Sunoikisis Digital Classics Spring 2019

Session 5, February 7, 2019

Introduction to Treebanking

Marja Vierros & Polina Yordanova
(University of Helsinki)

structure of the session

1. Introduction to Dependency Grammar and treebanking Latin and Greek (MV)

2. Treebanking in action - demonstration of the annotation tool Arethusa (PY)

NB. Sunoikisis Digital Classics Spring 2019, February 21

Session 7. Using treebanked corpora: Universal Dependencies

(Marco Passarotti & Timo Korkiakangas)

tree structure

- tree structure (tree diagram) is a graphical way of representing a hierarchical structure
- same hierarchical structure can be presented in multiple different ways

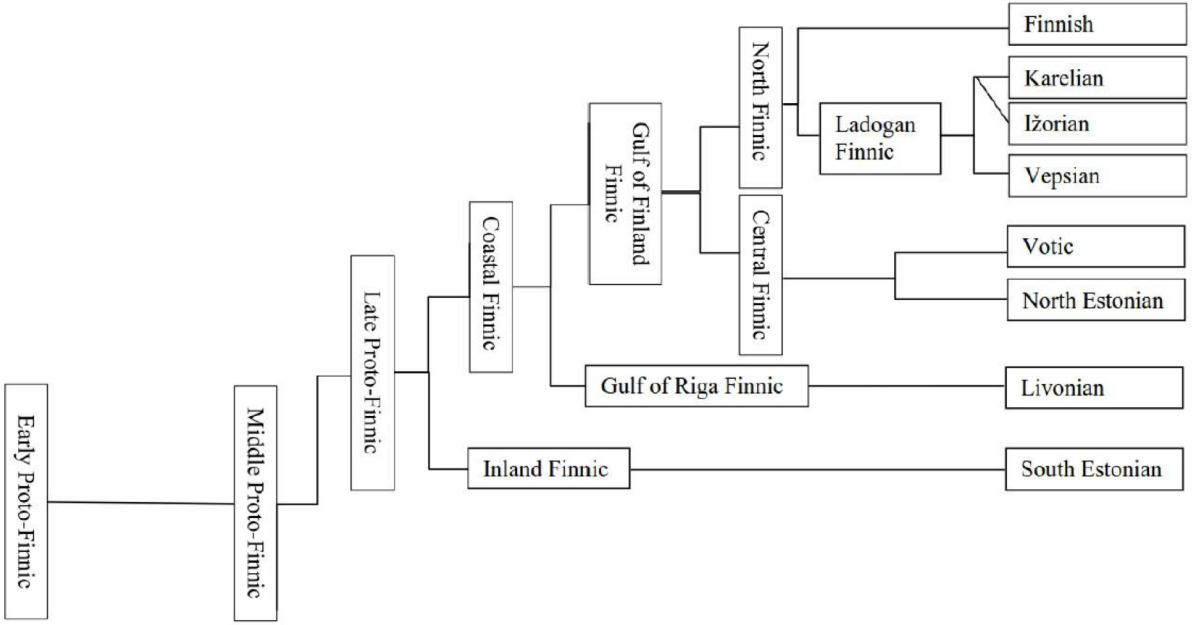
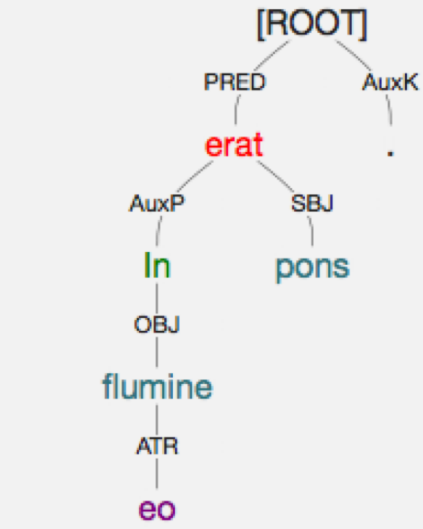


Figure 1. A Finnic language family tree (source Kallio 2014: 163).



In eo flumine pons erat.

Kallio, Petri. 2014. "The Diversification of Proto-Finnic". In Joonas Ahola & Frog with Clive Tolley (eds.). 2014. *Fibula, Fabula, Fact – The Viking Age in Finland.* Studia Fennica Historica 18. Helsinki: Finnish Literature Society, 155–168.
Tree edited in Frog & Saarikivi, Janne. 2015. "De situ linguarum fennicarum aetatis ferreae, Pars I" In Frog, Helen F. Leslie-Jacobsen and Joseph S. Hopkins (eds.), *Retrospective Methods Network Newsletter* 9: 64–115.

what is a treebank?

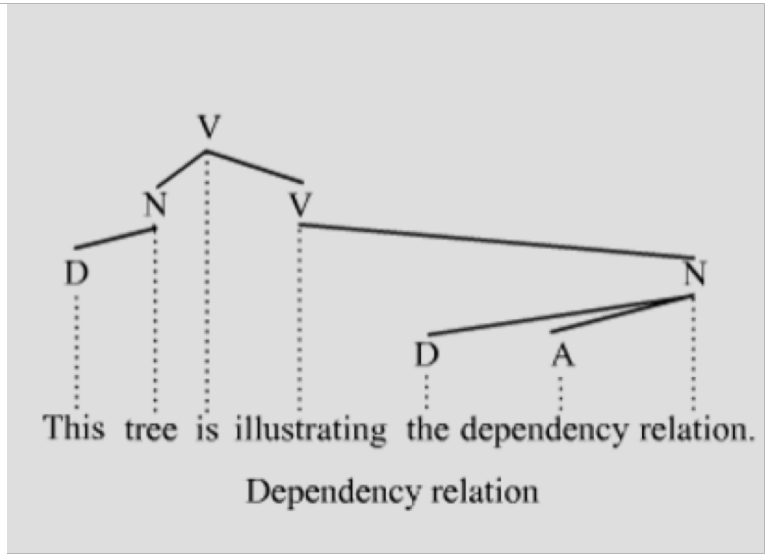
- = parsed text corpus that annotates syntactic or semantic sentence structure
- Treebanking = activity of building linguistic trees = linguistic annotation = adding linguistic information to a text corpus
- many different types of linguistic information can be annotated
  - WORD LEVEL
    - lemma annotation (lemma = basic form of a word, the 'dictionary entry')
    - morphological annotation (inflectional morphology like case, number, gender etc. OR derivational morphology e.g. parts of compound words)
    - POS annotation (word classes i.e. parts-of-speech)
  - SYNTACTIC ANNOTATION
    - different grammar formalisms: most important here: **dependency** and **constituent** structures
  - SEMANTIC ANNOTATION
    - meanings of words
    - meanings of phrases and sentences

See, e.g. S. Kübler & H. Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora.* Bloomsbury: London and New York.

dependency vs. constituency structures

- The dependency grammar presumes direct links between linguistic units (words). The (finite) verb is taken to be the center of the clause structure on which all other syntactic units *depend* either directly or indirectly.
  - each word has a *head*, on which it depends, except the main verb (or several coordinated verbs), considered as the *root*

- constituency grammar, also called as "phrase structure grammar", presumes that words are grouped into phrases (noun phrases, verbal phrases, adverbial phrases). Phrases are hierarchically grouped into larger phrases and finally clauses.
  - for each phrase, one word serves as a head, which determines the syntactic category of the phrase



dependency vs. constituency structures

- The dependency grammar presumes direct links between linguistic units (words). The (finite) verb is taken to be the center of the clause structure on which all other syntactic units *depend* either directly or indirectly.
  - each word has a *head*, on which it depends, except the main verb (or several coordinated verbs), considered as the *root*

- constituency grammar, also called as "phrase structure grammar", presumes that words are grouped into phrases (noun phrases, verbal phrases, adverbial phrases), Phrases are hierarchically grouped into larger phrases and finally clauses.
  - for each phrase, one word serves as a head, which determines the syntactic category of the phrase

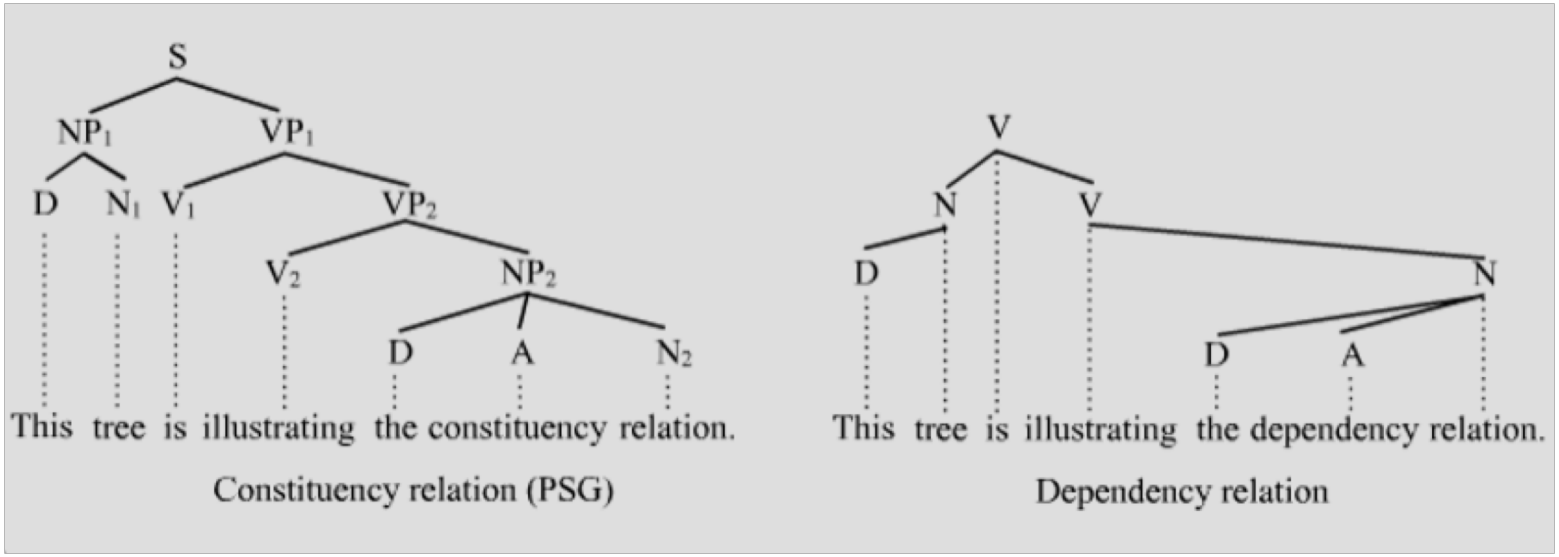


Image source: Wikimedia Commons (Tjo3ya)

different treebanks

- Several types, several languages, modern and historical, e.g.
  - The Penn Treebank (English, several sections, e.g. Wall Street Journal)
  - Index Thomisticus (Latin: Thomas Aquinas)
    - Search: http://www.corpusthomisticum.org/it/index.age
    - The ITT Project: http://itreebank.marginalia.it/
  - PROIEL Treebank corpus
    - parallel corpus of Old Indo-European languages (New Testament + other texts)
    - can be browsed via http://syntacticus.org/
    - can be queried via INESS

- Cf. Jonathan Robie, Biblical Humanities blog post (Dec 2017):

  "Nine Kinds of Ancient Greek Treebanks"

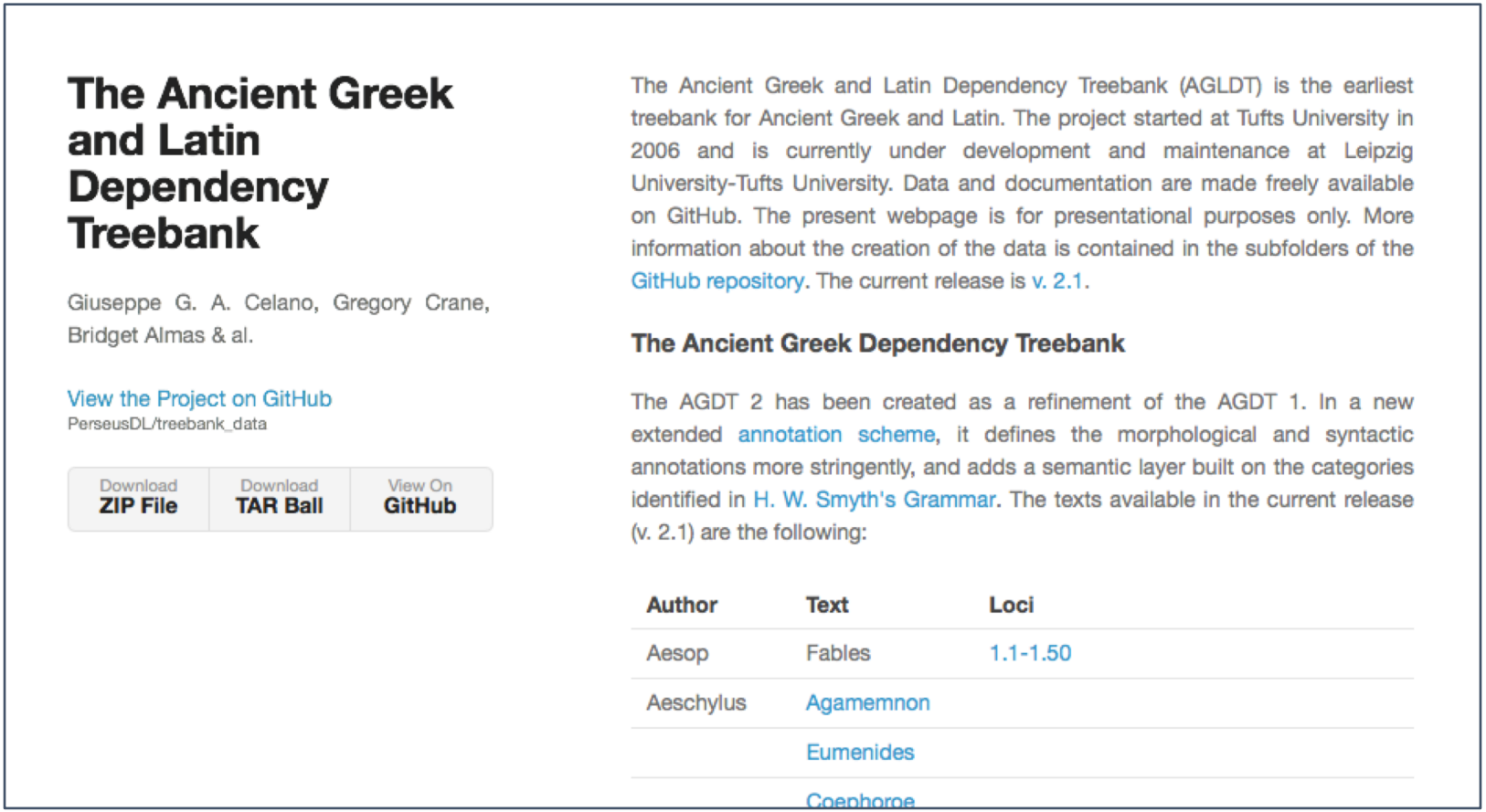
- http://universaldependencies.org/

AGLDT

https://perseusdl.github.io/treebank_data/

Greek: 15 authors (poetry and prose)

Latin: 12 authors (poetry and prose)



The Ancient Greek and Latin Dependency Treebank

Giuseppe G. A. Celano, Gregory Crane, Bridget Almas & al.

View the Project on GitHub
PerseusDL/treebank_data

Download ZIP File | Download TAR Ball | View On GitHub

The Ancient Greek and Latin Dependency Treebank (AGLDT) is the earliest treebank for Ancient Greek and Latin. The project started at Tufts University in 2006 and is currently under development and maintenance at Leipzig University-Tufts University. Data and documentation are made freely available on GitHub. The present webpage is for presentational purposes only. More information about the creation of the data is contained in the subfolders of the GitHub repository. The current release is v. 2.1.

The Ancient Greek Dependency Treebank

The AGDT 2 has been created as a refinement of the AGDT 1. In a new extended annotation scheme, it defines the morphological and syntactic annotations more stringently, and adds a semantic layer built on the categories identified in H. W. Smyth's Grammar. The texts available in the current release (v. 2.1) are the following:

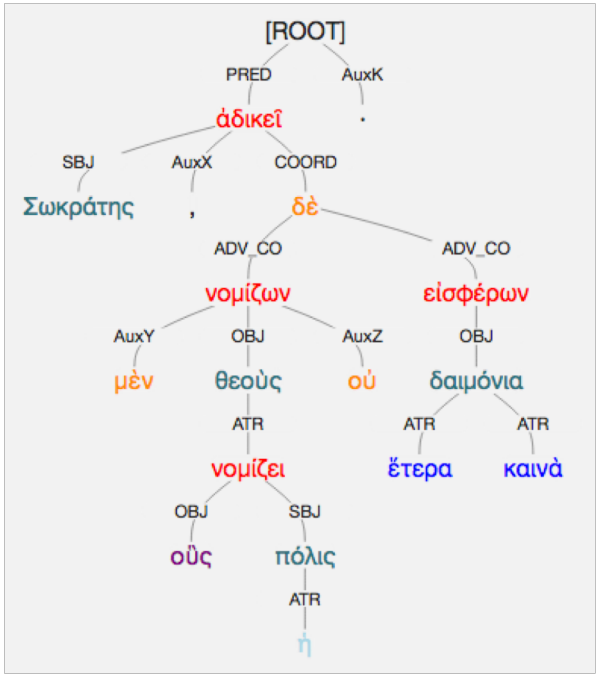
Author	Text	Loci
Aesop	Fables	1.1-1.50
Aeschylus	Agamemnon	
	Eumenides	
	Coephorae	

linguistic annotation according to AGLDT in Arethusa (The Perseids project)

- Tokenization (automatic)
- Lemmatization and morphological parsing (semi-automatic)
- Syntactic relations according to dependency grammar (manual)
- (Semantic layer optional; works in Greek only, Smyth's Grammar) (manual)

➢Following the GUIDELINES is essential!

➢for queries and for future developers of automatic parsers



ἀδικεῖ Σωκράτης οὓς μὲν ἡ πόλις νομίζει θεοὺς

οὐ νομίζων , ἕτερα δὲ καινὰ δαιμόνια εἰσφέρων ·

Socrates is guilty of rejecting the gods acknowledged by the state and of bringing in strange deities.
Xenophon *Mem.*1.1.1

underlying XML

```
<sentence id="3"
    document_id="http://perseids.org/cts5/nemo/citations/urn:cts:greekLit:tlg0032.tlg002.perseus-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="OBJ" head="7"/>
    <word id="4" form="μὲν" lemma="μέν" postag="d--------" relation="AuxY" head="10"/>
    <word id="5" form="ἡ" lemma="ὁ" postag="l-s---fn-" relation="ATR" head="6"/>
    <word id="6" form="πόλις" lemma="πόλις" postag="n-s---fn-" relation="SBJ" head="7"/>
    <word id="7" form="νομίζει" lemma="νομίζω" postag="v3spia---" relation="ATR" head="8"/>
    <word id="8" form="θεοὺς" lemma="θεός" postag="n-p---ma-" relation="OBJ" head="10"/>
    <word id="9" form="οὐ" lemma="οὐ" postag="d--------" relation="AuxZ" 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>
```

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

postag

- nine place string for the morphological annotation

ἀδικεῖ

postag="v3spia---"

1: verb
2: 3rd person
3: singular
4: present
5: indicative
6: active
7: -
8: -
9: -

1: part-of-speech
2: person
3: number
4: tense
5: mood
6: voice
7: gender
8: case
9: degree

πόλις

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

1: noun
2: -
3: singular
4: -
5: -
6: -
7: feminine
8: nominative
9: -

GREEK+LATIN POSTAG KEY

1st: pos	2nd: person	3rd: number	4th: tense	5th: mood	6th: voice	7th: gender	8th: case	9th: degree
noun (n)	1	singular	present (p)	indicative (i)	active	masc	nom (n)	comparative
verb (v)	2	plural	imperfect (i)	subjunctive (s)	passive	fem	gen (g)	superlative
participle (t)	3	dual	perfect (r)	optative (o)	medium	neutr	dat (d)	
adjective (a)			pluperfect (l)	infinitive (n)	med-pass.		acc (a)	
adverb (d)			future perfect (t)	imperative (m)	deponens (d)		voc (v)	
article (l)			future (f)	participle (p)			loc (l)	
(g particle)			aorist (a)	gerund (d)			abl (b)	
conjunction (c)				gerundive (g)				
preposition (r)				supine (u)				
pronoun (p)								
numeral (m)								
exclamation (e)								
punctuation (n)								

guidelines

- Greek:
  - G. Celano, Guidelines for the Ancient Greek Dependency Treebank 2.0
    - includes morphological, syntactic and semantic layer
  - D. Bamman & G. Crane (2008), Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1)
    - This older version should NOT be followed in every case, but is partly still valid and has more explanations; and is useful to know when performing queries on older treebanks
- Latin:
  - D. Bamman, M. Passarotti, G. Crane & S. Raynaud (2007), Guidelines for the Syntactic Annotation of Latin Treebanks (1.3)

- Examples in the Guidelines are limited
  - Compare to already existing annotations (with caution!)

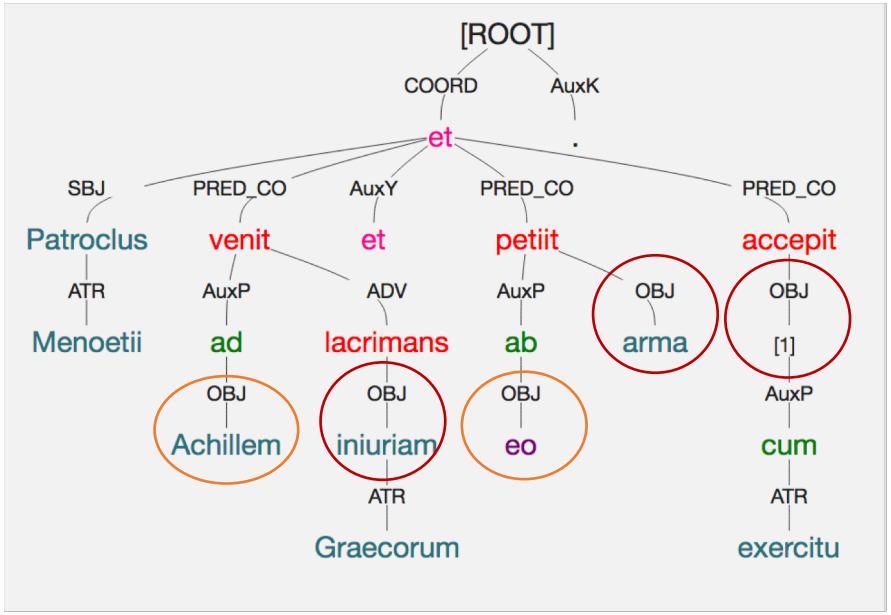
possible syntactic stumbling blocks

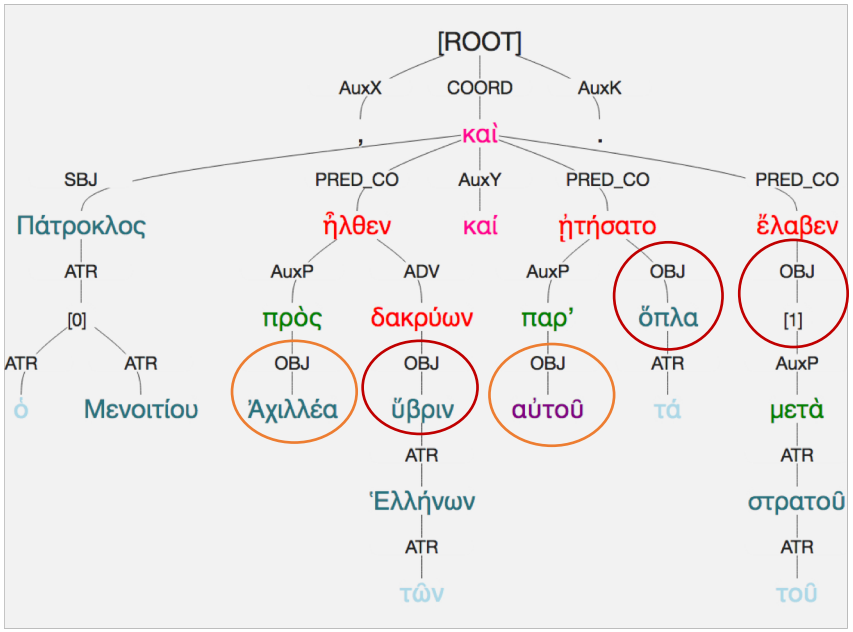
- OBJ (direct object, also indirect object and verbal complement i.e. required argument)
- ADV (= adverbial; specifies the circumstances under which a verb, adjective or adverb takes place. Can be an adverb, prep.phrase, noun, participle, subordinate clause)

Example: Bilingual school text (E. Dickey, *Learning Latin the Ancient Way. Latin Textbooks from the Ancient World.* 2016, pp. 133–135.)

Patroclus Menoetii venit ad Achillem lacrimans Graecorum iniuriam et petiit ab eo arma et accepit cum exercitu.

Πάτροκλος ὁ Μενοιτίου ἦλθεν πρὸς Ἀχιλλέα δακρύων τῶν Ἑλλήνων ὕβριν καί ἠτήσατο παρ’ αὐτοῦ τά ὅπλα, καὶ ἔλαβεν μετὰ τοῦ στρατοῦ.





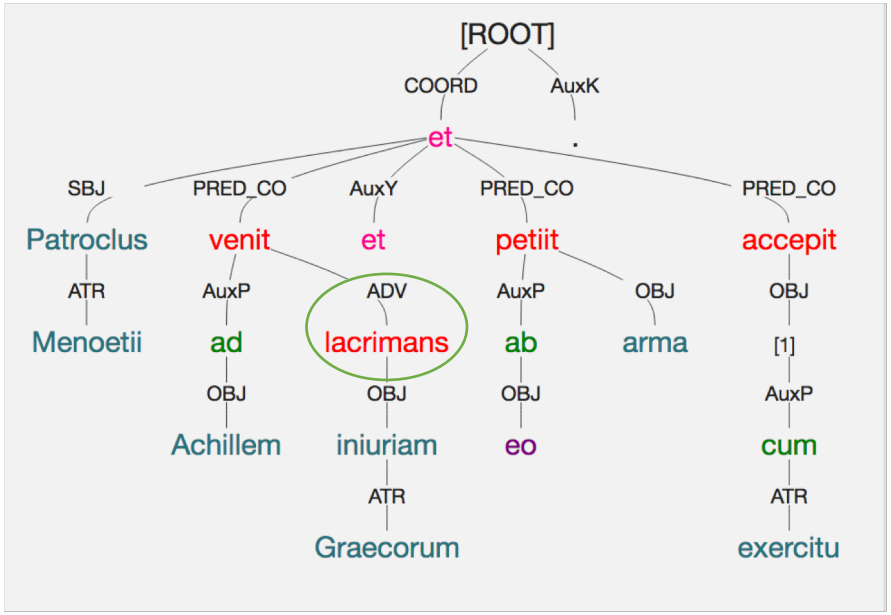
possible syntactic stumbling blocks

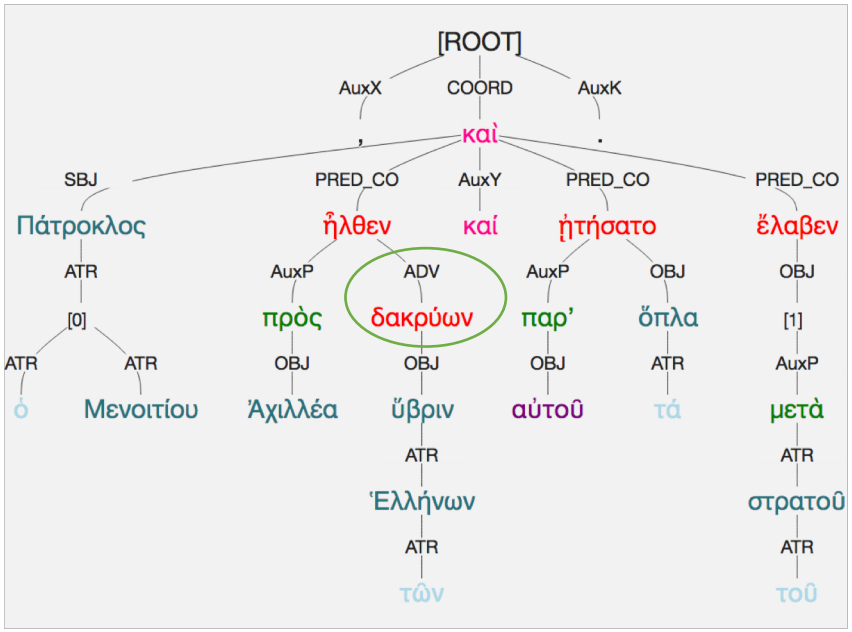
- OBJ (direct object, also indirect object and verbal complement i.e. required argument)
- ADV (= adverbial; specifies the circumstances under which a verb, adjective or adverb takes place. Can be an adverb, prep.phrase, noun, participle, subordinate clause)

Example: Bilingual school text (E. Dickey, *Learning Latin the Ancient Way. Latin Textbooks from the Ancient World.* 2016, pp. 133–135.)

Patroclus Menoetii venit ad Achillem lacrimans Graecorum iniuriam et petiit ab eo arma et accepit cum exercitu.

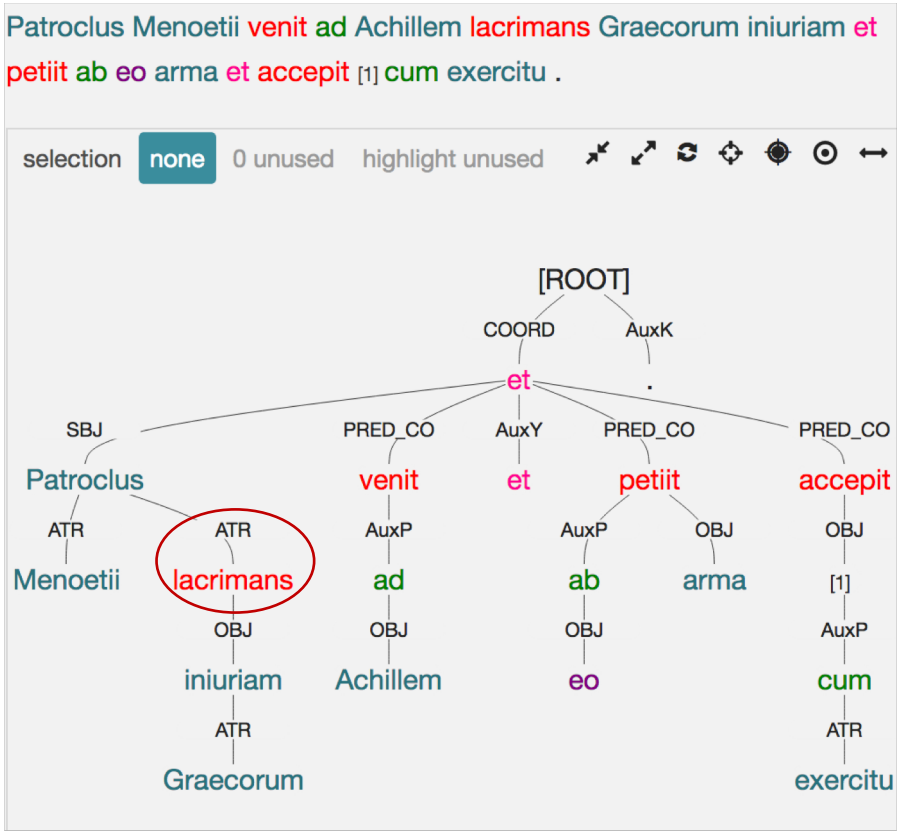
Πάτροκλος ὁ Μενοιτίου ἦλθεν πρὸς Ἀχιλλέα δακρύων τῶν Ἑλλήνων ὕβριν καί ἠτήσατο παρ’ αὐτοῦ τά ὅπλα, καὶ ἔλαβεν μετὰ τοῦ στρατοῦ.

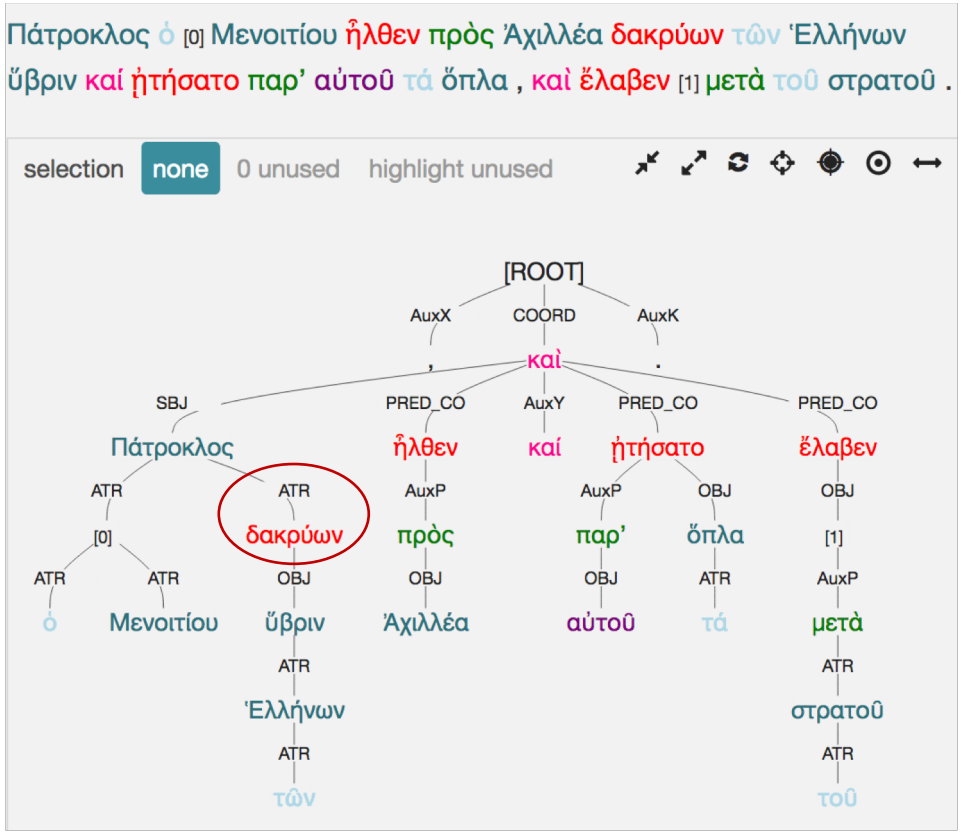




matters of interpretation, e.g.

- Participle (or adjective etc.) as ADV
  - depending on the PRED (previous slide; specifying circumstances of the verb)
- Participle as ATR
  - depending on the SBJ (here; specifying only its head, the subject)





annotating literary vs. documentary texts

LITERATURE

- manuscript tradition
- In annotation only one specific edition at a time (e.g. from the Perseus Digital Library)
- what about different readings (apparatus criticus)?
  - it is possible to change word forms or sentences during the annotation in the XML

DOCUMENTARY TEXTS

- direct source; often fragmentary
- papyri and ostraca available in TEI EpiDoc XML in papyri.info
- many inscriptions also available in TEI EpiDoc XML, but in different places
- preprocessing of the data needed in order to gain texts suitable for treebanking
  - Leiden mark-up contains important information (supplements and regularized forms)
  - Linguistically the original forms are the interesting ones

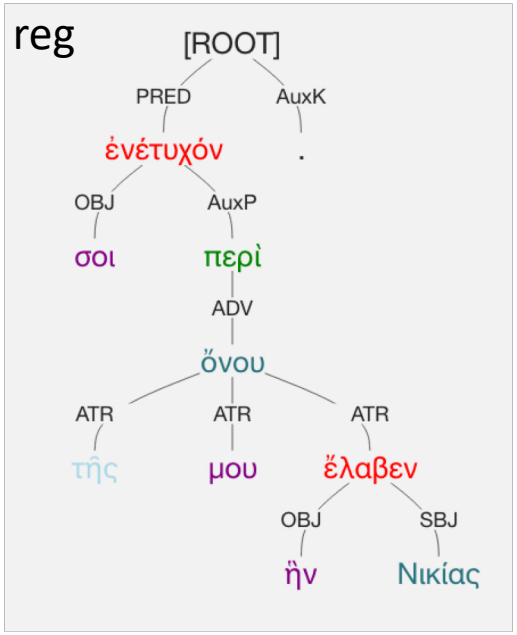
processing of papyrological texts

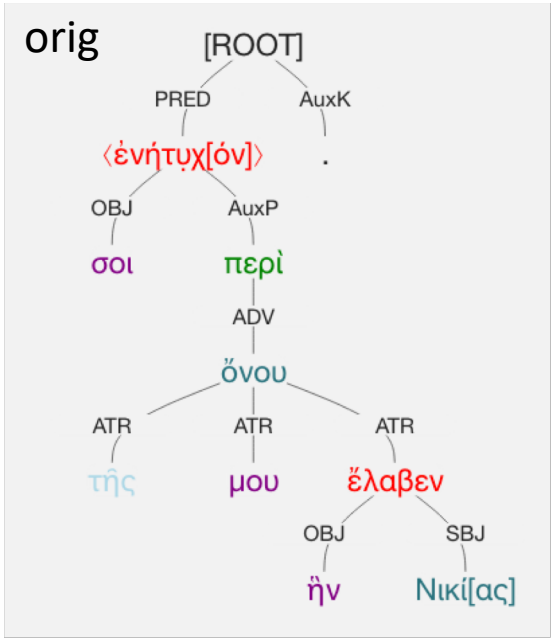
- Linguistically annotated corpora are based on tokenized plain text; TEI EpiDoc XML data is not plain: tags within the text
- Sematia tool was developed for getting rid of the tags but preserving the information they contain (now a newer version of Sematia (PapyGreek) is being developed)
  - ➢Two parallel layers for each text: ORIGINAL / REGULARIZED

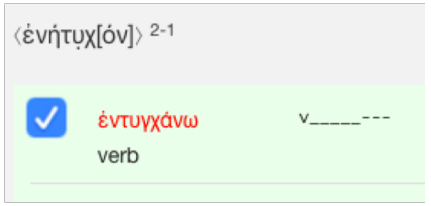
P.Mich. 1.29, 1–2

Ζήνωνι χαίρειν Σενχώνς. ἐνήτυχ[όν](*)
σοι περὶ τῆς ὄνου μου ἣν ἔλαβεν Νικί[ας].

To Zenon greeting from Senchons. I made a petition to you about my donkey which Nikias took.







Structure of the session

1. Introduction to Dependency Grammar and treebanking Latin and Greek (MV)

2. Treebanking in action - demonstration of the annotation tool Arethusa (PY)