SunoikisisDC Summer 2024

Session 9

Treebanking and annotation to study reception of ancient language and text

Marja Vierros (U Helsinki)

Alek Keersmaekers (KU Leuven)

Gabriel Bodard (U London)

Julia Balbilla poem 1

Ἰουλίας Βαλ<β>ίλλης·
ὄτε ἤκουσε τοῦ Μέμνο<νο>ς
ὀ Σεβαστὸς Ἀδριανός.

Μέμνονα πυνθανόμαν Αἰγύπτιον, ἀλίω αὔγαι
αἰθόμενον, φώνην Θηβαΐ<κ>ω ’πυ λίθω.
Ἀδρίανον δ’ ἐσίδων, τὸν παμβασίληα, πρὶν αὔγας
ἀελίω χαίρην εἶπέ <ϝ>οι ὠς δύνατον.
Τίταν δ’ ὄττ’ ἐλάων λεύκοισι δι’ αἴθερος ἴπποις
ἐ̣ν̣ὶ σκίαι ὠράων δεύτερον ἦχε μέτρον,
ὠς χάλκοιο τύπεντ[ο]ς ἴη Μέμνων πάλιν αὔδαν
ὀξύτονον· χαίρω[ν κ]αὶ τρίτον ἆχον ἴη.
κοίρανος Ἀδρίανο[ς τότ’ ἄ]λις δ’ ἀσπάσσατο καὖτος
Μέμνονα, κἀν [στά]λαι κάλ̣λ̣ι[π]εν ὀψ[ι]γόνοις
γρόππατα σαμαίνο[ν]τά τ’ ὄσ’ εὔϊδε κὤσσ’ ἐσάκουσε,
δῆλον παῖσι δ’ ἔγε[ν]τ’ ὤς <ϝ>ε φίλισι θέοι.

Julia Balbilla poem 2

ὄτε σὺν τῇ Σεβαστῇ Σαβείνη-
ι ἐγενόμην παρὰ τῷ Μέμνονι.
Αὔως καὶ γεράρω, Μέμνον, πάϊ Τιθώνοιο,
Θηβάας θάσσων ἄντα Δίος πόλιος,
ἢ Ἀμένωθ, βασίλευ Αἰγύπτιε, τὼς ἐνέποισιν
ἴρηες μύθων τῶν παλάων ἴδριες,
χαῖρε, καὶ αὐδάσαις πρόφρων ἀσπάσ̣δ̣ε[ο κ]αὔτ[αν]
τὰν σέμναν ἄλοχον κοιράνω Ἀδριάνω.
γλῶσσαν μέν τοι τμᾶξε [κ]αὶ ὤατα βάρβαρο̣ς ἄνηρ,
Καμβύσαις ἄθεος· τῶ ῥα λύγρῳ θαν̣άτῳ
δῶκέν τοι ποίναν τὤτωι ἄκ[ρῳ] ἄορι πλάγεις
τῷ νήλας Ἆπιν κάκτανε τὸν θέϊον.
ἀλλ’ ἔγω οὐ δοκίμωμι σέθεν τόδ’ ὄ̣λ̣εσθ’ ἆ̣ν ἄγαλμα,
ψύχαν δ’ ἀθανάταν λοίπ̣ο̣ν̣ ἔσωσ̣α̣ νόῳ.
εὐσέβεες γὰρ ἔμοι γένεται πάπποι τ’ ἐγένο̣ντο,
Βάλβιλλός τ’ ὀ σόφος κ’ Ἀντίοχος βασίλευς,
Βάλβιλλος γενέταις μᾶτρος βασιλήϊδος ἄμμας,
τῶ πάτ̣ερος δὲ πάτηρ Ἀντίοχος βασίλευς·
κήνων ἐκ γενέας κἄγω λόχον αἶμα τὸ κᾶλον,
Βαλβίλλας δ’ ἔμεθεν γρόπτα τάδ’ εὐσέβεος.

Julia Balbilla poem 3

ὅτε τῇ πρώτῃ ἡμέρᾳ οὐκ ἀκούσαμεν τοῦ Μέμνονος.

χθίσδον μέν Μέμνων σίγαις ἀπε[δέξατ’ ἀκ]οίτα[ν](?),
ὡς πάλιν ἀ κάλα τυῖδε Σάβιννα μ̣ό[λοι].
τ̣έρπει γάρ σ’ ἐράτα μόρφα βασιλήϊδος ἄμ̣μας,
ἐλθοίσαι δ’ [α]ὔ̣ται θήϊον ἄ̣χον ἴη,
μὴ καί τοι βασίλευς κοτ̣έσῃ, τό νυ δᾶρον ἀτά[ρβης]
τὰν σέμναν κατ̣έχ̣ες κουριδίαν ἄλοχον.
κὠ Μέμνων τρέσσαις μεγάλω μένος Ἀδρι[άνοιο]
ἐξαπίνας αὔδασ’, ἀ δ’ ὀΐοισ’ ἐχάρη.

Julia Balbilla poem 4

ἔκλυον αὐδήσαντος ἔγω ’πυ λίθω Βάλβιλλα
    φώνα<ς> τᾶς θεΐας Μέμνονος ἢ Φαμένωθ.
ἦλθον ὔμοι δ’ ἐράται βασιλήϊδι τυῗδε Σαβίννᾳ,
    ὤρας δὲ πρώτας ἄλιος ἦχε δρόμος.
κοιράνω{ι} Ἀδριάνω πέμπτῳ δεκότῳ δ’ ἐ/νιαύτῳ,
    <φῶτ>α δ’ ἔχεσκε<ν> Ἄθυρ εἴκοσι / καὶ πέσυρα.

- Encoding texts in EpiDoc
- Dialect analysis, text emendation
- Epigraphic + archaeological metadata
- Encoding/analysing metre
- Prosopography/genealogy, religion
- Image annotation
- Translation alignment
- Morphosyntactic annotation / syntax analysis
- Reception of archaic poetry in imperial literature

Treebanking & comparing syntax Case study: Julia Balbilla

Marja Vierros, University of Helsinki

SunoikisisDC Digital Classics and Byzantine Studies: Session 9
Monday June 10, 2024

Outline today

- Very short intro to treebanking *aka* linguistic annotation
- Very short intro to PapyGreek platform for annotating papyri (and inscriptions)
- Demonstration of treebanking Balbilla’s poem 1
    - For a full tutorial of treebanking, see the Sunoikisis Session 9 page (exercise), or
    - Other tutorial listed in https://wiki.digitalclassicist.org/Treebanking
    - Remember to read the guidelines, too!
- Comparing syntactic structures of Balbilla to Sappho (only preliminary, non-machine-facilitated glimpse)

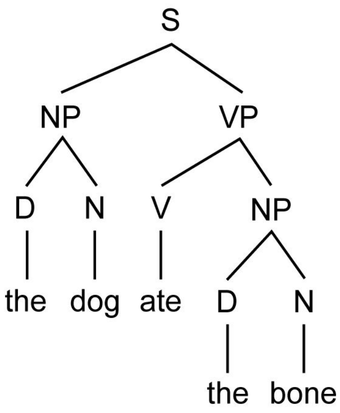
What is linguistic annotation?

- many different types of linguistic information can be added to a (digital) text
    - WORD LEVEL
        - lemma annotation (lemma = basic form of a word, the ’dictionary entry’)
        - morphological annotation (inflectional morphology like case, number, gender etc.)
        - POS annotation (word classes i.e. parts-of-speech)
    - SYNTACTIC ANNOTATION
        - different grammar formalisms: most important here: *dependency* and *constituent* structures
    - SEMANTIC ANNOTATION
        - Semantic classes based on meanings of words (or phrases or sentences), e.g. concrete / abstract nouns; animate / inanimate etc.
        - Annotating concepts (e.g. people, places, organizations, products or topics)

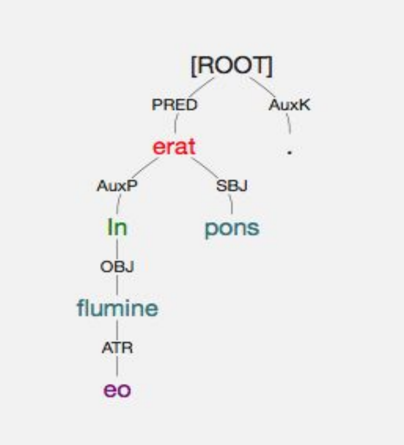
What is a treebank?

- tree structure (tree diagram) is a graphical way of representing a **hierarchical** structure, e.g. the syntax of a sentence

- Treebank = parsed text corpus that annotates syntactic (or semantic) sentence structure
  - Usually includes lemmatization and morphological annotation



Constituency tree



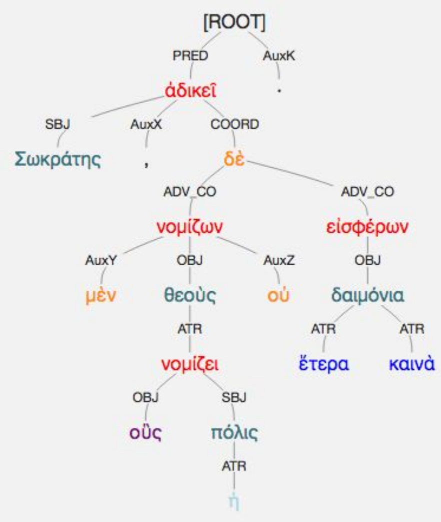
Dependency tree

Treebanks of Ancient Greek and Latin (=Dependency Trees)

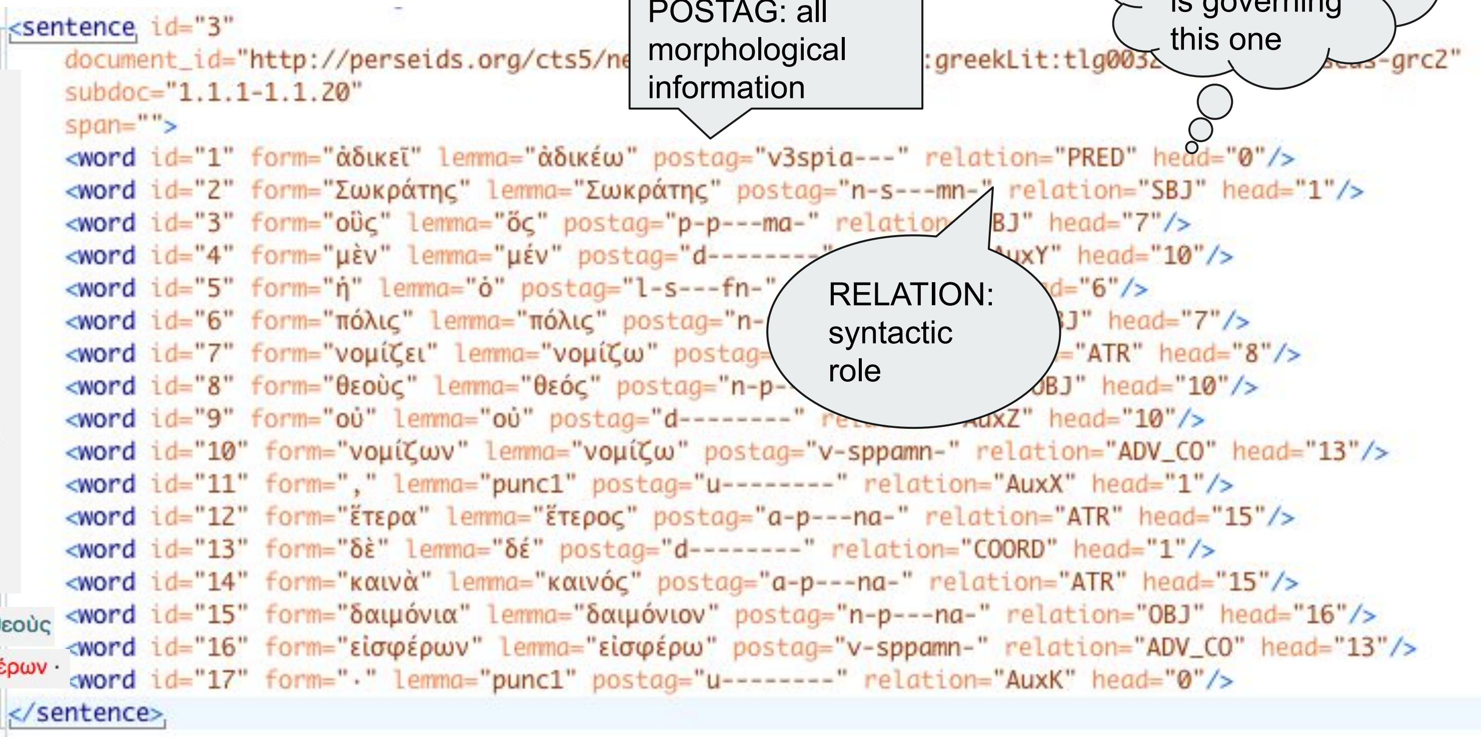
- The “Index Thomisticus Treebank” (inspired by the Prague Dependency Treebank)
- The “Perseus family” (AGLDT, Pedalion, Gorman Trees, PapyGreek …)
- PROIEL = Pragmatic Resources in Old Indo-European Languages

- **Universal Dependencies** (UD)

Underlying XML of a dependency tree (AGDT)



ἀδικεῖ Σωκράτης οὓς μὲν ἡ πόλις νομίζει θεοὺς
οὐ νομίζων , ἕτερα δὲ καινὰ δαιμόνια εἰσφέρων ·



```
<sentence id="3"
    document_id="http://perseids.org/cts5/ne                    :greekLit:tlg0032                    -grc2"
    subdoc="1.1.1-1.1.20"
    span="">
    <word id="1" form="ἀδικεῖ" lemma="ἀδικέω" postag="v3spia---" relation="PRED" head="0"/>
    <word id="2" form="Σωκράτης" lemma="Σωκράτης" postag="n-s---mn-" relation="SBJ" head="1"/>
    <word id="3" form="οὓς" lemma="ὅς" postag="p-p---ma-" relation    BJ" head="7"/>
    <word id="4" form="μὲν" lemma="μέν" postag="d--------"            uxY" head="10"/>
    <word id="5" form="ἡ" lemma="ὁ" postag="l-s---fn-"                 d="6"/>
    <word id="6" form="πόλις" lemma="πόλις" postag="n-                 J" head="7"/>
    <word id="7" form="νομίζει" lemma="νομίζω" postag                  ="ATR" head="8"/>
    <word id="8" form="θεοὺς" lemma="θεός" postag="n-p-                OBJ" head="10"/>
    <word id="9" form="οὐ" lemma="οὐ" postag="d--------"   re          uxZ" head="10"/>
    <word id="10" form="νομίζων" lemma="νομίζω" postag="v-sppamn-" relation="ADV_CO" head="13"/>
    <word id="11" form="," lemma="punc1" postag="u--------" relation="AuxX" head="1"/>
    <word id="12" form="ἕτερα" lemma="ἕτερος" postag="a-p---na-" relation="ATR" head="15"/>
    <word id="13" form="δὲ" lemma="δέ" postag="d--------" relation="COORD" head="1"/>
    <word id="14" form="καινὰ" lemma="καινός" postag="a-p---na-" relation="ATR" head="15"/>
    <word id="15" form="δαιμόνια" lemma="δαιμόνιον" postag="n-p---na-" relation="OBJ" head="16"/>
    <word id="16" form="εἰσφέρων" lemma="εἰσφέρω" postag="v-sppamn-" relation="ADV_CO" head="13"/>
    <word id="17" form="·" lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
</sentence>
```

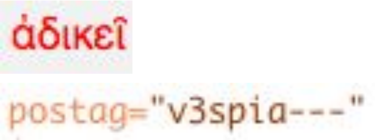
POSTAG: all morphological information

HEAD: which word is governing this one

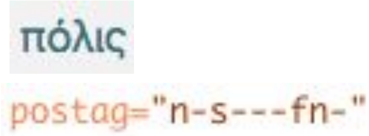
RELATION: syntactic role

postag

- nine place string including the morphological information


ἀδικεῖ

postag="v3spia---"


πόλις

postag="n-s---fn-"

1: verb	1: part-of-speech	1: noun
2: 3rd person	2: person	2: -
3: singular	3: number	3: singular
4: present	4: tense	4: -
5: indicative	5: mood	5: -
6: active	6: voice	6: -
7: -	7: gender	7: feminine
8: -	8: case	8: nominative
9: -	9: degree	9: -

Find the lettercodes here: https://github.com/gcelano/LemmatizedAncientGreekXML

European Research Council
Established by the European Commission

https://papygreek.com/

PAPY GREEK
DIGITAL GRAMMAR OF GREEK DOCUMENTARY PAPYRI

PapyGreek

A platform for the linguistic study of Greek papyri, including a grammar, annotated texts, and a search tool.

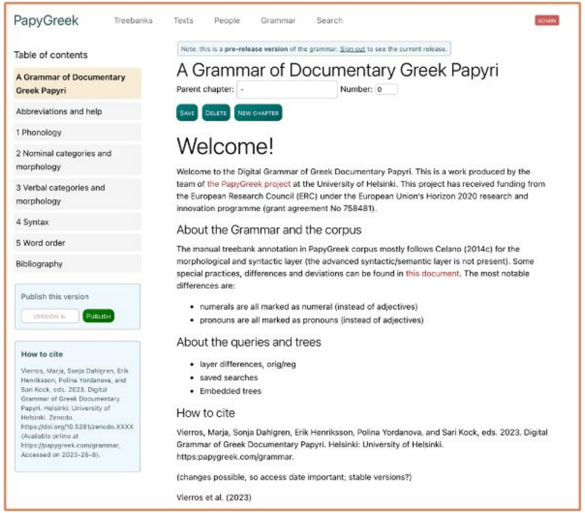


Table of contents

- A Grammar of Documentary Greek Papyri
- Abbreviations and help
- 1 Phonology
- 2 Nominal categories and morphology
- 3 Verbal categories and morphology
- 4 Syntax
- 5 Word order
- Bibliography

Publish this version

VERSION N. PUBLISH

How to cite

Vierros, Marja, Sonja Dahlgren, Erik Henriksson, Polina Yordanova, and Sari Kock, eds. 2023. Digital Grammar of Greek Documentary Papyri. Helsinki: University of Helsinki. Zenodo. https://doi.org/10.5281/zenodo.XXXX (Available online at https://papygreek.com/grammar, Accessed on 2023-26-8).

Note: this is a **pre-release version** of the grammar. Sign out to see the current release.

A Grammar of Documentary Greek Papyri

Parent chapter: - Number: 0

SAVE DELETE NEW CHAPTER

Welcome!

Welcome to the Digital Grammar of Greek Documentary Papyri. This is a work produced by the team of the PapyGreek project at the University of Helsinki. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758481).

About the Grammar and the corpus

The manual treebank annotation in PapyGreek corpus mostly follows Celano (2014c) for the morphological and syntactic layer (the advanced syntactic/semantic layer is not present). Some special practices, differences and deviations can be found in this document. The most notable differences are:

- numerals are all marked as numeral (instead of adjectives)
- pronouns are all marked as pronouns (instead of adjectives)

About the queries and trees

- layer differences, orig/reg
- saved searches
- Embedded trees

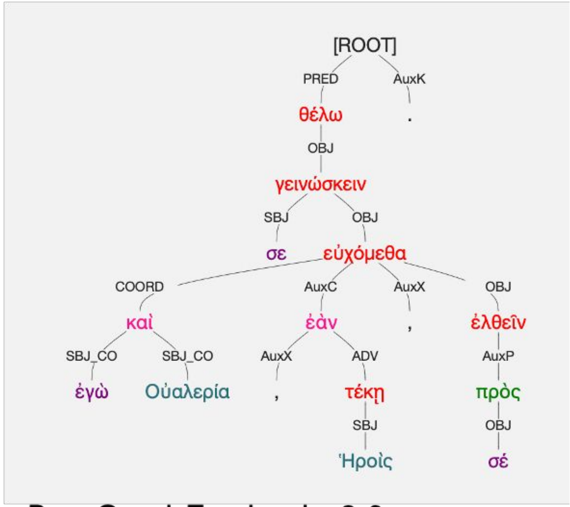
How to cite

Vierros, Marja, Sonja Dahlgren, Erik Henriksson, Polina Yordanova, and Sari Kock, eds. 2023. Digital Grammar of Greek Documentary Papyri. Helsinki: University of Helsinki. https:papygreek.com/grammar.

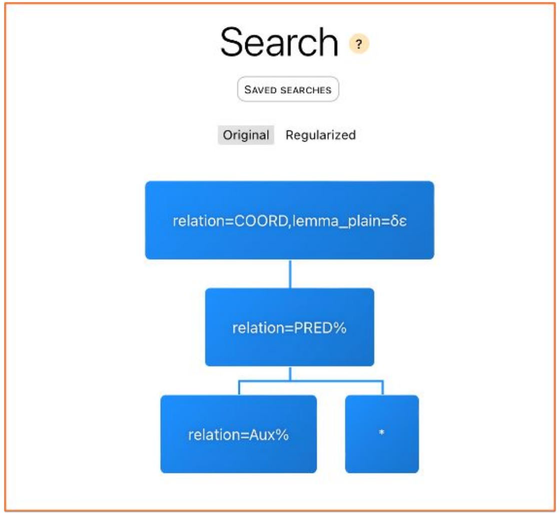
(changes possible, so access date important; stable versions?)

Vierros et al. (2023)

Grammar not public yet



PapyGreek Treebanks 3.0:
https://zenodo.org/records/8428823



relation=COORD,lemma_plain=δε

relation=PRED%

relation=Aux%

*

Whole corpus of papyri for phonology and morphology; treebanked corpus for syntax

PapyGreek platform for treebanking papyri (and inscriptions)

- Greek Documentary Papyri all there (synchronized with Papyri.info)
- TEI EpiDoc XML preprocessed to get annotatable texts
  - Papyri and inscriptions are fragmentary, contain markup, and cannot be annotated as is
- –> works also for inscriptions (although some differences in markup)
- Arethusa annotation tool implemented within PapyGreek
  - Arethusa originally exists in Perseids' page (Tufts University): https://sosol.perseids.org/sosol/

Demo on Julia Balbilla's inscribed poems to follow

Julia Balbilla’s poem 1 (beginning)

Ἰουλίας Βαλ<β>ίλλης·
ὅτε ἤκουσε τοῦ Μέμνο<νο>ς
ὁ Σεβαστὸς Ἁδριανός.

Μέμνονα πυνθανόμαν Αἰγύπτιον, ἀλίω αὔγαι
αἰθόμενον, φώνην Θηβαΐ<κ>ω ʼπυ λίθω.
Ἀδρίανον δʼ ἐσίδων, τὸν παμβασίληα, πρὶν αὔγας
ἀελίω χαίρην εἶπέ <ϝ>οι ὡς δύνατον.

[Composed] by Julia Balbilla,
When the august Hadrian
heard Memnon.

I’ve heard tell that Memnon the Egyptian, warmed by the sun’s rays,
utters a loud sound from the Theban stone.
Seeing Hadrian, the greatest of kings, and before greeting
the sun’s rays, he (Memnon) addressed him as well as he could.

Transl. Patricia A. Rosenmeyer

Introducing GLAUx

Alek Keersmaekers

University of Leuven

The GLAUx corpus

- The **G**reek **L**anguage **Au**tomated
- 26M tokens, literary Ancient Greek (with the papyri: 32M)
- Ranging from Homer to the Bible to mathematical texts to narrative texts to …
- 8th century BC – 3rd/4th century AD
- Automatic annotation:
  - Lemmas
  - Morphology
  - Syntax
  - Future: Semantics
- https://glaux.be/, https://github.com/alekkeersmaekers/glaux

GLAUx annotation

- ‘Trained’ on manually annotated treebanks: AGDT, PROIEL, Gorman, Pedalion, Harrington, Yordanova (1,5M words) → about 5% of GLAUx
- Machine learning (AI) to predict annotations automatically
- Lemmas: ~99% accurate
- Morphology: ~97% accurate
- Syntax: ~80% accurate