Sunoikisis Digital Classics Spring 2018

Session 6, March 1, 2018

Treebanking 1

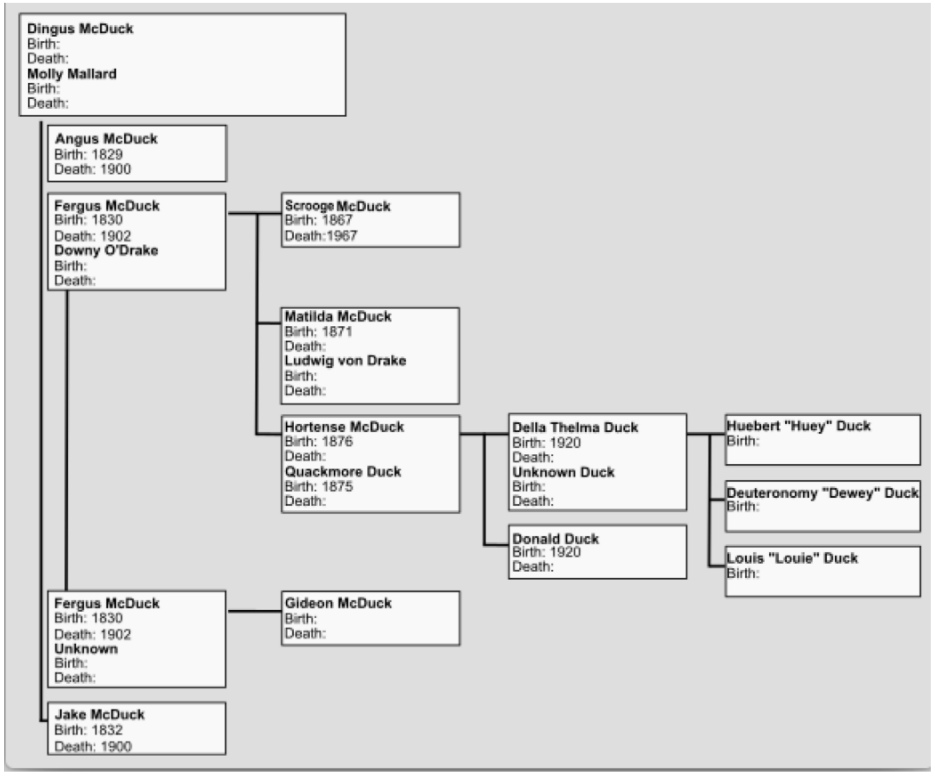
Polina Yordanova (Sofia & London)
Marja Vierros (Helsinki)

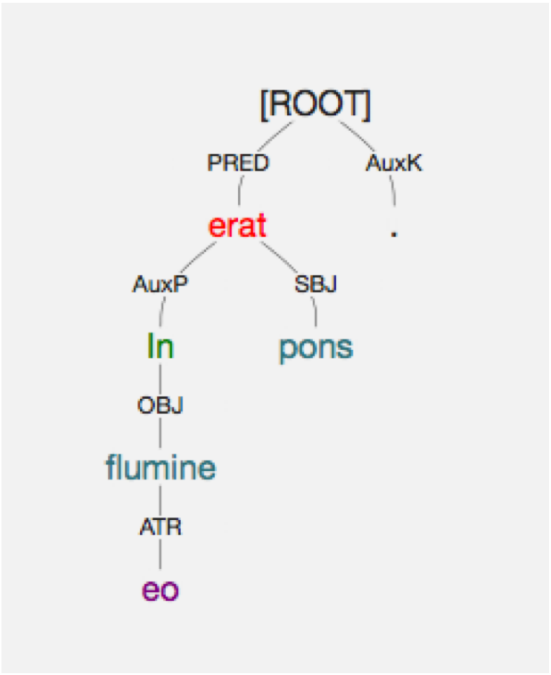
Structure of the session

1. Introduction to treebanks and linguistic annotation (MV)
2. How to annotate using Arethusa (PY)
3. How to annotate papyri, ostraca, tablets (MV)

TREE

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





In eo flumine pons erat .

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)
    - 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, verb 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

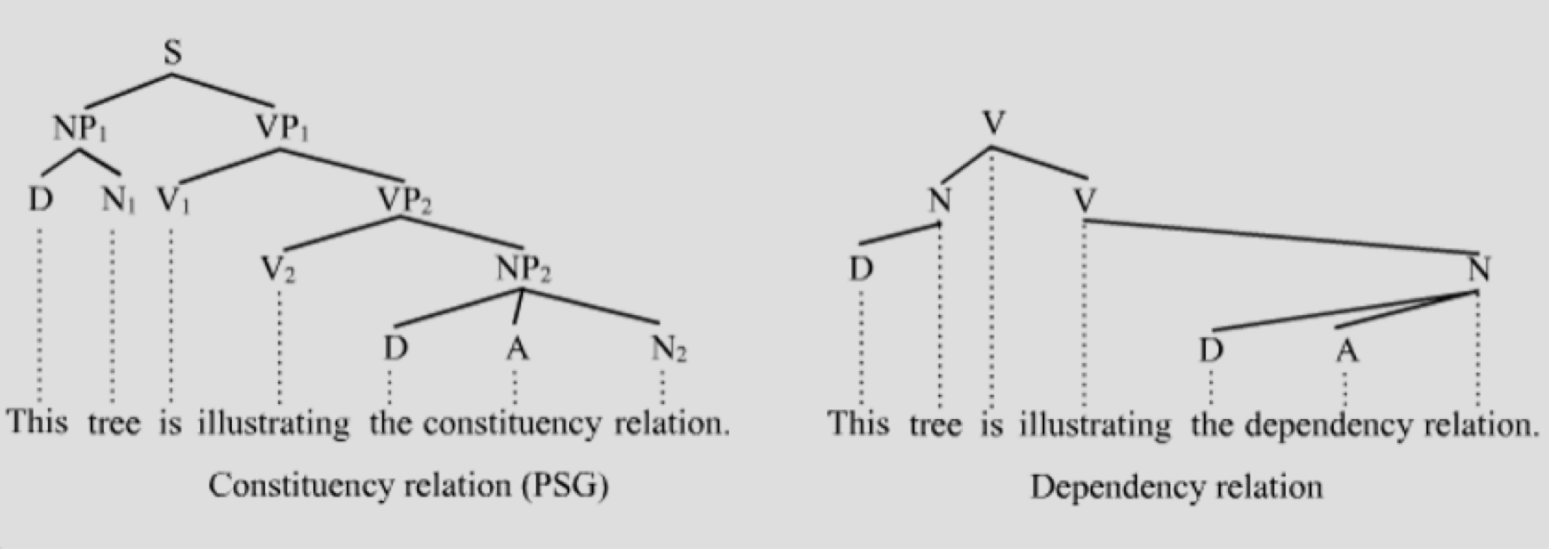


Image source: Wikimedia Commons (Tjo3ya)

TREEBANKS

- Several types, several languages, modern or 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/

- Jonathan Robie, Biblical Humanities blog post "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:

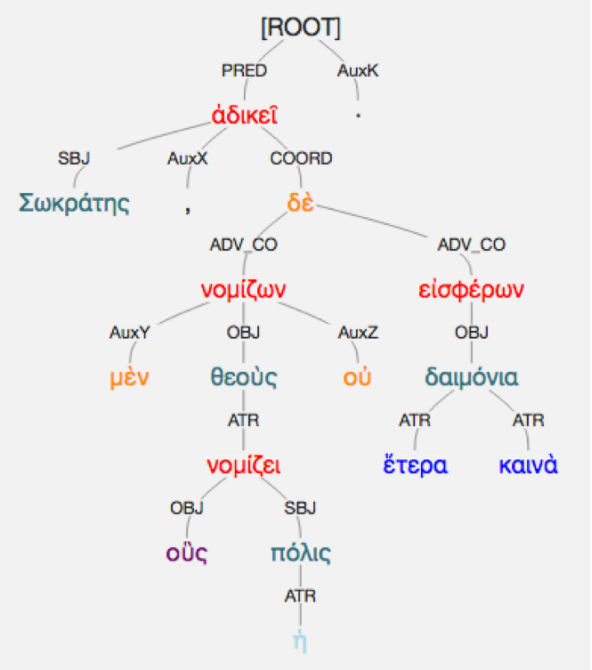
Author	Text	Loci
Aesop	Fables	1.1-1.50
Aeschylus	Agamemnon	
	Eumenides	
	Coephorae	

Linguistic annotation according to AGLDT in Arethusa

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

➢Following the guidelines is essential!

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



Socrates is quilty 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: -

1st: pos	2nd: person	3rd: number	4th: tense	5th: mood	6th: voice	7th: gender	8th: case	9th: degree
noun	1	singular	present	indicative	active	masc	nom	comparative
verb	2	plural	imperfect	subjunctive	passive	fem	gen	superlative
participle	3	dual	perfect	optative	medium	neutr	dat	
adjective			pluperfect	infinitive	med-pass.		acc	
adverb			future perfect	imperative			voc	
article			future	participle			loc	
g particle			aorist	gerund				
conjunction				gerundive				
preposition								
pronoun								
numeral								
interjection								
exclamation								
punctuation								

Structure of the session

1. Introduction to treebanks and linguistic annotation (MV)
2. How to annotate using Arethusa (PY)
3. How to annotate papyri, ostraca, tablets (MV)

Structure of the session

1. Introduction to treebanks and linguistic annotation (MV)
2. How to annotate using Arethusa (PY)
3. How to annotate papyri, ostraca, tablets (MV)

Why treebank papyri?

- Papyrological corpus offers different type of linguistic source for Greek and Latin than the literary corpora
  - Historical development of linguistic forms, linguistic variation and language change
  - Sociohistorical linguistics
- The whole existing corpus (DDbDP) of documentary papyri is digitally available at papyri.info (EpiDoc XML)
  - Now /soon also literary and paraliterary papyri (DCLP) at litpap.info!

Preprocessing needed

- Leiden mark-up for editorial information in print editions = TEI EpiDoc XML mark-up in digital form
- 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
  - ➢Two (three) parallel layers for each text / act of writing

ORIGINAL	Editorial STANDARD
Presents the original forms which the ancient writer wrote and nothing more	Presents supplied text, regularized spellings and expanded abbreviations
Includes **dummy markers** for the supplied forms in the standard layer in order to keep tokenization identical in both layers	Includes mostly whole text, but can also include some **dummy markers**

Read more: Vierros, Marja and Erik Henriksson. 2017. "Preprocessing Greek Papyri for Linguistic Annotation." *Journal of Data Mining and Digital Humanities.* Available: http://jdmdh.episciences.org/paper/view/id/1385

SEMATIA STEPS FOR ANNOTATION

https://sematia.hum.helsinki.fi/docs/how_to_use.html

1. **Sign in** to Sematia (https://sematia.hum.helsinki.fi/user/) using Google account (You will also need an account in Perseids (http://sosol.perseids.org/sosol/))
2. **Import the document** you have selected to work with from the papyri.info by clicking the button **New**
3. **Sematia creates automatically three layers** and divides the document into acts of writing by hand shifts
4. **To expand the view** of your imported document (i.e. to see the layer options), click the **plus** symbol.
5. **Select a layer of your document** (click the buttons “original” or “standard”; ”variation” layer is not yet in use)
6. **Export the plain text** to the annotation framework Arethusa by clicking the paper plane icon.
7. **Annotate** the text and submit to Sematia board.
8. Annotate **the other layer**, too. (In fact, I recommend annotating both layers simultaneously in two windows)