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

A Simple Sentence

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

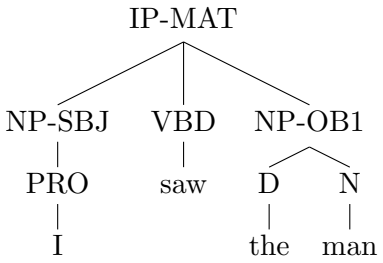
(1)
```
(IP-MAT (NP-SBJ (PRO I))
        (VBD saw)
        (NP-OB1 (D the)
                (N man)))
```

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.)

Graphical Representation of Penn Treebank Trees

(2)



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))          << displaced element
               (C 0)
        (IP-SUB (NP-OB1 *T*-1)       << co-indexed trace
                (VBD did)               indicating functional
                (NP-SBJ (PRO you))      position
                (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'

English Verbal POS Tags

There are just 7 verbal POS tags in the Penn Parsed 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.

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.

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

The -P Extension

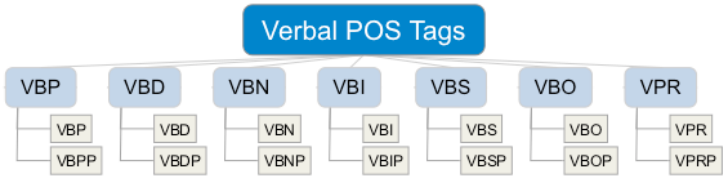


Figure 1: The 7 basic verbal POS tags plus their middle/passive voice extensions with -P.

Representing Case on Participles

- `VPR` = nominative participle
- `VPR$` = genitive participle
- `VPRA` = accusative participle
- `VPRD` = dative participle

Aspect, Tense, and Voice Tags

- `-IMPF` = imperfective
- `-AOR` = perfective
- `-PRF` = perfect
- `-FUT` = future
- `-PASS` = *syntactic* passive

Marking Syntactic Passives

- `-P` marks verbal forms whose *morphology* is non-active.
- `-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 kithōni 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 *gegraptai* dia tu profētu. . .
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 pneumatos 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) tuto de holon *gegonen* hina
this but all happen-3SG.PRF.IND.**act** that
plērōthē to rhēthen hupo Kuriu dia tu
fulfill-3SG.AOR.SBJV.**pass** the thing-spoken by God through the
profētu...
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 *mimnēskō* 'remember' takes a genitive object [13, §1356], as do compounds built from this verb:

(11) ...epimnēsomai amfoterōn homoiōs.
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 telōnai kai hamartōloi elthontes
and behold many tax.collectors and sinners having.come
*sun*anekēnto 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	[The [boy**s**]$_N$ I met]$_{NP}$ waved to me.
Possessive	*[The [boy**'s**]$_N$ I met]$_{NP}$ bike...
	Morpheme attaches at phrase edge
Plural	*[The [boy]$_N$ I met]$_{NP}$**s** waved to me.
Possessive	[The [boy]$_N$ I met]$_{NP}$**'s** bike...

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

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

`CLPRT` Example

(13)
```
( (IP-MAT (NP-1 (DS$ ton)
                (CLPRT de)        << intervening clitic particle
                (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))
```


Intervening Clitic Pronouns and Verbs

(14) nun de **amfoterōn** *me* **tutōn** apoklēisas echēs…
now but both-GEN me-ACC these-GEN barred have
'But now you have barred me from both of these…' (Hdt. 1.37.2)

(15) **kurios** gar *estin* **tu** **sabbatu** ho huios tu anthrōpu
Lord for is the-GEN Sabbath-GEN the son the-GEN man-GEN
'For the Son of Man is Lord of the Sabbath.' (Matthew 12.8)

`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)               << trace of clitic
              (NP-OBP (Q$ amfoteron)           pronoun
                      (NP-CL-1 (CLPROA me)) << intervening clitic
                      (DS$ touton))            pronoun
              (VPR-AOR apokleisas)
              (NP-SBJ *pro*)
              (VBP-IMPF eches)
              (, ,))
 (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/`
- The annotation manual for my parsed corpus of Ancient Greek: `http://www.ling.upenn.edu/~janabeck/PPCHiG_Annotation_Manual.xhtml`
- My academic blog, where updates on corpus construction are posted: `http://greekings.wordpress.com/`
- My GitHub repository for my parsed corpora of Ancient Greek: `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`

References I

[1] David Bamman and Gregory Crane.
Guidelines for the syntactic annotation of the Ancient Greek dependency treebank.
Technical report, The Perseus Project, Tufts University, 2008.

[2] Helma Dik and the ARTFL Project at the University of Chicago.
Perseus under PhiloLogic.
Online resource, 2011.

[3] Gregory Crane (editor-in chief).
Perseus Digital Library.
Online resource, 2011.

[4] Jan Hajič.
Building a syntactically annotated corpus: The Prague dependency treebank.
In Eva Hajičová, editor, *Issues of Valency and Meaning: Studies in Honor of Jarmila Panevová*, pages 106–132. Karolinum, Prague, 1998.

[5] Dag Trygve Truslew Haug.
PROIEL guidelines for annotation.
Online, June 2010.

[6] Dag Trygve Truslew Haug.
Pragmatic Resources in Old Indo-European Languages (PROIEL).
Online resource, 2011.

References II

[7] Anthony Kroch, Beatrice Santorini, and Lauren Delfs.
The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME).
Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, 2004.
URL: http://www.ling.upenn.edu/hist-corpora/.

[8] Anthony Kroch, Beatrice Santorini, and Ariel Diertani.
The Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE).
Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, 2010.
URL: http://www.ling.upenn.edu/hist-corpora/.

[9] Anthony Kroch and Ann Taylor.
The Penn-Helsinki Parsed Corpus of Middle English (PPCME2).
Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, 2000.
URL: http://www.ling.upenn.edu/hist-corpora/.

[10] Caitlin Light.
Excerpts from the Tyndale New Testament.
Unpublished parsed corpus, 2011.

[11] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz.
Building a large annotated corpus of English: the Penn treebank.
In Susan Armstrong, editor, *Using Large Corpora*, pages 273–290. MIT Press, Cambridge, 1994.

References III

[12] Beatrice Santorini.
Annotation manual for the Penn historical corpora and the PCEEC.
Online, 2006.

[13] Herbert Weir Smyth.
Greek Grammar.
Harvard University Press, 1956.
Revised by Gordon M. Messing.

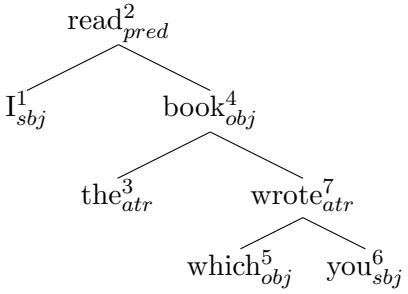
Phrase-Structure vs. Dependency Annotation I

- 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:

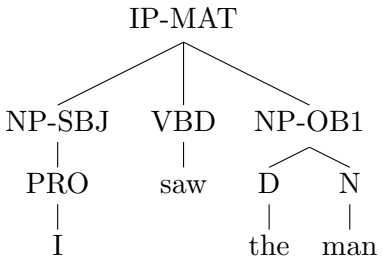
(17)



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)



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.