Adapting Penn Treebank-style Annotation for Ancient Greek

Jana E. Beck

November 20, 2011

Goals:

- introduce a work-in-progress: a syntactically parsed corpus of historical Greek
- introduce Penn Treebank-style annotation, a type of phrase-structure annotation
- discuss three major modifications of Penn Treebank-style annotation necessary for annotating Ancient Greek:
  1. additions to the verbal Part-of-Speech (POS) tag set
  2. additions to the types of NP objects
  3. strategies for representing the position of clitic elements

Introduction to Penn Treebank-style Annotation

A Simple Sentence

The simple sentence ‘I saw the man’ is represented in Penn Treebank-style annotation as follows:

\[(1) \quad (IP-MAT \ (NP-SBJ \ (PRO \ I))) \]
\[
(VBD \ saw) \]
\[
(NP-OB1 \ (D \ the)) \]
\[
(N \ man))) \]

where...

- a pair of parentheses ( ) delineates each level
- each level contains two components:
  1. a label on the left (e.g., a phrase label, a POS label, etc.)
  2. content on the right (e.g., phrase(s), word(s), etc.)
Representing Discontinuities

Discontinuities are represented by means of placeholders—traces—in the structure that:

- show the origin of the displaced element
- indicate the connection between the displaced element and the trace by numerical co-indexation

Example: *T* Traces for “wh-” Movement

(3) (CP-QUE (WNP-1 (WPRO What)))
    (IP-SUB (NP-OB1 *T*-1)
        (VBD did)
        (NP-SBJ (PRO you))
        (VB see)))

Search over Linguistic Accuracy

- A final motivating principle behind Penn Treebank-style annotation that is important to understand is that the primary goal of the annotation is facilitation of automated search, not linguistically-accurate markup [12].
- A corollary: Labels used in the annotation system should not be taken as descriptive claims about the language but as atheoretical tools to aid in the automatic classification of sentences according to various patterns and properties.

Analytic Verbal Forms

- Penn Treebank-style annotation was originally designed for modern and historical English [11, 9, 7, 8], a language that expresses the verbal concepts of tense, mood, and voice in an analytic fashion, via combinations of distinct verbs—that is, one or more auxiliary verbs together with a main verb in participial form.
  - simple past: I wrote.
  - present progressive: I am writing.
  - present perfect passive: It has been written.
Synthetic Verbal Forms

- In contrast to languages like English, Ancient Greek expresses these verbal concepts within one synthetic verbal form, the main verb of the sentence.
  - *egrapsa* ‘I wrote’
  - *grafo* ‘I write/I am writing’
  - *gegraptai* ‘It has been written’

The “Dash” Tag Strategy

- “Dash” tags separated from the main verbal tag by a hyphen can be used to add information about different verbal features without exploding the number of verbal tags.
- Using this strategy reduces the number of distinct verbal POS tags for Ancient Greek to 27.

English Verbal POS Tags

There are just 7 verbal POS tags in the PennParsed Corpora of Historical English [12]:

- **VAG** = present participle
- **VAN** = passive participle
- **VB** = infinitive
- **VBD** = past
- **VBI** = imperative
- **VBN** = perfect participle
- **VBP** = present

Adopting the same strategy for Ancient Greek, using a single tag to represent each distinct tense, aspect, mood, voice, and finiteness combination, would require over 100 distinct tags.

Basic Verbal POS Tags for Ancient Greek

- **VBP** = primary sequence verb (includes present, future, and present perfect)
- **VBD** = secondary sequence verb (includes imperfect/past imperfective, aorist/past perfective, and pluperfect)
- **VBN** = infinitive
- **VBI** = imperative
- **VBS** = subjunctive
- **VBO** = optative
- **VPR** = participle
Adapting Penn Treebank-style Annotation for Ancient Greek Verbal POS Tags

The -P Extension

![Verbal POS Tags Diagram](image)

Figure 1: The 7 basic verbal POS tags plus their middle/passive voice extensions with -P.

Representing Case on Participles

- $\text{VPR} = \text{nominative participle}$
- $\text{VPR}$ = genitive participle
- $\text{VPRA} = \text{accusative participle}$
- $\text{VPRD} = \text{dative participle}$

Aspect, Tense, and Voice Tags

- $\text{-IMPF} = \text{imperfective}$
- $\text{-AOR} = \text{perfective}$
- $\text{-PRF} = \text{perfect}$
- $\text{-FUT} = \text{future}$
- $\text{-PASS} = \text{syntactic passive}$

Marking Syntactic Passives

- $\text{-P}$ marks verbal forms whose *morphology* is non-active.
- $\text{-PASS}$ marks verbal forms in a clause where the *syntax* involves the promotion of a typical object in an active construction to the subject of the sentence.

Syntactic passives can have morphological forms that are either ambiguous between middle and passive voice or unambiguously passive, but the converse is not true: there are verb forms that are unambiguously passive with respect to their morphology but that have active (intransitive) syntax, not passive syntax.
Middle/passive morphology, active syntax: VBPP-IMPF

(4) . . . hama de kithōn ekduomenō at.the.same.time but tunic taking.off
    sunekduetai kai tēn aidō take.off.with-3SG.PRS.mid/pass also the-ACC modesty-ACC
    gunē woman-NOM
    ‘...but at the same time as she removes her tunic, a woman
    dispenses with her modesty too.’ (Hdt. 1.8.3)

Middle/passive morphology, passive syntax: VBPP-PRF-PASS

(5) . . . hutōs gar geγrapτai dia tu profētū... thus for write-3SG.PRF.mid/pass through the prophet
    ‘...for thus it has been written through the prophet...’
    (Matthew 2.6)

Passive morphology, intransitive syntax: VBDP-AOR

(6) . . . angelos Kuriu kat’ onar efanē autō... angel lord-GEN in dream appear-3SG.AOR.pass to.him
    ‘...an angel of the Lord appeared to him in a dream...’
    (Matthew 1.20)

Comparison with Tyndale Bible: Agreement

(7) . . . heurethē en gastri exusa ek pneumatōs hagiu. find-3SG.AOR.pass in stomach having from spirit holy
    ‘...[Mary] was found to be pregnant by the holy spirit.’
    (Matthew 1.18)

(8) Early Modern English: . . . she was foude-3SG.pass with chylde by ye holy goost. (Tyndale Matthew 1.18 [10])
Comparison with Tyndale Bible: Disagreement

(9) τοῦτο δὲ θέλει γέγονεν ἵνα
 this but all happen-3SG.PR.F.IND. act that
πληροθέ ὑπὸ Κυρίου διὰ τὸ
 to rhēthen hupo Kuriou dia tu fulfill-3SG.AOR.SBJV. pass the thing-spoken by God through the profētē... prophet
‘All this has happened in order that it might be fulfilled what was spoken by God through the prophet...’ (Matthew 1.22)

(10) Early Modern English: All this was done-
3sg. pass to fulfill-INF. act yt which was spoken of the Lorde by the Prophet...’
(Tyndale Matthew 1.22 [10])

Two Basic Noun Phrase Objects Tags

Penn Treebank-style annotation includes two basic tags for distinguishing between types of noun phrase objects:
▶ NP-OB1 for direct objects  
‘I gave John the book.’
▶ NP-OB2 for indirect objects  
‘I gave John the book.’

Additional Object Types in Ancient Greek

Ancient Greek has (at least) two additional types of objects:
▶ objects that appear in a “quirky” case
▶ objects that derive their case from a prepositional prefix on the verb

NP-OBQ for Objects in a Quirky Case

The Ancient Greek verb μιμνῄσκω ‘remember’ takes a genitive object [13, §1356], as do compounds built from this verb:

(11) ἐπιμνῄσκομαι ἁμορτερὸν homoīos.  
mention-1SG.FUT.MID both-gen alike
‘...[I] will mention both alike.’ (Hdt. 1.5.4)
NP-OBP for Objects of a Prepositional Prefix

The Ancient Greek verb *sunanakēmai* ‘sit down with’ takes a dative object, just as the preposition *sun* ‘with’ does [13, §1545]:

(12) kai idu polloi telēnai kai hamartōloi elthontes and behold many tax.collectors and sinners having.come
sunanakēnto tō iēsu kai tois sit.down.with-3PL.IMPF.MID the-dat Jesus-dat and the-dat
mathētais autu disciples-dat his
‘And behold, many tax collectors and sinners, having come, sat down with Jesus and his disciples.’ (Matthew 9.10)

What are clitics? I

Clitics are:

- prosodically weak (unstressed) elements whose position in a clause is highly constrained
- form a unit with some neighboring word on the right or left, but they can’t be considered affixes because they also exhibit syntactic independence

What are clitics? II

| Morpheme attaches to head noun | Plural | Possessive
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>[The [boy]NP I met]NP waved to me.</strong></td>
<td><em>The [boy’s]NP I met]NP bike.</em></td>
<td></td>
</tr>
</tbody>
</table>
| **Morpheme attaches at phrase edge** | Plural | Possessive
| *The [boy]NP I met]NP’s waved to me.* | * |

Table 1: The distribution of clitics vs. affixes in English (* indicates that the sentence is ungrammatical)

Clitics in Ancient Greek

Ancient Greek clitics can be divided into (roughly) two groups based on their behavior:

- clitic particles
- clitic pronouns and verbs
Adapting Penn Treebank-style Annotation for Ancient Greek

Clitic Position

Special POS Tags for Clitic Particles

- CLTE for the conjunctive clitic particle *te*
- CLGE for the emphatic clitic particle *ge*
- CLPRT for all other clitic particles

Intervening Clitic Pronouns and Verbs

(14) nun de *amfoterôn me* *tutôn* apokleîsas echês...
‘But now you have barred me from both of these.’ (Hdt. 1.37.2)

(15) *kurious* gar *estin tu* *sabbatu* ho huios tu anthrôpu
‘For the Son of Man is Lord of the Sabbath.’ (Matthew 12.8)

CLPRT Example

(13) ( (IP-MAT (NP-1 (DS$ ton)
   (CLPRT de)
   (Q$ amfoteron))
   (PP (P es)
   (NP (DA+ADJA touto)))
   (NP-SBJ (NP-ATR *ICH*-1)
   (DS hai)
   (NS gnomai))
   (VBD-AOR sunedramon)
   (, ,))
   (ID Herodotus,Histories.489))

CLPROA Example

Two additions to the Penn Treebank annotation system are employed to represent the proper hierarchical position of clitic elements in a clause:

- a -CL dash tag
- a distinct *CL* trace

(16) ( (IP-MAT-SPE (ADVP-TMP (ADV nun))
   (CLPRT de)
   (NP-OB1 *CL*-1)
   (NP-OBP (Q$ amfoteron)
   (NP-CL-1 (CLPROA me))
   (DS$ touton))
   (VPR-AOR apokleîsas)
   (NP-SBJ *pro*)
   (VBP-IMPF echês)
   (, ,))
   (ID Herodotus,Histories.370))
Acknowledgements I

I would like to thank the following groups and individuals for helpful comments and input both on the construction of a parsed corpus of Ancient Greek and on earlier drafts of this paper: Beatrice Santorini, Tony Kroch, Caitlin Light, Joel Wallenberg, Aaron Ecay, Anton Ingason, Akiva Bacovcin, Constantine Lignos, and the Treebanks Lab at the University of Pennsylvania. All errors remain my own.

Acknowledgements II

Although my parsed texts of the Greek New Testament and Herodotus' *Histories* have not yet been released to the public (although an alpha version of the GNT is available by request), I would also like to acknowledge the various open-source resources that I have used in the construction of these parsed texts:

- Creative Commons-licensed texts from the Perseus Digital Library at Tufts University [3].
- Morphological information extracted from the PROIEL dependency treebank of the Greek New Testament [6].
- Morphological information from the Perseus under PhiloLogic project at the University of Chicago [2].
- Morphological information extracted from James Tauber's MorphGNT.

Links

- My website: [http://www.ling.upenn.edu/~janabeck/](http://www.ling.upenn.edu/~janabeck/)
- The annotation manual for my parsed corpus of Ancient Greek: [http://www.ling.upenn.edu/~janabeck/PPCHiGAnnotation_Manual.xhtml](http://www.ling.upenn.edu/~janabeck/PPCHiGAnnotation_Manual.xhtml)
- My academic blog, where updates on corpus construction are posted: [http://greekings.wordpress.com/](http://greekings.wordpress.com/)
- My GitHub repository for my parsed corpora of Ancient Greek: [https://github.com/janabeck/PPCHiG](https://github.com/janabeck/PPCHiG)
- The website for CorpusSearch 2, the software used to search Penn Treebank-style parsed historical corpora: [http://corpussearch.sourceforge.net/CS.html](http://corpussearch.sourceforge.net/CS.html)

References I

Adapting Penn Treebank-style Annotation for Ancient Greek

References II


References III


Phrase-Structure vs. Dependency Annotation I

- The main focus of dependency annotation is to indicate the relation that each word bears to the word that it is dependent on (see e.g., [4, 1, 5]).
- For example, the abbreviation atr is used to indicate that a word is in an attributive relationship with the word it depends on, sbj indicates a subject, obj an object, etc.
- Theses relations are indicated by subscript text in italics in the example on the next slide.

Phrase-Structure vs. Dependency Annotation II

- Dependency annotation only indirectly represents word order, usually via ordered identification numbers for the words (in this example indicated with superscripts).
- The structural graphs it produces do not necessarily preserve the order of words in a sentence:

(17) read pred
    1
    2
    3
    4
    5
    6
   book subj
   the
   wrote
   which
   you

Chicago Colloquium on Digital Humanities and Computer Science, 19-21 Nov. 2011 37/42
Phrase-Structure vs. Dependency Annotation III

In contrast, the graph produced via phrase-structure annotation always preserves the order of the words in the sentence as well as representing functional relationships between units and sub-units (phrases and words) in the sentence.

(18)

```
IP-MAT
  /\    \/
 NP-SBJ  VBD  NP-OB1
    /\    |    |    |
   PRO  saw D    N
       |    |    |
       I    the man
```

Phrase-Structure Annotation and Theory-Neutrality

- By its nature, phrase-structure annotation is less theory-neutral than dependency annotation since some choices must be made as to what types of phrases exist for grouping words together hierarchically.
- In the Penn Treebank style of phrase-structure annotation, an effort is made to keep the phrase structure as minimal as possible—resulting in somewhat “flat” trees compared to modern syntactic theories—to reduce the number of controversial decisions about phrase boundaries that are necessary.