Papyrology and Linguistic Annotation

How can we make TEI EpiDoc XML corpus and Treebanking work together?

Marja Vierros, University of Helsinki

Digital Classicist Summer Seminars, ICS London
July 25, 2014

Papyrology and Linguistic Annotation

How can we make TEI EpiDoc XML corpus and Treebanking work together?

Marja Vierros, University of Helsinki

Digital Classicist Summer Seminars, ICS London
July 25, 2014

SEMATIA: Linguistic Annotation of the Greek Documentary Papyri – Detecting and Determining Contact-Induced, Dialectal and Stylistic Variation

ACADEMY OF FINLAND

Papyri as a source for studying Greek language

Papyri = historical fieldwork

- Documentary papyri preserve us texts written for **everyday use** on perishable material; as close as we can get to the ancient speakers
- Archive = speech community
- Written text = covered with education, formulaic language

Examples



Variation

- languages change all the time
- synchronic and diachronic varieties

- studying different varieties tell about
  a) history of the language
  b) community and speakers

Variation

- languages change all the time
- synchronic and diachronic varieties

- studying different varieties tell about
  a) history of the language
  b) community and speakers

Papyri = historical fieldwork

- Documentary papyri preserve us texts written for **everyday use** on perishable material; as close as we can get to the ancient speakers
- Archive = speech community
- Written text = covered with education, formulaic language

Examples

Phonology / orthography

πυροῦ wheat.GEN	/pyrú/	
πουροῦ	/purú/	*O.Narm.* 42 and 47
ποιροῦ	/pyrú/	*O.Narm.* 46 and 86

- Merging of /y/ and /oi/ internal Greek development
- Egyptian did not have front vowel /y/
  → /u/ and /y/ often confused by Egyptians writing Greek

Dahlgrén, S. (in preparation). *Effects of language contact: impact of Egyptian phonology onto Greek vowel orthography. Study of the Narmouthis Greek ostraka collection*

Syntax and beyond

καλῶς πυή[σεις], ἄδελφε, πέμψε μοι τὸ τοῦτο
Please, brother, send me this (sum) here.

O. Claud. II 243, 2–3

πέμψε: AOR. IND. 3SG
πέμψαι: AOR. INF.
πέμψον: AOR. IMP. 2SG

Pronunciation for all of the above: /pémpsa/ ?

Leiwo, Martti. 2010. "Imperatives and Other Directives in the Greek Letters from Mons Claudianus" In T. V. Evans, D. D. Obbink (eds.) *The Language of the Papyri.* (Oxford: Oxford University Press), 97–119.

Morphosyntax

ὁμολογεῖ	Νεχθανοῦπις	Πατσεοῦτος …	συνκεχωρηκέναι	
agree:Pr.3sg	Nechthanoupis:Nom	Patseous:Gen	cede:Pf.Inf	
Πετεαρσεμθεῖ	Πανοβχούνιος	καὶ	τοῖς	ἀδελφοῖς
Petehersemtheus:Dat	Panobchounis:Gen	and	Art.Dat.Pl	brothers:Dat.Pl
Πετεσοῦχος	καὶ	Φαγῶνις	καὶ	Ψεννῆσις …
Petesouchos:Nom	and	Phagonis:Nom	and	Psennesis:Nom

Nechthanoupis, son of Patseous, agrees … to have ceded to Petehersemtheus, son of Panobchounis, and to the brothers Petesouchos and Phagonis and Psennesis…

P.Grenf. 2.25, 4–7 (103 BCE)

Not knowing cases? – or rather a pragmatic strategy of phrase-initial inflection (combined with native language of no case inflection)

Vierros, M. 2012. *Bilingual Notaries in Hellenistic Egypt. A Study of Greek as a Second Language.* Brussel.

Phonology / orthography

πυροῦ wheat.GEN - - -	/pyrú/	
πουροῦ	/purú/	*O.Narm.* 42 and 47
ποιροῦ	/pyrú/	*O.Narm.* 46 and 86

- Merging of /y/ and /oi/ internal Greek development
- Egyptian did not have front vowel /y/
  → /u/ and /y/ often confused by Egyptians writing Greek

Dahlgrén, S. (in preparation). *Effects of language contact: impact of Egyptian phonology onto Greek vowel orthography. Study of the Narmouthis Greek ostraka collection*

Morphosyntax

ὁμολογεῖ	Νεχθανοῦπις	Πατσεοῦτος …	συνκεχωρηκέναι
agree:PR.3SG	Nechthanoupis:NOM	Patseous:GEN	cede:PF.INF

Πετεαρσεμθεῖ	Πανοβχούνιος	καὶ	τοῖς	ἀδελφοῖς
Peteharsemtheus:DAT	Panobchounis:GEN	and	ART.DAT.PL	brothers:DAT.PL

Πετεσοῦχος	καὶ	Φαγῶνις	καὶ	Ψεννῆσις …
Petesouchos:NOM	and	Phagonis:NOM	and	Psennesis:NOM

Nechthanoupis, son of Patseous, agrees … to have ceded to Peteharsemtheus, son of Panobchounis, and to the brothers Petesouchos and Phagonis and Psennesis…

P.Grenf. 2.25, 4–7 (103 BCE)

Not knowing cases? – or rather a pragmatic strategy of phrase-initial inflection (combined with native language of no case inflection)

Vierros, M. 2012. *Bilingual Notaries in Hellenistic Egypt. A Study of Greek as a Second Language*. Brussel.

Syntax and beyond

καλῶς πυή[σεις], ἄδελφε, πέμψε μοι τὸ τοῦτο

O. Claud. II 243, 2–3

Please, brother, send me this (sum) here.

πέμψε: AOR. IND. 3SG

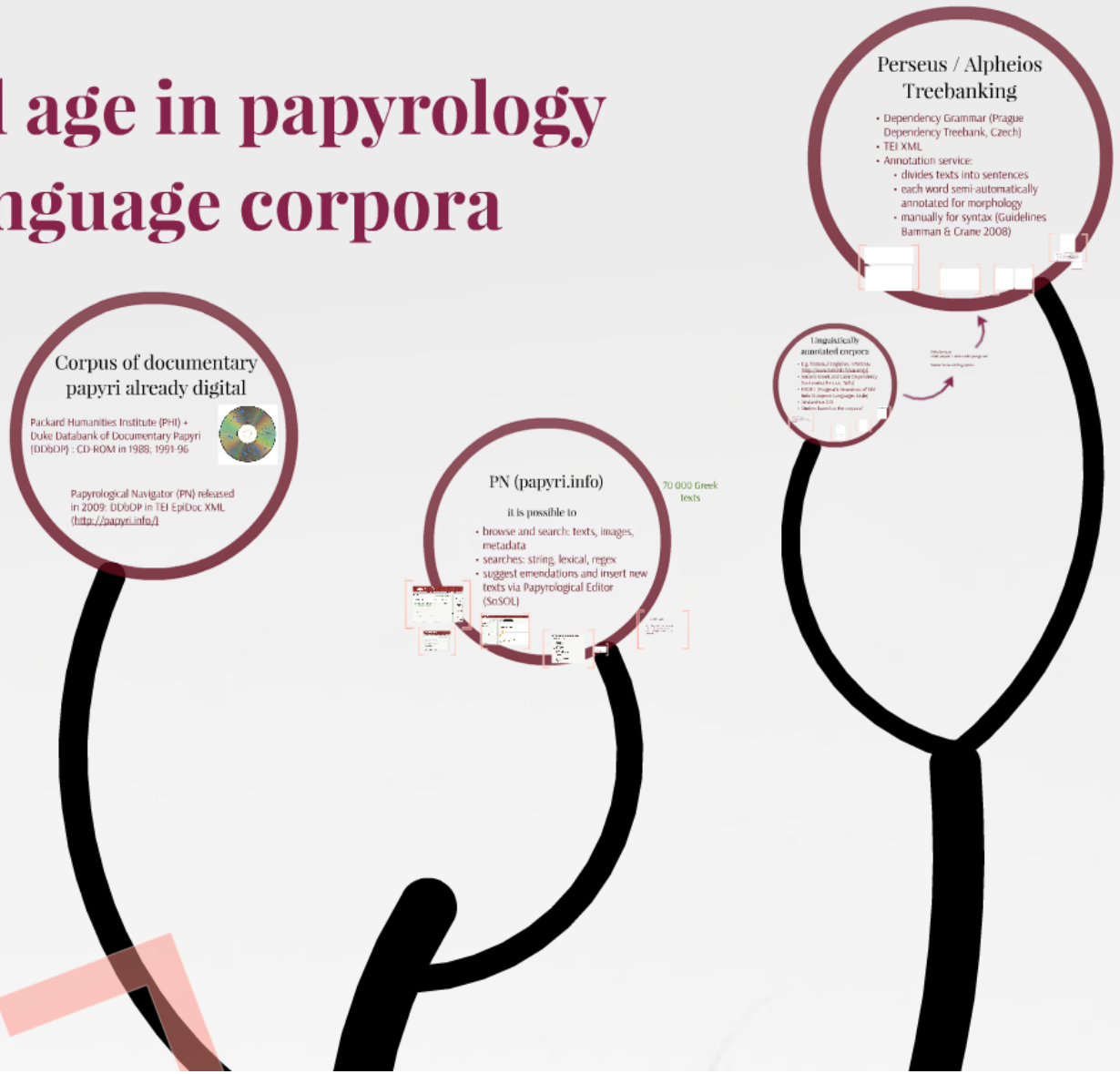
πέμψαι: AOR. INF.

πέμψον: AOR. IMP. 2SG

Pronunciation for all of the above: /pémpsə/ ?

Leiwo, Martti. 2010. "Imperatives and Other Directives in the Greek Letters from Mons Claudianus" in T. V. Evans, D. D. Obbink (eds.) *The Language of the Papyri.* (Oxford: Oxford University Press), 97–119.

Digital age in papyrology and language corpora



Corpus of documentary papyri already digital

Packard Humanities Institute (PHI) + Duke Databank of Documentary Papyri (DDbDP) : CD-ROM in 1988; 1991-96

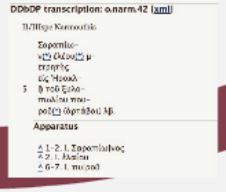
Papyrological Navigator (PN) released in 2009: DDbDP in TEI EpiDoc XML (http://papyri.info/)

PN (papyri.info)

70 000 Greek texts

it is possible to

- browse and search: texts, images, metadata
- searches: string, lexical, regex
- suggest emendations and insert new texts via Papyrological Editor (SoSOL)



it is NOT possible to

- find orthographic variation, except when searching certain words
- search for linguistic structures or anomalies


Papyri.info
Browse: DDbDP HGV APIS TM Number or Search: Data Bibliography
Search
New Search
Search
LEX ἀγορανόμος
within
chars
and
or
not
then
near
lex
clear
regex
abbr
start-not
end-not
Convert from betacode as you type
ignore capitalization
ignore diacritics/accents
Text
Metadata
Translations
<<
Please se
Selecting a
documents
the control
adding new
narrowed a
More ab
More ab
More ab
More ab
More ab
More ab


Papyri.info
Browse: DDbDP HGV APIS TM Number or Search: Data Bibl
Search
New Search
Search
REGEX παρ[εη]μην
within
chars
and
or
not
then
near
lex
clear
regex
abbr
start-not
end-not
Convert from betacode as you type
ignore capitalization
ignore diacritics/accents
Text
Metadata
Translations

Papyri.info

Browse: DDbDP HGV APIS TM Number or Search: Data Bibliography

Refine Search

New Search | Search

within chars

and or not then near lex clear -

regex abbr start-not end-not

- [ ] Convert from betacode as you type
- [x] ignore capitalization
- [x] ignore diacritics/accents

(•) Text ( ) Metadata ( ) Translations

Series --- All values --- Vol.

or

Collection --- All values --- ID #

Provenance --- All values ---

Nome Arsinoites (4)

Date on or after --- All values --- BCE (•) CE ( )

Date before --- All values --- BCE (•) CE ( )

(•) Loose ( ) Strict

Language Ancient Greek (8)

Translation language None (8)

Show only records with images from:

Papyri.info | Other sites | Print publications | Go

<<

Substring: #πουρο X
Target: text
No Caps: On
No Marks: On

8 hits.

Identifier	Title	Location	Date
o.bodl 2 1033	keiner	Theben	197 CE
1. Πουροῦσιος κ̣αὶ μ(έτοχοι) ἐπιτη(ρηταὶ) τέ(λους) γερδ(ίων)			
o.narm 42	Registrazione ...	Narmuthis (Arsinoites)	101 CE - 300 CE
4. εἰς Ἡρακλ– 5. ᾶ τοῦ ξυλο– 6. πωλίου που– 7. ρο̣ῦ(*) (ἀρτάβαι) λβ.			
o.narm 47	Notizen ...	Narmuthis (Arsinoites)	154 CE - 210 CE
3. πόλι(*) που– 6. ρο̣ῦ(*) 5. διετίαν ὅ ἐστιν που– 6. ρο̣ῦ(*) ἀρτάβας δέκα κα– 7. τ' ἔτος καὶ δραχμὰς			
p.berl.leihg 1 12	Konto ...	Euhemeria (Arsinoites)	210 CE
9. ὁ αὐτὸς Παῦνι θ δη(μοσίων) ἄλλαι πουρο̣ῦ(*) ἀρτάβαι εἴ̣κ̣[ο]σι δύο· γ(ίνονται) (ἀρτάβαι) κβ.			
p.kron 44	Ricevuta ...	Tebtynis (Arsinoites)	148 CE - 149 CE
12. τὰς τοῦ πουρο̣ῦ(*) ἀρτάβας κ̣αθὼς̣ π[ρόκ(ειται)][.]			
p.lond 5 1673	Account ...	Ibion (Antaiopolites oder Hermopolites)	501 CE - 600 CE
81. / Πουροσε Ἰ(*)ωάννου ν(όμισμα) α π(αρὰ) ζ̣ ⊾			
p.lond.herm 1	A ...	Hermopolis	546 CE - 547 CE

DDbDP transcription: o.narm.42 [xml]

II/IIIspc Narmouthis

Σαραπίω-
ν(*) ἐλέου(*) μ-
ετρητὴς
εἰς Ἡρακλ-
ᾶ τοῦ ξυλο-
πωλίου που-
ροῦ(*) (ἀρτάβαι) λβ.

Apparatus

^ 1–2. l. Σαραπίω|νος
^ 2. l. ἐλαίου
^ 6–7. l. πυ|ροῦ

```
<TEI n="0035;;42" xml:id="o.narm.42" xml:lang="en"><teiHeader><fileDesc>
    <ab>
    <lb n="1"/><choice><reg>Σαραπίω
    <lb n="2" break="no"/>νος</reg><orig>Σαραπίω
    <lb n="2" break="no"/>ν</orig></choice>
        <choice><reg>ἐλαίου</reg><orig>ἐλέου</orig></choice> μ
    <lb n="3" break="no"/>ετρητὴς
    <lb n="4"/>εἰς Ἡρακλ
    <lb n="5" break="no"/><unclear>ᾶ</unclear> τοῦ ξυλο
    <lb n="6" break="no"/>πωλίου
        <choice><reg>πυ
    <lb n="7" break="no"/>ροῦ</reg><orig>που
    <lb n="7" break="no"/>ροῦ</orig></choice>
        <expan><ex>ἀρτάβαι</ex></expan><num value="32">λβ</num>.
    </ab></div></body></text></TEI>
```

it is NOT possible to

- find orthographic variation, except when searching certain words
- search for linguistic structures or anomalies

Linguistically annotated corpora

- E.g. historical Englishes: VARIENG (http://www.helsinki.fi/varieng/)
- Ancient Greek and Latin Dependency Treebanks (Perseus, Tufts)
- PROIEL (Pragmatic Resources of Old Indo-European Languages, Oslo)
- Aristarchus 2.0
- Studies based on the corpora?

- Varieng ePublication series
- 14 volumes, starting from year 2007
- http://www.helsinki.fi/varieng/series/volumes/

- Varieng ePublication series
- 14 volumes, starting from year 2007
- http://www.helsinki.fi/varieng/series/volumes/

BRILL

Journal of Greek Linguistics 9 (2009) 56–92

JGL

brill.nl/jgl

Breaking down and putting back together: analysis and synthesis of New Testament Greek

Dag T. T. Haug[a], Hanne M. Eckhoff[b], Marek Majer[c], Eirik Welo[d]
University of Oslo, Norway
[a]daghaug@ifikk.uio.no
[b]h.m.eckhoff@ifikk.uio.no
[c]marek.majer@ifikk.uio.no
[d]eirikwelo@gmail.com

Abstract
In this paper we first briefly describe the design of a corpus containing the Koine Greek original text of the New Testament and its translations in to Gothic, Latin, Old Church Slavic and Armenian. We then discuss extensively the annotation that we have applied in each layer of annotation: morphology and syntax, information structure, animacy, and token alignment. For each type of annotation we provide some preliminary results and applications that draw on it, often in combination with other layers of annotation.

Keywords
corpus linguistics, information structure, New Testament Greek, pragmatics, syntax

1. Background

Pragmatic Resources in Old Indo-European Languages (PROIEL)[1] is a project based at the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo. The main objective of the project is to investigate what morphological and syntactic resources different older Indo-European (IE) languages utilize for expressing categories related to information structure. In particular we focus on how word order, definiteness, anaphoric expressions, discourse particles and participles are employed in information packaging.

The goal of PROIEL is to study comparatively how information packaging works in old IE languages. In order to do this, we have developed a richly annotated corpus[2] consisting of the Koine Greek original of the New Testament

[1] http://www.hf.uio.no/ifikk/proiel
[2] Available at http://foni.uio.no:3000

DOI 10.1163/156658409X12529372103308

Non-projectivity in the Ancient Greek Dependency Treebank

Francesco Mambrini
The Center for Hellenic Studies
Washington, DC
fmambrini@chs.harvard.edu

Marco Passarotti
Università Cattolica del Sacro Cuore
Milano, Italy
marco.passarotti@unicatt.it

Abstract

In this paper, we provide a quantitative analysis of non-projective constructions attested in the Ancient Greek Dependency Treebank (AGDT). We consider the different types of formal constraints and metrics that have become standardized in the literature on non-projectivity (planarity, well-nestedness, gap-degree, edge-degree). We also discuss some of the linguistic factors that cause non-projective edges in Ancient Greek. Our results confirm the remarkable extension of non-projectivity in the AGDT, both in terms of quantitative incidence of non-projective nodes and for their complexity, which is not paralleled by the corpora of modern languages considered in the literature. At the same time, the usefulness of other constraint (especially well-nestedness) is confirmed by our researches.

1 Introduction

The "free" word-order of Ancient Greek (AG) is a notorious problem for philologists and linguists. In spite of several studies devoted to the subject, the tendencies that govern the disposition of words and constituents in the sentence still lack a comprehensive explanation. Strictly connected to the word-order issue is the relevant amount of discontinuous constituents, which even casual readers of AG texts can experience[1].

The dependency-based treebanks of Classical languages (AG and Latin) that have been recently made available enable us to reconsider this long debate in the light of the abundant work on non-projective structures in dependency trees. Non-projectivity (see 2 for a formal definition) is a key issue in dependency grammar, both from the formal point of view and from a more descriptive linguistic perspective. From the standpoint of natural language processing, non-projectivity is also known to affect the efficiency of dependency parsers.

In a first attempt to improve parsing performances on AG, Mambrini and Passarotti (2012) reported that the amount of non-projective arcs occurring in the available treebanks of Classical languages is significantly higher than that attested in the corpora of modern languages used for CoNLL-X (Buchholz and Marsi, 2006, 155, tab. 1) and CoNLL 2007 shared tasks (Nivre et al., 2007, 920, tab. 1). Furthermore, the non-projective rate in the Ancient Greek Dependency Treebank is higher than in Classical and Medieval Latin (Passarotti and Ruffolo, 2010, 920, tab. 1).

In this paper, we want to discuss this claim in depth and substantiate it by applying to AG data the standard metrics for the different kinds of non-projective constructions established in the literature.

The paper is organized as follows. Section 2 provides a definition of the formal constraints considered and of the metrics that will be used: non-projectivity, planarity, well-nestedness, on the one hand, and gap-degree and edge-degree on the other. Section 3 introduces the corpus that will be tested, the Ancient Greek Dependency Treebank (AGDT).

Section 4 presents the evidence provided by the data. In 4.1 we report the results for the different constraints and metrics defined in section 2. Results for the distribution of non-projectivity in the different genres of the corpus are given and commented in 4.2.

In section 5, we discuss some of the linguistic issues that cause non-projectivity. Finally, section 6 reports our conclusions and sketches possible directions for additional research.

[1] On AG word-order see more recently Dik (1995; 2007), with bibliography of previous studies. On discontinuous structures see Devine and Stephens (2000).

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 177–186,
Prague, August 27–30, 2013.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 101 APRIL 2014 97–110

A computational study on preverbal and postverbal accusative object nouns and pronouns in Ancient Greek

Giuseppe G. A. Celano

Tufts University, USA

Abstract

Many studies try to determine whether Ancient Greek is an OV or VO language. All of them, however, fail to conduct a research whose method is entirely clear. This paper presents the first attempt to quantify the number of verbs governing preverbal or postverbal accusative object nouns or pronouns in single or coordinate independent clauses in Homer's Iliad and Odyssey, Herodotus' Histories, and the New Testament, by providing results which are fully verifiable and reproducible. I prove that as for the parameter OV vs. VO there is great variation in the texts, which suggests a change over time from OV order in Homer to VO order in the New Testament. The figures for Herodotus' Greek prove a quasi-exact match between OV order and VO order.

1. Introduction

Ancient Greek (AG) is an Indo-European language allowing great freedom of word order at both clausal and subclausal level. A great variety of studies were conducted on the position of subject (S), verb (V), and object (O) to establish the "normal" order of such constituents (see, among others, Ebeling (1902); Friederich (1975); Cervin (1990); Kwong (2005)). They however provide discordant results, which are impossible to evaluate (see, for example, Cervin (1990); Taylor (1994)): the sample analyzed is often limited and, what is worse, the method employed to count the instances of a given word order is usually not precisely defined: e.g., Friederich (1975) counted 195 constructions in Iliad 5.1–296, but it is not clear what exactly he means by a construction.

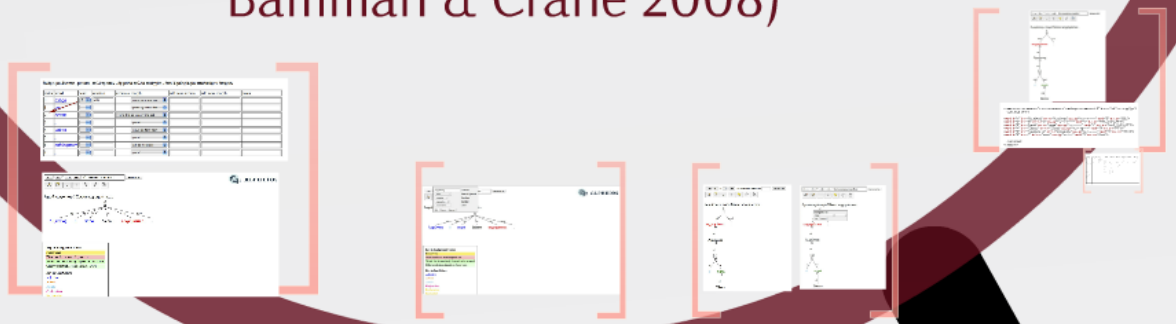
Corresponding author: giuseppe.celano@tufts.edu
Cite as: Giuseppe G. A. Celano. A computational study on preverbal and postverbal accusative object nouns and pronouns in Ancient Greek. The Prague Bulletin of Mathematical Linguistics No. 101, 2014, pp. 97–110.
doi: 10.2478/pralin-2014-0006.

Ratio between
what you put in versus what you get out

Reason to use existing system

Perseus / Alpheios Treebanking

- Dependency Grammar (Prague Dependency Treebank, Czech)
- TEI XML
- Annotation service:
  - divides texts into sentences
  - each word semi-automatically annotated for morphology
  - manually for syntax (Guidelines Bamman & Crane 2008)



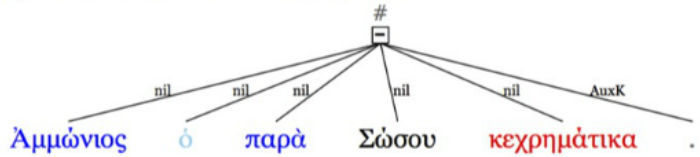
ἄνδρα μοι ἔννεπε , μοῦσα , πολύτροπον , ὃς μάλα πολλὰ πλάγχθη , ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσεν

index	word	head	relation	lemma + morph	add new lemma	add new morph	notes
0	ἄνδρα	2	OBJ	noun sg masc acc			
1	μοι			pron sg masc dat			
2	ἔννεπε			verb 3rd sg imperf ind act			
3	,			punc*			
4	μοῦσα			noun sg fem nom			
5	,			punc*			
6	πολύτροπον			adj sg neut acc			
7	,			punc*			

1 ◂◂ 6 ◂ 7 ▸ 8 ▸▸ 8 Go to sentence number Sentence list

ALPHEIOS

Ἀμμώνιος ὁ παρὰ Σώσου κεχρημάτικα .



Key to Background Colors

Focus word

Word that focus word depends on

Words that immediately depend on focus word

Other words that depend on focus word

Key to Text Colors

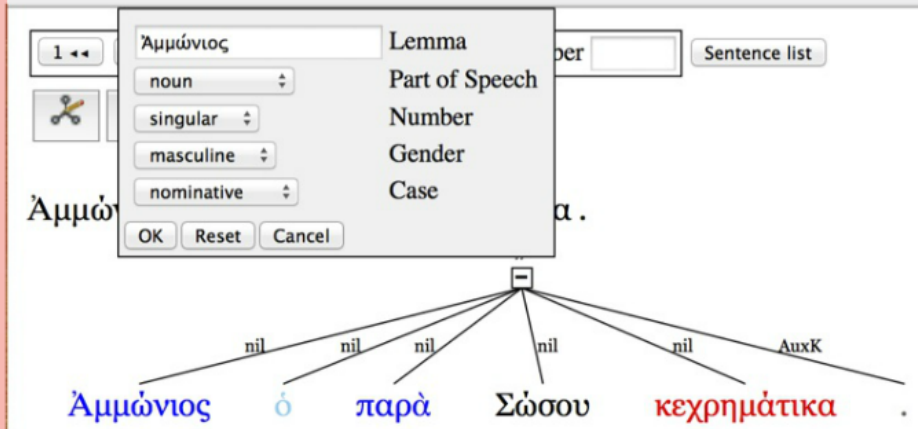
Adjective

Adverb

Article

Conjunction

Exclamation



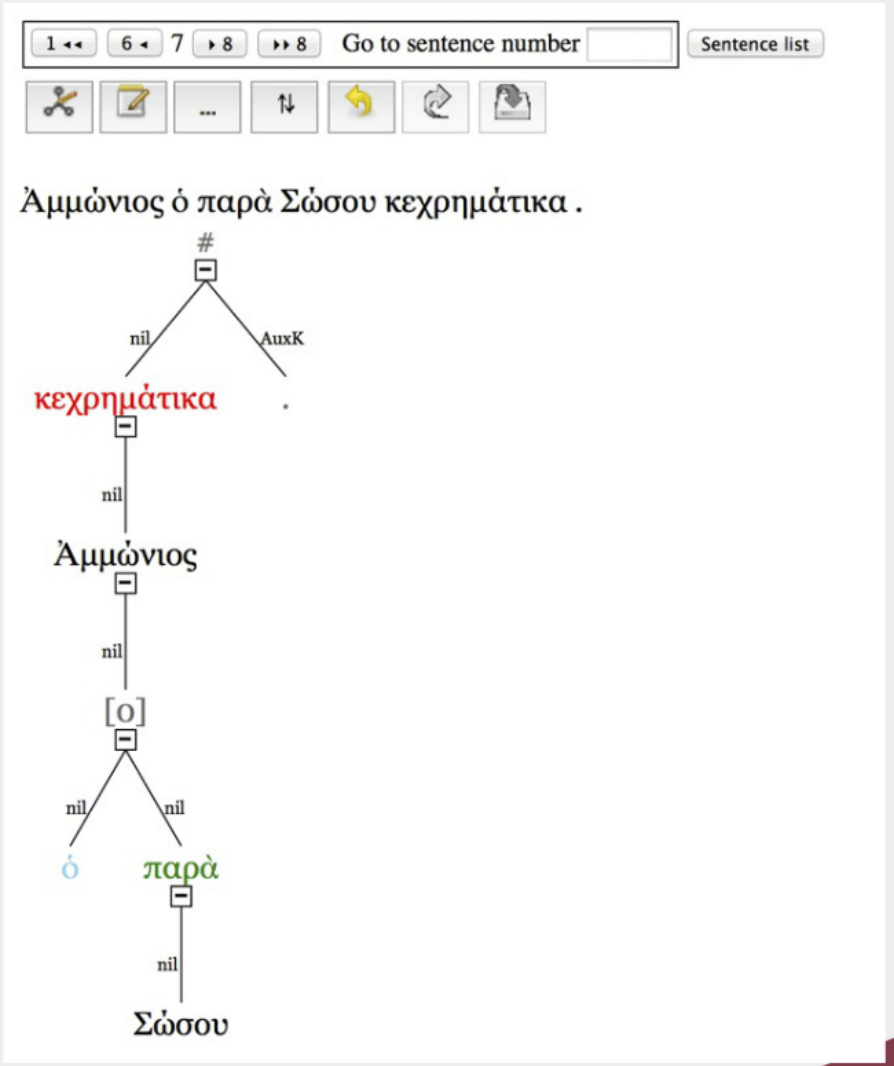
ALPHEIOS

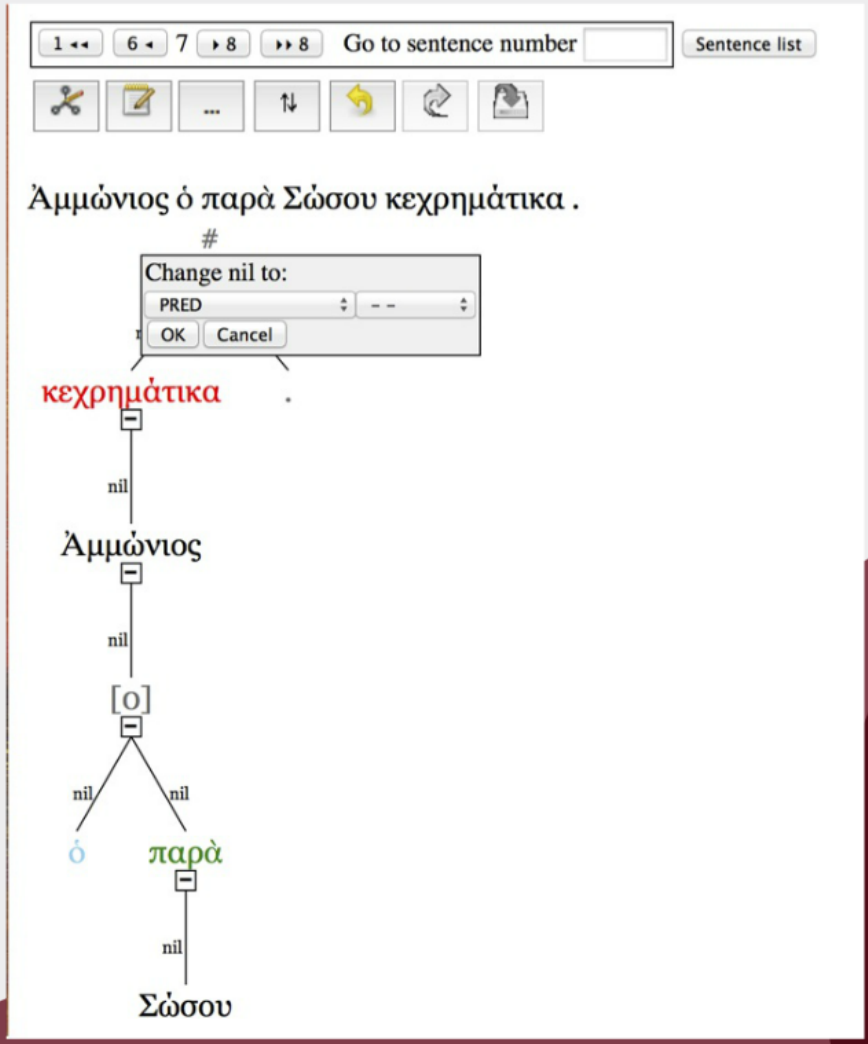
Key to Background Colors

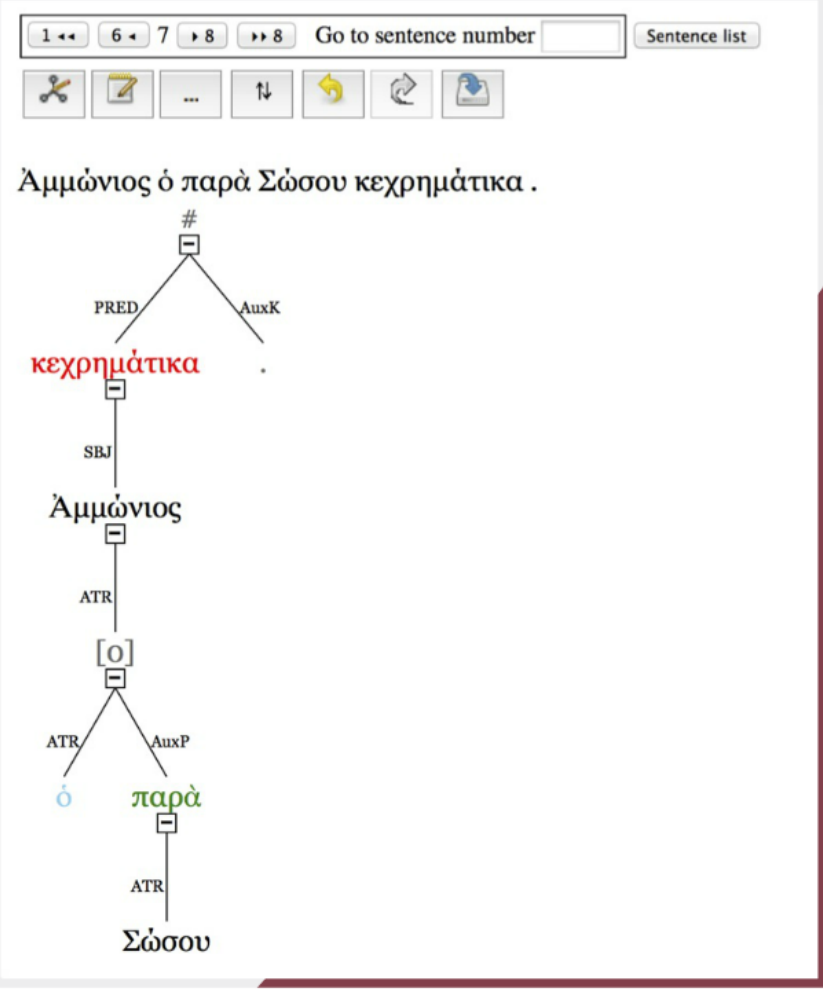
Focus word
Word that focus word depends on
Words that immediately depend on focus word
Other words that depend on focus word

Key to Text Colors

Adjective
Adverb
Article
Conjunction
Exclamation
Interiection


1 ◂◂ 6 ◂ 7 ▸ 8 ▸▸ 8 Go to sentence number Sentence list
Ἀμμώνιος ὁ παρὰ Σώσου κεχρημάτικα .
#
nil AuxK
κεχρημάτικα .
nil
Ἀμμώνιος
nil
[o]
nil nil
ὁ παρὰ
nil
Σώσου


1 ◂◂ 6 ◂ 7 ▸ 8 ▸▸ 8 Go to sentence number Sentence list
Ἀμμώνιος ὁ παρὰ Σώσου κεχρημάτικα .
#
Change nil to:
PRED - -
OK Cancel
κεχρημάτικα .
nil
Ἀμμώνιος
nil
[o]
nil nil
ὁ παρὰ
nil
Σώσου




```
<treebank xmlns:treebank="http://nlp.perseus.tufts.edu/syntax/treebank/1.5" format="aldt" xml:lang="grc">
   <sentence id="1">

<word id="1" form="Ἀμμώνιος" lemma="Ἀμμώνιος" postag="n-s---m-------" head="5" relation="SBJ"/>
<word id="2" form="ὁ" lemma="ὁ" postag="l-s---mn------" head="1" relation="ATR_ExD0_ATR"/>
<word id="3" form="παρὰ" lemma="παρά" postag="r----------" head="1" relation="AuxP_ExD0_ATR"/>
<word id="4" form="Σώσου" lemma="Σῶσος" postag="n-s---mg------" head="3" relation="ATR"/>
<word id="5" form="κεχρημάτικα" lemma="χρηματίζω" postag="v--r----------" head="0" relation="PRED"/>
<word id="6" form="." lemma="punc1" postag="u--------" head="0" relation="AuxK"/>

   </sentence>
</treebank>
```

POSTAG

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
t 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								

Treebanking and papyri – Problems and wishes

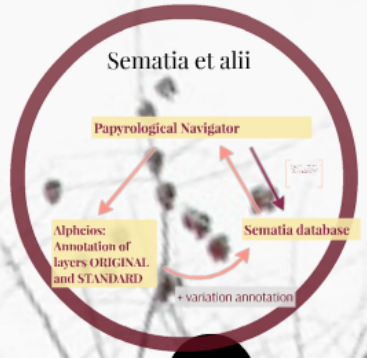
Challenges of papyrological material in annotation service

- Test corpus of 50 documents into annotation service
- Annotation service does not support EpiDoc XML -> XSLT to strip away markup that breaks up words -> loss of information
- Choice was made to use <orig> and <sic> tags instead of <reg> and <corr> tags -> both would be better
- Abbreviations in expanded form -> not truthful
- supplied words in lacunae...

Proposal: Layers and variation tagging

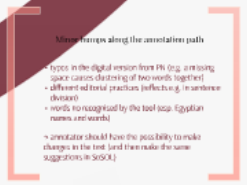
- Several layers of annotation of the same text:
  - ORIGINAL (only what is in the papyrus: <orig>, <sic>, no expanded abbreviations, no forms whose markers are in a lacuna)
  - STANDARD (the emendations of the editors: <corr> and <reg>; abbreviations expanded and supplied forms in lacunae. Sublayers for competing emendations and supplements?)
  - VARIATION (new tagset; base text ORIGINAL)

Sematia et alii

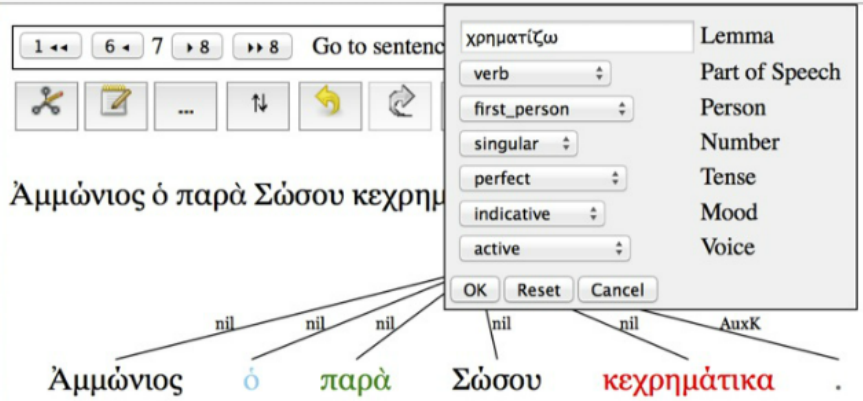


Challenges of papyrological material in annotation service

- Test corpus of 50 documents into annotation service
- Annotation service does not support EpiDoc XML -> XSLT to strip away markup that breaks up words -> loss of information
- Choice was made to use <orig> and <sic> tags instead of <reg> and <corr> tags -> both would be better
- Abbreviations in expanded form -> not truthful
- supplied words in lacunae...



Abbreviations



ALPHEIOS

Ἀμμώ(νιος) ὁ παρὰ Σώσου κεχρη(μάτικα).

Key to Background Colors

- Focus word
- Word that focus word depends on
- Words that immediately depend on focus word
- Other words that depend on focus word

Key to Text Colors

- Adjective
- Adverb
- Article
- Conjunction
- Exclamation
- Interjection

Lacunae

P. Adl. 1 II, 7–8:

προπωληταὶ καὶ βεβαιωταὶ τῶν κατὰ τὴν ὠνὴν ταύτην πάντων Πτόλλις καὶ Νεχοῦθις καὶ Χαλῆις καὶ Ἀρ[πα]ῆσις οἱ ἀ̣π̣ο̣δ̣ό̣μ̣[ενοι, οὓς ἐδέξατο Ἰσίδωρος] ὁ πριάμ̣ε̣ν̣ος.

Minor bumps along the annotation path

- typos in the digital version from PN (e.g. a missing space causes clustering of two words together)
- different editorial practices (reflects e.g. in sentence division)
- words no recognised by the tool (esp. Egyptian names and words)

→ annotator should have the possibility to make changes in the text (and then make the same suggestions in SoSOL)

Proposal: Layers and variation tagging

- Several layers of annotation of the same text:
  - ORIGINAL (only what is in the papyrus: <orig>, <sic>, no expanded abbreviations, no forms whose markers are in a lacuna)
  - STANDARD (the emendations of the editors: <corr> and <reg>; abbreviations expanded and supplied forms in lacunae. Sublayers for competing emendations and supplements?)
  - VARIATION (new tagset; base text ORIGINAL)

- Database management system (e.g. eXist?)
- Perseids platform? http://sites.tufts.edu/perseids/

Variation tagset?

- Element <var>
- type: pho, (mor, syn)
- value: the characters in Betacode, postag...
- For example:

- Database management system (e.g. eXist?)
- Perseids platform? http://sites.tufts.edu/perseids/

Variation tagset?

- Element <var>
- type: pho, (mor, syn)
- value: the characters in Betacode, postag…
- For example:

```
<var type=”pho” value=”ou” not=”u”>πουροῦ</var>
<var type=”pho” value=”e” not=”?”>πέμψε</var>

Or even:
<var type=”pho” value=”ou” not=”u” before=”p” after=”r”>πουροῦ</var>
<var type=”pho” value=”e” not=”?” before=”y” after=”#”>πέμψε</var>


<var type=”mor” value=”n-s---mn-” not=”n-s---md-”>Πετεσοῦχος</var>
```

Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material

Timo Korkiakangas – Matti Lassila

University of Helsinki – University of Tampere

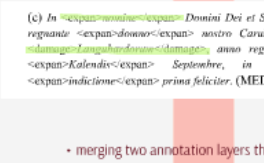
Abstract

This article proposes a method that makes possible the linguistic study of textually difficult hand-written materials which are imperfectly preserved. These materials include medieval manuscripts, letters, and legal as well as private documents. With these, the normal treebanking procedure is not sufficient. We present the case of medieval Latin charter texts, i.e., private documents, that 1) are partly fragmentary and 2) exhibit massive use of abbreviations, e.g., *chartul* for *chartulam* 'charter'. In addition, 3) charter texts are highly formulaic and display passages that differ from each other in their language use. It is not possible to ascertain the inflexional endings of most of the fragmentary and abbreviated words, so a method of excluding them from morphological (but not from syntactic) analysis is needed. Moreover, due to the varying degree of formulaicity in certain parts of charter texts, the language of these parts must be studied separately. Therefore, a method of merging two XML layers is introduced. One layer that contains lemmatic, morphological, and syntactic analysis according to the Perseus Latin Dependency Treebank standard is aligned with the other layer that contains textual information (abbreviations, fragmentary words, diplomatic segmentation).

1 Encoding textual data in treebanks

Before the invention of printing, all texts were written by hand. In the Middle Ages, the period through which, for example, all the Classical Latin literature was transmitted, it became more and more customary to abbreviate certain common Latin words or inflexional endings. The scribes wrote, for example, *dns* for *dominus* 'lord' or *chartulā*, with a small horizontal stroke over the final *a*, for *chartulam* 'charter', where the abbreviated *-m* stands for an accusative ending. The practice of abbreviating poses limitations to how the abbreviated words can be used in linguistic analysis. If correctly applied, linguistic analysis of medieval Latin texts can tell us, for example, how well the scribes managed their Latin, which traits of spoken language infiltrated the written code, and whether there was regional variation.

Along with the abbreviations, the state of preservation of the physical object, on which the text is written, may affect linguistic study. Both the medieval literary texts and the texts that were written for practical purpose,


(c) *In* <expan>nomine</expan> *Domini Dei et S*
regnante <expan>domno</expan> *nostro Caru*
<damage>*Langubardorum*</damage>, *anno reg*
<expan>*Kalendis*</expan> *Septembre, in*
<expan>*indictione*</expan> *prima feliciter.* (MEI

• merging two annotation layers th

ACRH 2013 (Annotation of Corpora for Research in the Humanities Workshop) http://www.bultreebank.org/ACRH-3/ACRH-3Proceeding.pdf

(c) *In* <expan>*nomine*</expan> *Domini Dei et Salvatori nostri Iesu Christi regnante* <expan>*domno*</expan> *nostro Carulo rex Francorum seo et* <damage>*Langubardorum*</damage>, *anno regni eius in Etalia quinto,* <expan>*Kalendis*</expan> *Septembre, in natale sancti Reguli,* <expan>*indictione*</expan> *prima feliciter*. (MED 172)

- merging two annotation layers through XPointers

Sematia et alii

Papyrological Navigator

Alpheios: Annotation of layers ORIGINAL and STANDARD

Sematia database

+ variation annotation

- Word-id's or sentence-id's as "stand-off" mark-up into PN xml?
- When the PN text changes (someone suggest a new reading *vel. sim.* via the Editor), keeping Sematia in sync needs action, too

helsinki.fi

THANKS

Bridget Almas (Perseus/Tufts)

Hugh Cayless & Josh D. Sosin
(The Duke Collaboratory for
Classics Computing, the DC3)


contact:
marja.vierros[ät]helsinki.fi
T
Bridg
Hugh
(The
Class

Papyrology and Linguistic Annotation

How can we make TEI EpiDoc XML corpus and Treebanking work together?

Marja Vierros, University of Helsinki

Digital Classicist Summer Seminars, ICS London
July 25, 2014