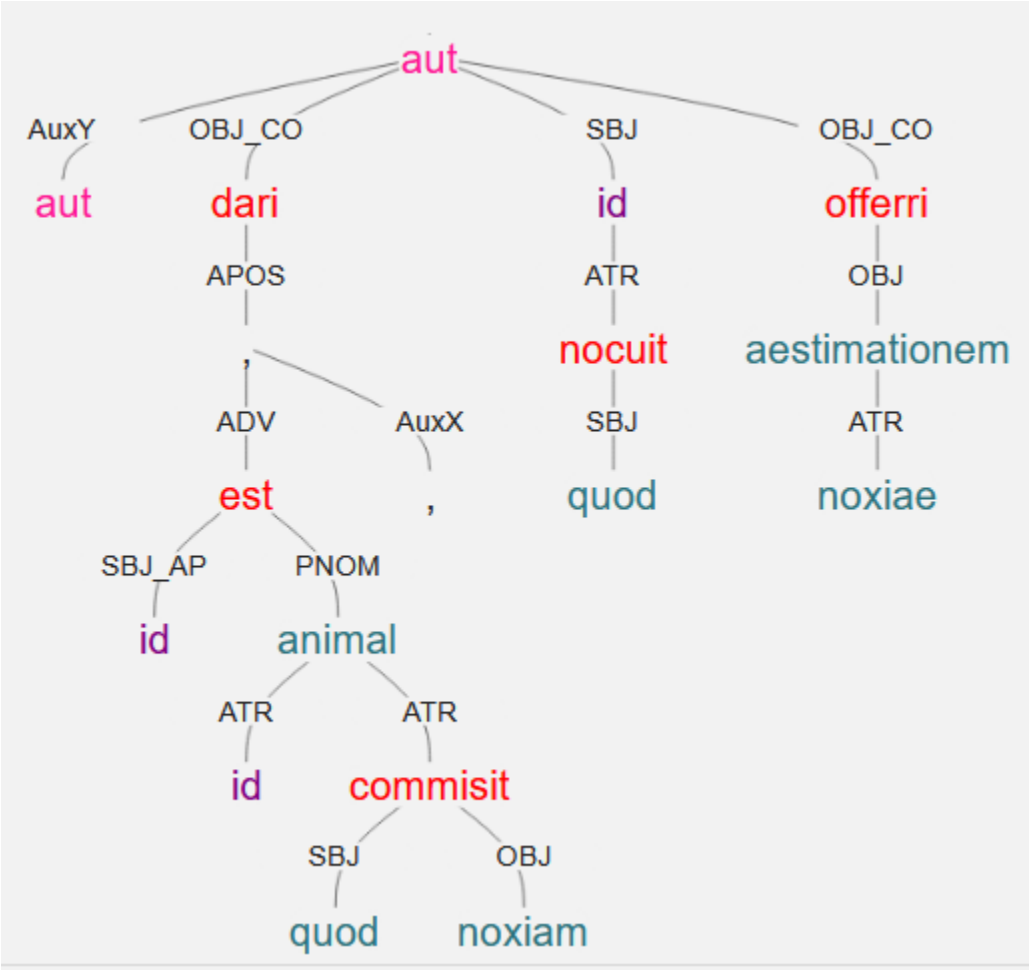
‘If a four-footed animal shall be said to have caused “pauperies”, [loss, legal action for the same is derived from the Law of the Twelve Tables], this law sanctioned either the surrender of the thing which damaged, that is the animal which committed the damage, or else the offer of assessment for the damage.’

Si quadrupes pauperiem fecisse dicetur , quae lex voluit aut dari id quod nocuit , id est id animal quod noxiam commisit , aut aestimationem noxiae offerri . VIII 6

selection none 0 unused highlight unused



The *quod* relative clauses in VIII 6



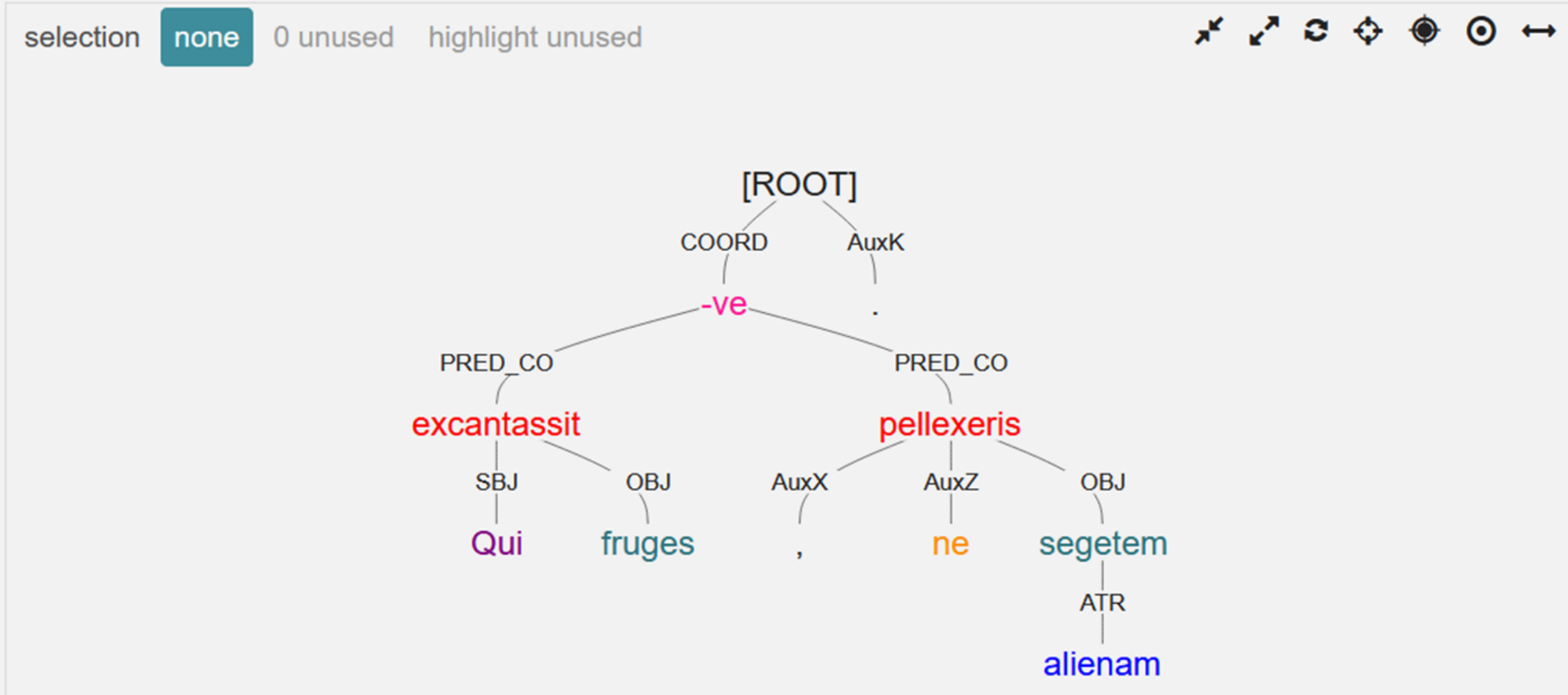
Treebanking VIII 8a-b

*Qui fruges excantassit, neve alienam segetem pellexeris**

**pellexeris* is likely meant to be *pellexerit*

‘Whoever has enchanted crops away, or enticed not another’s corn’

Qui fruges excantassit , ne -ve alienam segetem pellexeris . (VIII 8a–b)



Sources

Primary Sources:

- Cicero, *De re publica*
- Horace, *Satirae*
- Pliny, *Historia naturalis*

Secondary Sources:

- Rives, James B., “Magic in the XII Tables Revisited,” *Classical Quarterly* 52 (2002): 270–90.

RESEARCH WITH TREEBANKS: PAPYGREEK

Sunoikis Digital Classics, Fall 2021
Session 5: Research with Treebanks
Nov 4, 2021

MARJA VIERROS
UNIVERSITY OF HELSINKI

marja.vierros @ helsinki.fi

Digital Grammar of Greek Documentary Papyri (PapyGreek)
ERC Starting Grant, No. 758481

Marja Vierros, PI (morphosyntax, treebanking)

Sonja Dahlgren
postdoc
(phonology)

Erik Henriksson
(developing
infrastructure)

Polina Yordanova
PhD candidate
(word order, treebanking)

Research assistants (shorter time periods, treebanking)
Arttu Alaranta, Iida Huitula, Sari Kock, Petri Lahtinen, Jamie Vesterinen

Project website: https://www.helsinki.fi/en/researchgroups/digital-grammar-of-greek-documentary-papyri
Portal: https://papygreek.hum.helsinki.fi/ (tools and data in progress)

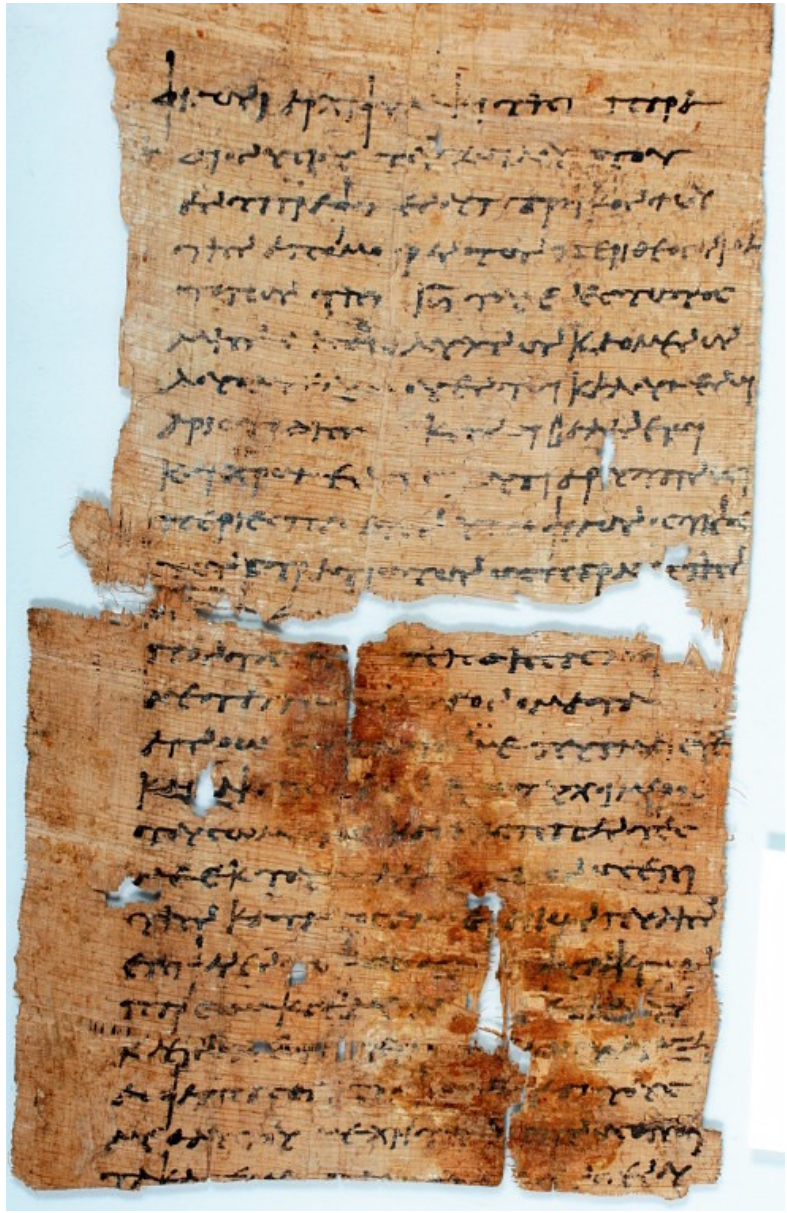
PAPY GREEK
DIGITAL GRAMMAR OF GREEK DOCUMENTARY PAPYRI

DATA RELEASE AND FORTHCOMING PUBLICATIONS

- Stable data release **PapyGreek Treebanks 1.01** (July 2021), DOI: https://doi.org/10.5281/zenodo.5074307
  - 44000 tokens (395 texts, 3102 sentences)
  - Vierros, Marja and Erik Henriksson. Forthcoming. "PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri." *Journal of Open Humanities Data* 55.
  - PapyGreek portal always has the latest versions of treebanks available for download, but stable releases should be used in in publications for replicability
- Marja Vierros and Polina Yordanova. Forthcoming. "Querying syntactic constructions in Ancient Greek parsed corpora: a case study on the genitive absolute in literature and documentary papyri." In Novokhatko, Anna, Stylianos Chronopoulos and Felix Maier (eds) *Classics@* CHS
- Henriksson, E., S. Dahlgren and M. Vierros. In preparation. "Phonological variants in the Greek documentary papyri."

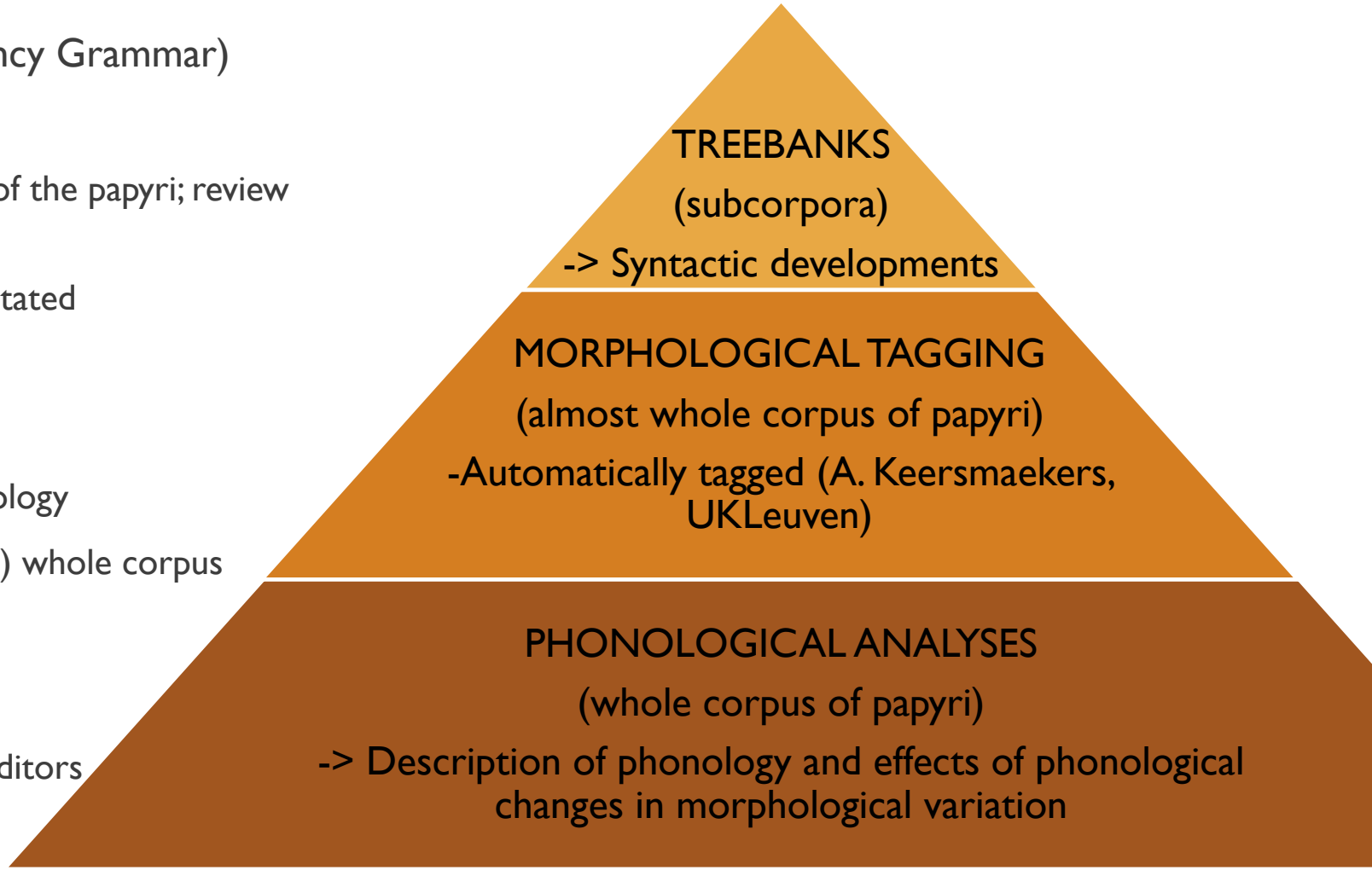
OUR SOURCE: DOCUMENTARY PAPYRI – LIMITATIONS AND ADVANTAGES

- restricted, fragmentary corpus, ca. 50 000 documents (5M tokens) with digital transcription in DDbDP (papyri.info)
- have several documentary text types (letters, contracts, administrative texts, etc.)
- direct source, wide time range c. 300BCE–600CE
- sociolinguistics (writers, authors, registers etc.)
- contact linguistics
- variation (spelling, morphology, syntax)
- language change (chronology, speed, directions)



DATA,
ITS ANNOTATION AND ANALYSIS

1. Morphosyntactic annotation (Dependency Grammar)
   - semimanually compiled
   - qualitatively analysed with close reading of the papyri; review process
   - **original** layer + **regularized** layer annotated
     - comparing these: VARIATION
2. Morphology
   - treebanks (subcorpora): checked morphology
   - automatic morphological tagging: (almost) whole corpus
3. Phonological / orthographic search tool
   - dynamic and takes context into account
   - NB: relies on the regularizations of the editors



SAMPLE RESEARCH QUESTION

KOINE -ΑΣ > -ΕΣ (NOMINAL)

- ”In nominal morphology, various trends ushering in the Medieval (and ultimately Modern) state of affairs are observable in private inscriptions and Egyptian papyri. Here belong the use of -es for -as in the accusative plural of athematic nouns.” (Bubenik 2013)
- ”The nom. pl. –ες is *occasionally* used for the acc. pl. of masculine and feminine nouns of the third declension. [19+ examples, all CE] This use […] reflects a middle stage in the process by which the nominative supplanted the accusative in the plural. […] Many of the above examples are found in connection with numerals and designations of quantity, in which the nom. –ες first came to be used for the acc. –ας in various ancient dialects.” (Gignac 1981, 46–47, my italics)
- On adjective πᾶς, πᾶσα, πᾶν: ”The acc. pl. masc. is *frequently* replaced by the –ες of the nom. as in dental stem noun of the third declension. […] Conversely, the acc. pl. –ας is used *sporadically* for the nom.” (Gignac 1981, 134, my italics)

→ -ες instead of –ας, and sporadically -ας instead of –ες

KOINE -ΑΣ > -ΕΣ (VERBAL: AORIST/PERFECT 2SG)

- interparadigmatic leveling of irregular morphology of the 'strong' aorist where Classical forms eîpon, eîp**es** 'said' are being replaced with eîpa, eîp**as** by analogy to the sigmatic forms -sa, -sas… (Bubenik 2013)
- The second person singular indicative ending -εϛ of the second aorist and imperfect is *quite often* substituted for the ending –αϛ of the first aorist throughout the Roman and Byzantine periods. […]
  - widely paralleled in inscriptions
  - not very common in Ptolemaic papyri or in Koine literature
  - may have come through the influence of the common third person ending –ε(ν)
  - Led to the retention of this ending –εϛ in the predominantly –α- inflection of the Modern Greek aorist. (Gignac 1981, 348–349)
- By the beginning of the LMedG period a common set of past-tense endings, combining elements from the aorist, the imperfect and the perfect, was already firmly in place for the active voice: -α, -εϛ, -ε(ν) –αμε(ν), -ετε (/-ατε), -αν/-ασι(ν) (Holton et al. 2019, 1613; Horrocks 2010: 318–19)

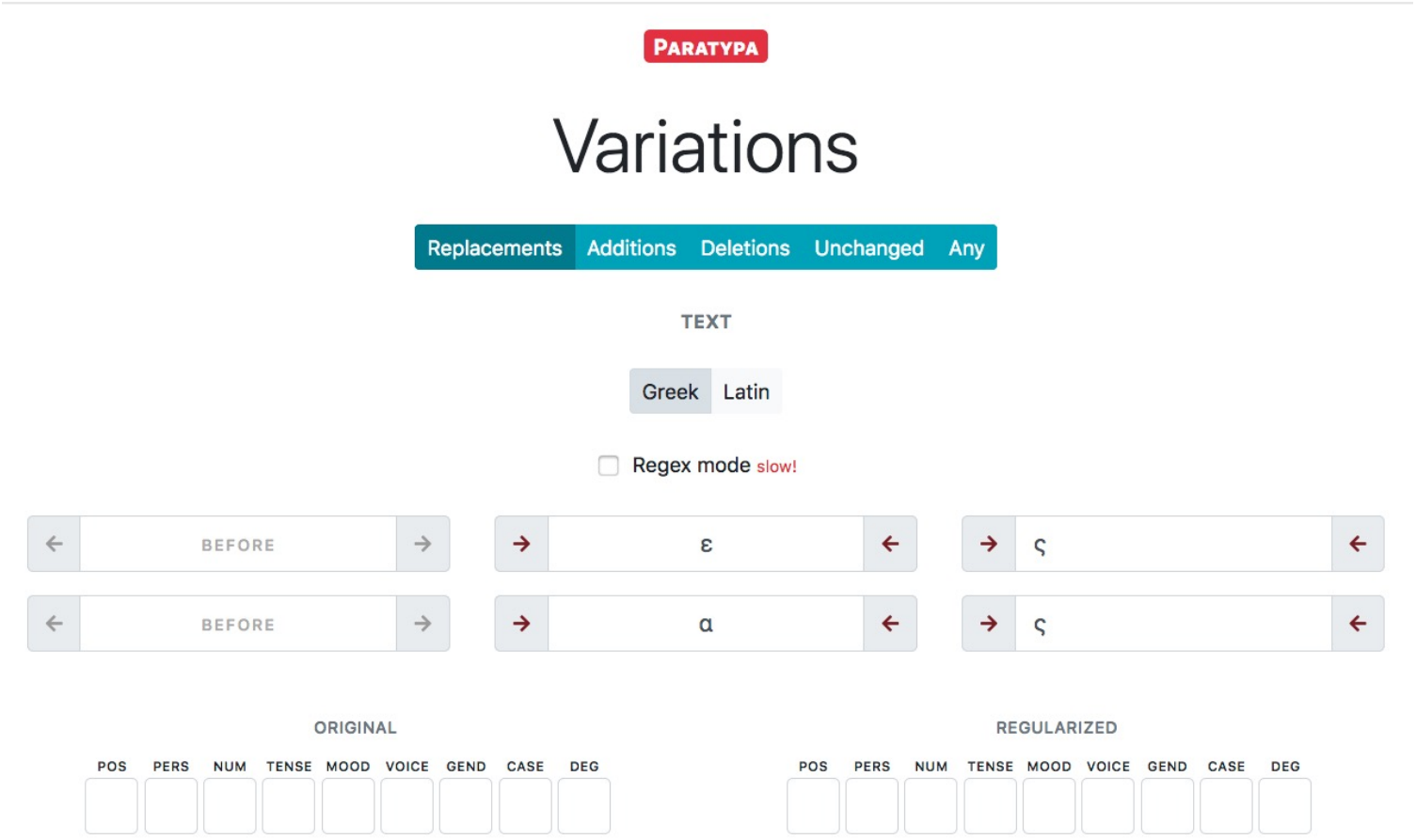
→ -εϛ instead of -αϛ
→ -αϛ instead of -εϛ

QUERYING -ΑΣ / -ΕΣ

- PapyGreek Variations search tool (Paratypa)
- searches orthographic regularisations compared to the original writing form
- morphological tagging integrated (can be combined in queries)

All orthographic regularisarions of –ες written for –ας:

original ε followed directly by ς
regularised α followed directly by ς



Showing 51 to 100 of 912 entries Filter:

Original	Pos	Rel	Regularized	Pos	Rel	Document	≤	≥		Line	Text type	People
βόες			βόας	n-p---fa-		p.abinn.60.xml	346	346	Dionysias (Arsinoites)	6		
γεγράφηκες			γεγράφηκας			p.wuerzb.21.xml	101	200	Koptos (?)	16		
γοιναῖκες			γυναῖκας	n-p---fa-		o.edfou.3.468.xml	1	100	Apollonopolis	4		
γυγγυλιδες			γογγυλίδας	n-p---fa-		o.wadi.hamm.29.xml	1	100	Wadi Hammamat	3		
γυναῖκες			γυναῖκας			chr.mitt.372.xml	142		Alexandria oder Arsinoites	2		
δέδωκες			δέδωκας	v2sria---		bgu.12.2144.xml	401	500	Hermopolis	6		
δέδωκες			δέδωκας	v2sria---		bgu.12.2191.xml	501	600	Hermopolis	1		
δέδωκες			δέδωκας			bgu.12.2195.xml	501	600	Hermopolis	2		
δέδωκες			δέδωκας			bgu.12.2196.xml	501	600	Hermopolis	1		
δέδωκες			δέδωκας			bgu.19.2769.xml	375	376	Hermopolis	4		
δέδωκες			δέδωκας	v2sria---		bgu.19.2781.xml	401	500	Hermopolis	5		
δέδωκες			δέδωκας			cpr.24.4.xml	401	450	Hermopolites	10		

The other way around:
–ας written for –ες:
68

BY MORPHOLOGY

–ες written for –ας

- Nouns: 78
- Participles: 44
- Adjectives: 66 (64 = πάντες)
- Numerals 329 (esp. τέσσαρες)

- Verb 2nd sg: 76 (aorist and perfect)

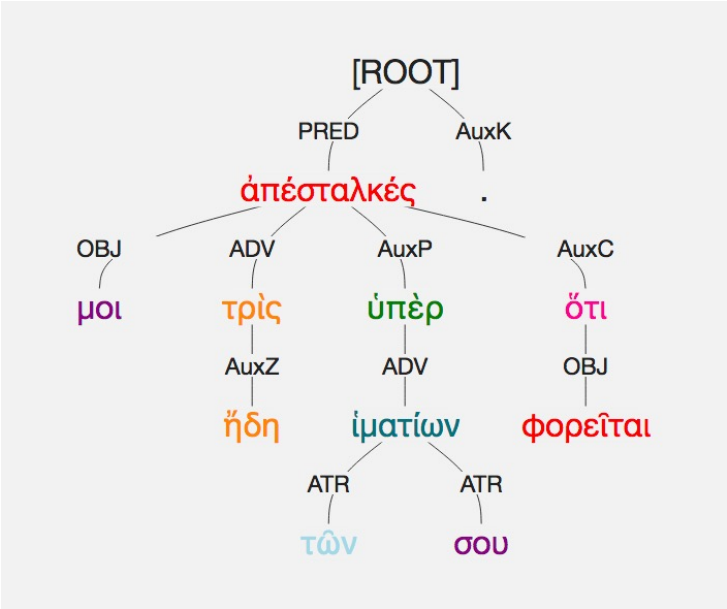
- We do find significantly more instances than what the old grammars give us

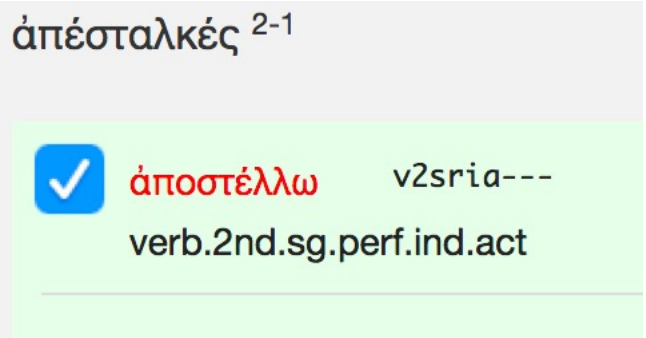
- Since we do not yet have morphological info in all documents, decisive conlusions cannot be drawn
- We could manually go through all 912 cases; would we then have found all?

Case I

- No, because all editors have not regularised these variants (as they are to be expected in the time period)

"You have already thrice sent <word> to me about your clothes, to have them brought."
O.Claud.1.160 (c. 100–120 CE)

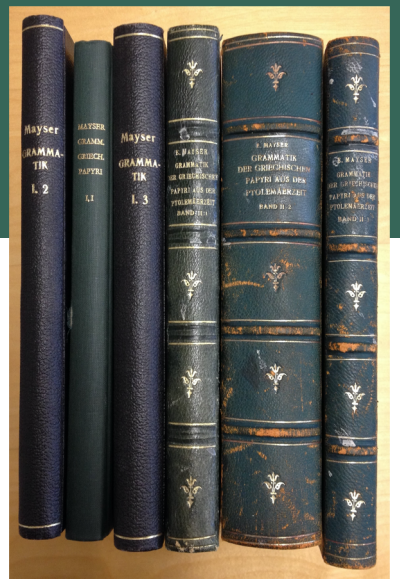


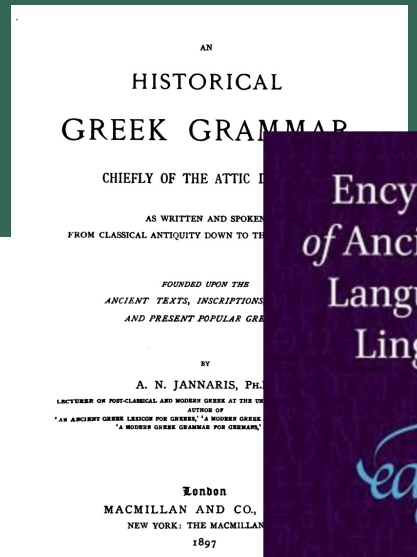


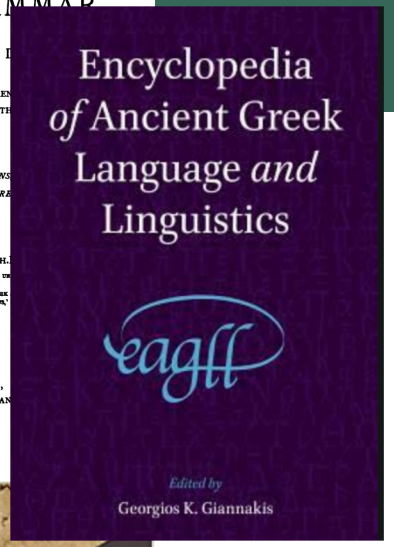
When have all treebanked, then yes, we find them all in some way…

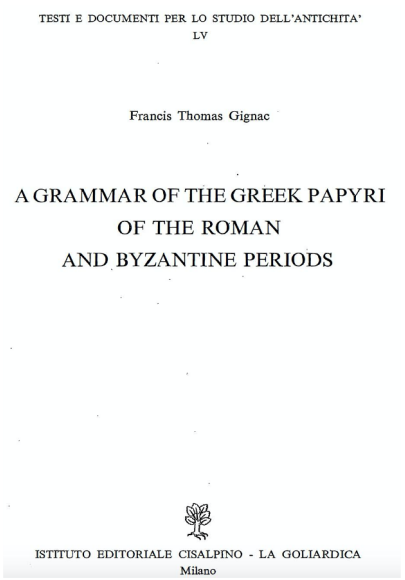
NEED FOR ANOTHER GRAMMAR?

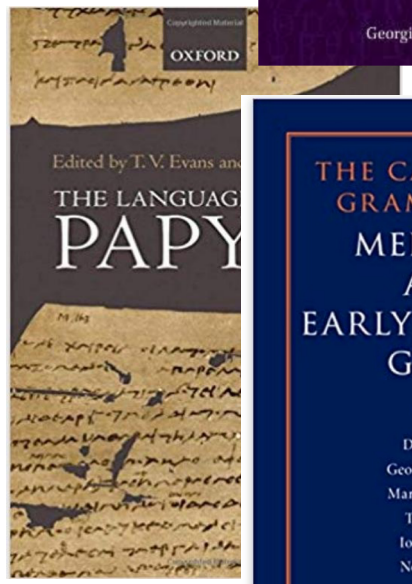
- Jannaris, Antonius N. 1968 [1897]. *An historical Greek grammar chiefly of the Attic dialect as written and spoken from Classical Antiquity down to present time founded upon the ancient texts, inscriptions, papyri and present popular Greek.*
- Edwin Mayser, *Grammatik der griechischen Papyri aus der Ptolemäerzeit. 6 vols.*
- Gignac, F.T., *A Grammar of the Greek Papyri of the Roman and Byzantine Periods. Vol. I Phonology.* (1976) *Vol. II Morphology.* (1981)
- T.V. Evans, D. D. Obbink (eds.) 2010. *The Language of the Papyri.*
- Giannakis (ed.), EAGLL, and Holton & al. 2019, CGMEMG

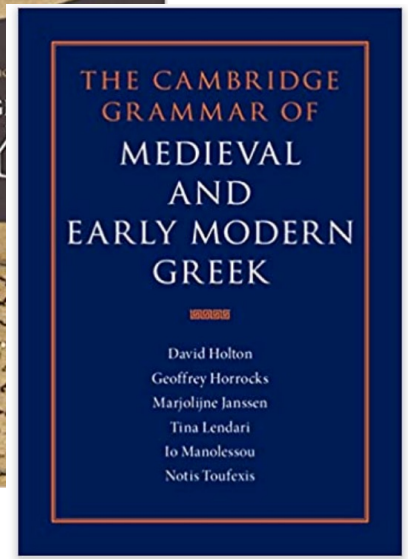












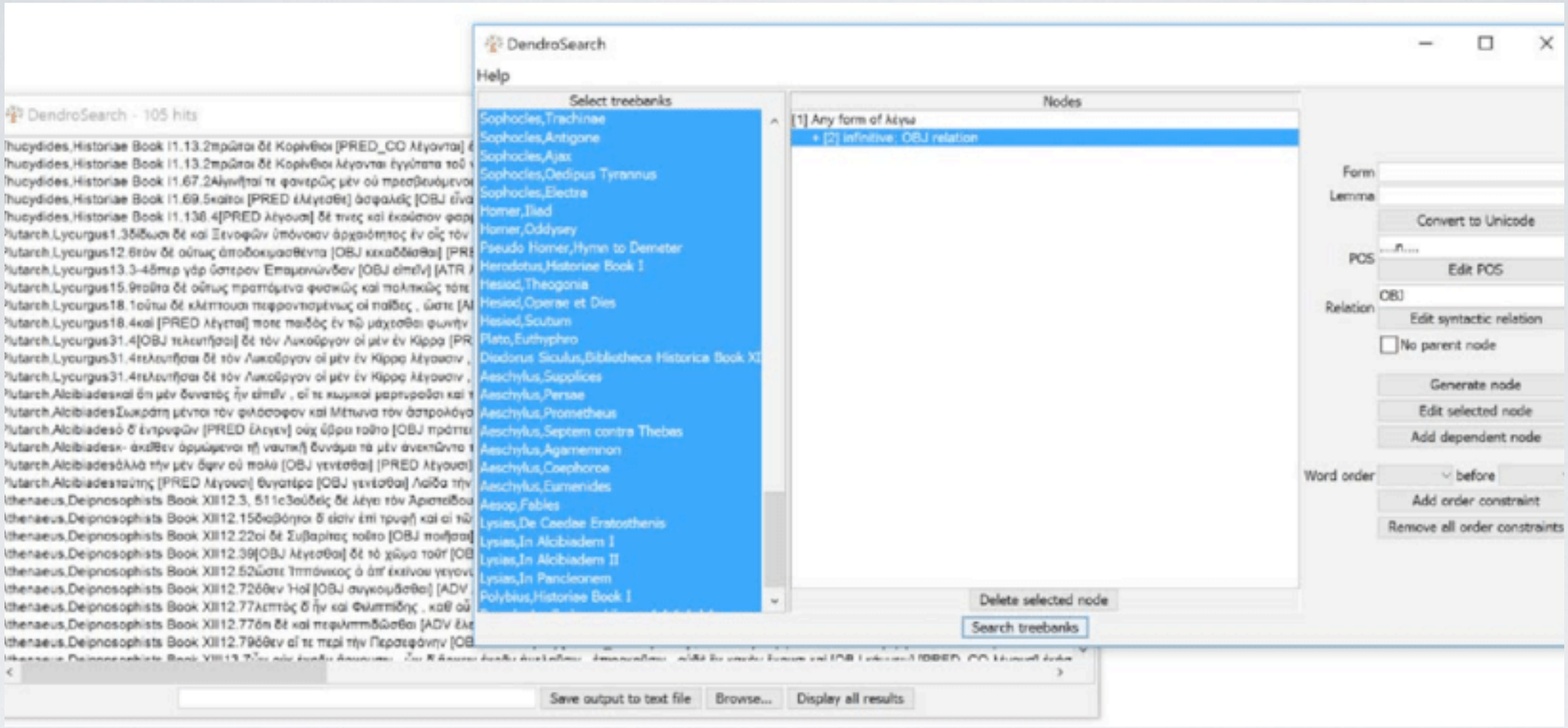
NEED FOR ANOTHER GRAMMAR?

- Previous grammars (on papyri):
  - new data has been published since
  - quantification vague (e.g. Gignac uses ”sporadically”, occasionally”, ”often”, ”frequently”)
  - some levels missing
- What we aim at:
  - new description of phonology (clarity on chronology and impact of language contact)
  - more syntax (with sprinkles of semantics and pragmatics); lexicon?
  - descriptions of (diachronic, diastratic and diatopic) variation
  - dynamic digital grammar portal, where users can query online data also according to their own interests
  - possibility for quantification
  - grammar, that also meets the needs of general linguists/typologists

REFERENCES

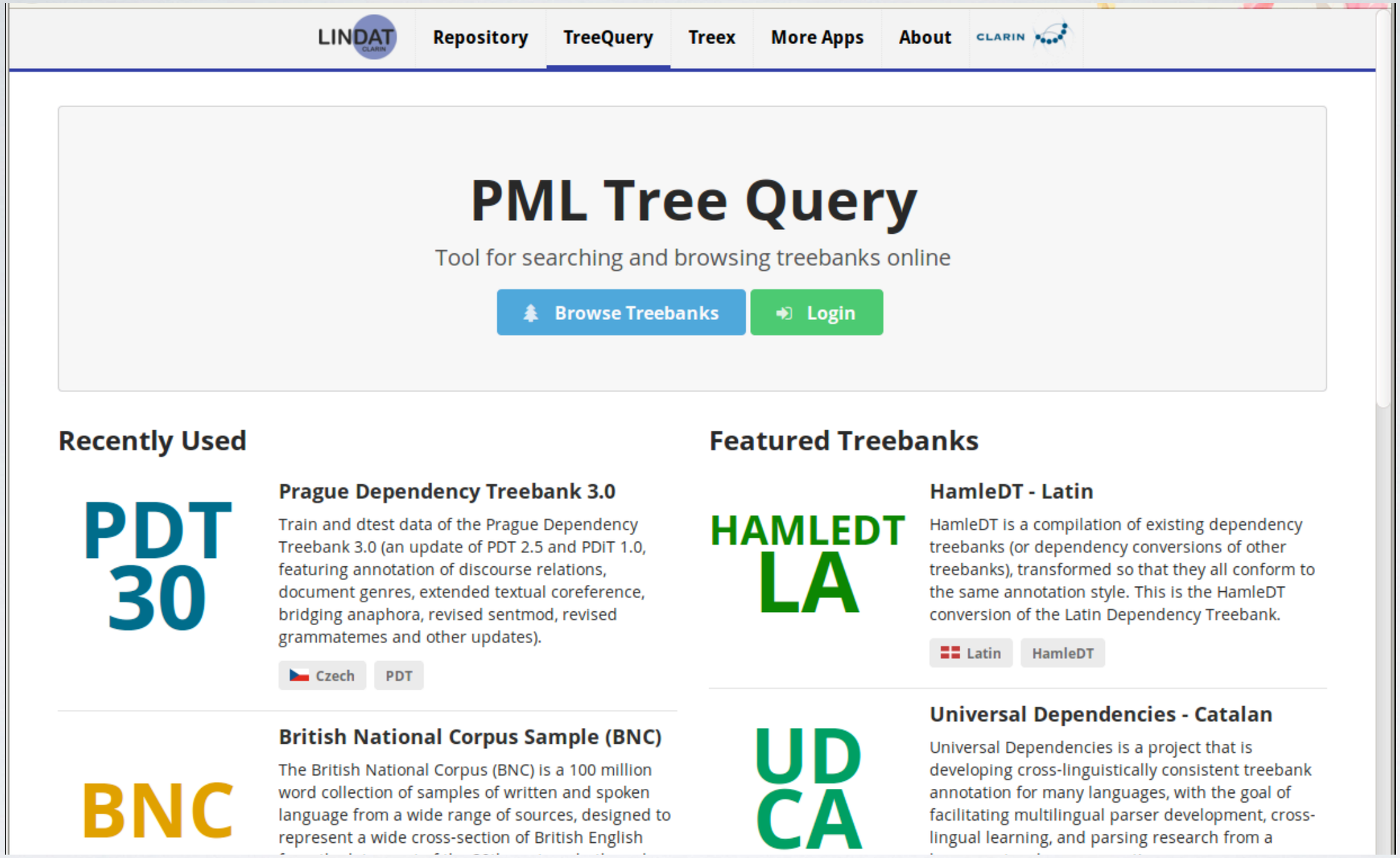
- Bubenik, Vit "Koine, Origins of", in: Encyclopedia of Ancient Greek Language and Linguistics, General Editor: Georgios K. Giannakis
- Dixon, R.M.W. 2010. *Basic Linguistic Theory Volume 1 : Methodology*. OUP Oxford.
- Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://wals.info, Accessed on 2021-02-23.)
- Holton D., G. Horrocks, M. Janssen, T. Lendari, I. Manolessou and N. Toufexis. 2019. *The Cambridge Grammar of Medieval and Early Modern Greek*.
- Manolessou, Io & Geoffrey Horrocks. 2007. The development of the definite article in Greek.
- Torallas Tovar, Sofia "Koine, Features of", in: Encyclopedia of Ancient Greek Language and Linguistics, General Editor: Georgios K. Giannakis

TOOLS



DENDROSEARCH

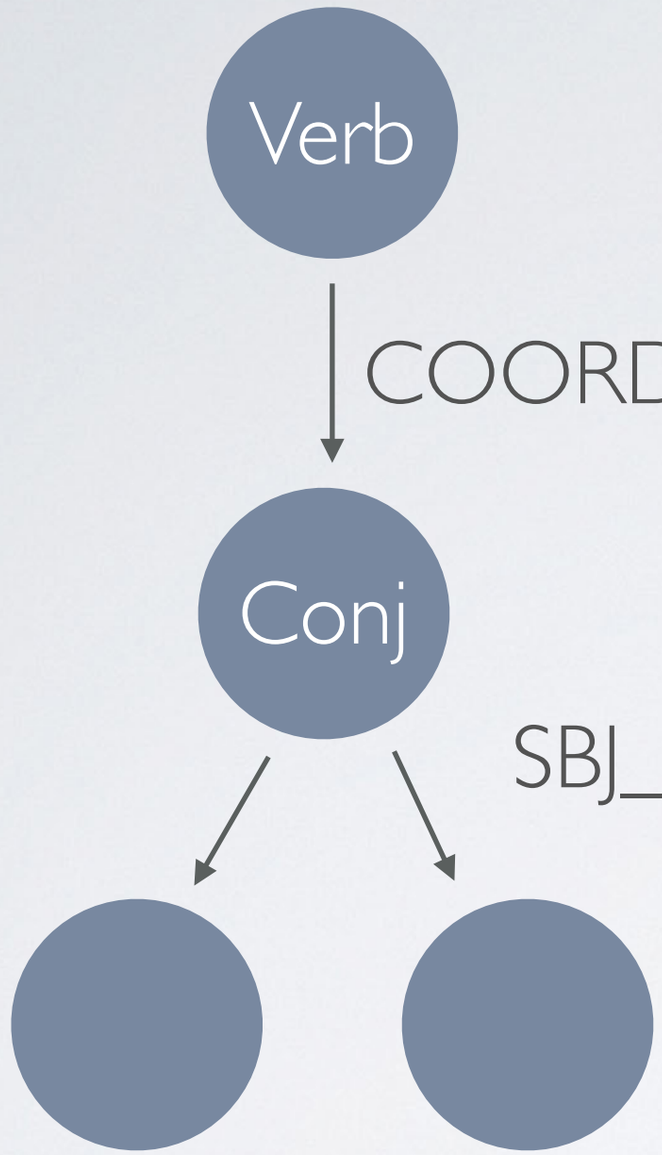
By the Pedalion project



PML-TQ

By the Pragued Dependency Treebank team

EXERCISE



What About UD?

VERB + COORDINATE SUBJECTS

Find the construction

- Use Dendrosearch to select these nodes (Head with PoS=VERB, child with relation COORD, child of COORD with relation SBJ_CO)

- Search the PROIEL treebank using DendroSearch

- Can you do the same query in UD, using PML-TQ? (select PROIEL Greek as treebank)

- Play around a bit with word-order? E.g. find sentences where the verb precedes/follows the subjects

HINTS

- Read the guidelines!
- Here is what you need for UD
- Here is the link to the Perseus-style guidelines: Latin and Greek