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

- Languages change all the time
- Synchronic and diachronic varieties

Variation
- Studying different varieties will tell us about the history of the language, community, and speakers

Papyri = historical fieldwork
- Documentary papyri preserve texts written for everyday use and as close as we can get to the ancient speakers
- Archive = speech community
- Written text = covered with education, formalic language

Examples

- [Image of papyri]
• languages change all the time
• synchronic and diachronic varieties

Variation

• 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
Phonology / orthography

Examples

Syntax and beyond

Morphosyntax

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

πυροὖ /pyrû/
wheat.GEN

O.Narm. 42 and 47

πουρο⇡ /purû/

O.Narm. 46 and 86

ποιρο⇡ /pyrû/

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

Morphosyntax

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

καλῶς πυῆ[σεῖς], ἀδελφε, πέμψε μοι τὸ τοῦτο
Please, brother, send me this (sum) here.

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

Pronunciation for all of the above: /pέmpso/ ?

Digital age in papyrology and language corpora

- Perseus / Alpheios Treebanking
  - TEI XML
  - Annotation service
  - database tools for researchers
  - text and Logos systematically annotated for morphology
  - research software (Golubovskii 2006)

- PN (papyri.info)
  - It is possible to:
    - browse and search texts, images, metadata
    - search, using lexical, logical
    - search the whole site and insert new text via hypertext editor (pawiki)

- Corpus of documentary papyri already digital
  - Papyrus Institute (University of Oxford)
  - Digital Edition of documentary papyri
  - CD-ROM in PDF & HTML

  - CD-ROM with hypertext (http://papyrus.info)
Corpus of documentary papyri already digital


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

it is possible to

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

70 000 Greek texts
Search

REGEX παρ[η]μην

within
chars

and or not then near lex
regex abbr start-not end-not

clear

Convert from betacode as you type
ignore capitalization
ignore diacritics/accents

Text Metadata Translations
<table>
<thead>
<tr>
<th>Identifier</th>
<th>Title</th>
<th>Location</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>o.bodi 2 1033</td>
<td>keiner</td>
<td>Theben</td>
<td>197 CE</td>
</tr>
<tr>
<td>o.narm 42</td>
<td>Registrazione ...</td>
<td>Narmuthis (Arsinoites)</td>
<td>101 CE - 300 CE</td>
</tr>
<tr>
<td>o.narm 47</td>
<td>Notizen ...</td>
<td>Narmuthis (Arsinoites)</td>
<td>154 CE - 210 CE</td>
</tr>
<tr>
<td>p.berl.leih 112</td>
<td>Konto ...</td>
<td>Euhemeria (Arsinoites)</td>
<td>210 CE</td>
</tr>
<tr>
<td>p.kron 44</td>
<td>Ricevuta ...</td>
<td>Teblynis (Arsinoites)</td>
<td>148 CE - 149 CE</td>
</tr>
<tr>
<td>plond 5 1673</td>
<td>Account ...</td>
<td>Ibion (Antaiopolites oder Hermopolites)</td>
<td>501 CE - 600 CE</td>
</tr>
<tr>
<td>plond.herm 1</td>
<td>Α ...</td>
<td>Hermopolis</td>
<td>546 CE - 547 CE</td>
</tr>
</tbody>
</table>
Π/ΠΙ Ispc Narmouthis

Σαραπίων
ν(*) ἐλέου(*) μ-
ετρητῆς
εἰς Ἡρακλ-
5 ἃ τοῦ ξυλο-
πωλίου που-
ροῦ(*) (ἀρτάβαι) λβ.

Apparatus

^ 1–2. ἰ. Σαραπίων|νος
^ 2. ἰ. ἐλαίου
^ 6–7. ἰ. πυρμοῦ
<TEI n="0035;42" xml:id="o.narm.42" xml:lang="en"><teiHeader><fileDesc>
  <ab>
    <lb n="1"/><choice><reg>Σαραπίω</reg><orig>Σαραπίω</orig></choice>
    <lb n="2" break="no"/>νος</reg><orig>Σαραπίω</orig></choice>
    <lb n="2" break="no"/>ν</orig></choice>
    <choice><reg>ἐλαίου</reg><orig>ἐλέου</orig></choice> μ
    <lb n="3" break="no"/>ἐτρητής
    <lb n="4"/>εἰς Ἡρακλ
    <lb n="5" break="no"/><unclear>ἀ</unclear> τοῦ ἔυλο
    <lb n="6" break="no"/>πωλίου
      <choice><reg>πυ</reg>
      <lb n="7" break="no"/>ροῦ</reg><orig>που</orig></choice>
      <lb n="7" break="no"/>ροῦ</orig></choice>
      <expan><ex>ἀρτάβας</ex></ex><expan><num value="32">λβ</num>.</expan>
  </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/
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 into 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/c3000

© Koninklijke Brill NV, Leiden, 2009 DOI 10.1631/jfgel095x12525372103308
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.

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.
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.
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)
Άνδρα μοι ἐννεπε, μούσα, πολύτροπον, ὡς μάλα πολλὰ πλάγχη, ἔπει Τροίης ἱερὸν πτολεψίδον ἔπεσεν.
Αμμώνιος ὁ παρὰ Σώσου κεχρημάτικα.
Αμμώνιος ὁ παρὰ Σώσου κεχρημάτικα.

[Diagram on the left]

Αμμώνιος ὁ παρὰ Σώσου κεχρημάτικα.

[Diagram on the right]
Αμμώνιος ὁ παρὰ Σώσου κεχρημάτικα.

<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="παρὰ" postag="r------------" head="1" relation="AuxP_ExD0_ATR"/>
    <word id="4" form="Σώσου" lemma="Σώσου" 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>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>1</td>
<td>singular</td>
<td>present</td>
<td>indicative</td>
<td>active</td>
<td>masc</td>
<td>nom</td>
<td>comparative</td>
</tr>
<tr>
<td>verb</td>
<td>2</td>
<td>plural</td>
<td>imperfect</td>
<td>subjunctive</td>
<td>passive</td>
<td>fem</td>
<td>gen</td>
<td>superlative</td>
</tr>
<tr>
<td>t participle</td>
<td>3</td>
<td>dual</td>
<td>perfect</td>
<td>optative</td>
<td>medium</td>
<td>neutr</td>
<td>dat</td>
<td></td>
</tr>
<tr>
<td>adjective</td>
<td></td>
<td></td>
<td>pluperfect</td>
<td>infinitive</td>
<td>med-pass.</td>
<td></td>
<td>acc</td>
<td></td>
</tr>
<tr>
<td>adverb</td>
<td></td>
<td></td>
<td>future</td>
<td>imperative</td>
<td></td>
<td></td>
<td></td>
<td>voc</td>
</tr>
<tr>
<td>article</td>
<td></td>
<td></td>
<td>perfect</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>loc</td>
</tr>
<tr>
<td>g particle</td>
<td></td>
<td></td>
<td></td>
<td>participle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conjunction</td>
<td></td>
<td></td>
<td></td>
<td>gerund</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preposition</td>
<td></td>
<td></td>
<td></td>
<td>gerundive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pronoun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>numeral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interjection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exclamation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Challenges of papyrological material in annotation service

- Test corpus of 50 documents into annotation service
- Annotation service does not support Epitext XML, so XSLT to strip away markup that breaks up words and loss of information
- Choice was made to use <orig> and <src> tags instead of <seg> and <com> tags, both would have worked
- Abbreviations in expanded form (not fruitful supplied words in lacunae)

Proposal: Layers and variation tagging

- Several layers of annotation of the same text
  - ORIGINAL (only what is in the papyrus
    - strip lines, no expanded abbreviations, no form whose words are in a lacuna)
  - STANDARD (the emendations of the editors:
    - corruptions, emendations, expanded and supplied forms in lacunae. Sub-layers for competing emendations and supplements)
  - SUPPLIERS (how T4G et al., tagged ORIGINAL)

Treebanking and papyri – Problems and wishes

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...
`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
- Interjection
Lacunae

P. Adl. 1 II, 7–8:

προσωληται καὶ βεβαιωται τῶν κατὰ τὴν ὀψιν ταύτην πάντων Πτόλλις καὶ Νεκοὺθις καὶ Χαλῆς καὶ Ἁρ[πα]ήσος οἱ ἀποδοῦσ[ενοι, οὖσ ἐδέξατο Ἱσίδωρος] ὁ πριάμενος.
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/](http://sites.tufts.edu/perseids/)
Variation tagset?

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

```html
<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-">Πετεσσοῦχος</var>
```
Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material

Timo Korkialangas – 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., 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 syntactic) analysis is needed. Moreover, due to the varying degree of formality 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 lemmatized, 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, dvs for dominus ‘lord’ or chartulā, with a small horizontal stroke over the final ā, for chartulam ‘charter’, where the abbreviated -ān 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 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)
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
THANKS

Bridget Almas (Perseus/Tufts)

Hugh Cayless & Josh D. Sosin
(The Duke Collaboratory for Classics Computing, the DC3)
contact: marja.vierros[at]helsinki.fi
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