Adapting Penn Treebank-style annotation for Ancient Greek: The Benefits and Challenges

Jana E. Beck

October 21, 2011

Abstract

This paper introduces a work in progress: a syntactically-annotated corpus of historical Greek in the Penn Treebank style built on the Perseus Digital Library's Creative Commons-licensed texts of the Greek New Testament and Herodotus, with more texts to follow. The main purpose of this corpus is to facilitate research on syntactic variation within specific time periods of Greek (e.g., Classical Greek), as well as syntactic change over the course of the history of Greek. In this paper, I discuss adaptations of the Penn Treebank phrase-structure annotation style that I have implemented in order to encode information about the morphology and syntax of Greek that is necessary to conduct research on syntactic variation and change.

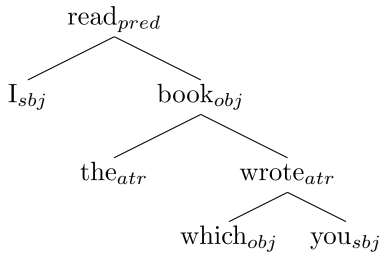
1 Introduction

Following the construction of the Prague Dependency Treebank (Hajič, 1998), it has generally been argued that a dependency annotation system that encodes functional information about the relationships between words in a sentence without directly encoding intermediate-level syntactic constituency is the best system for building a treebank for a language that has "free word order." What is meant by "free word order" in these cases is an abundance of discontinuous phrases:

> [Dependency grammar] is an especially appropriate manner of representation for languages with a moderately free word order (such as Greek, Latin and Czech), where the linear order of constituents is broken up with elements of other constituents. (Bamman and Crane, 2008)

A dependency graph of the sentence "I read the book which you wrote," with the relation each dependent bears to its parent node indicated with a subscripted tag, is shown in (1)[1].

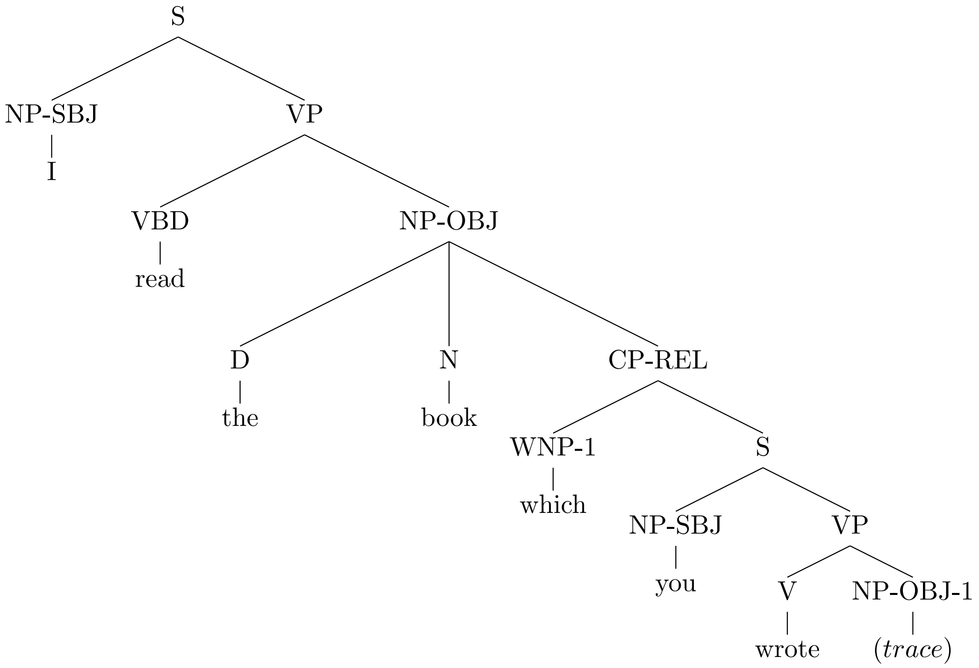
(1)



[1] The *atr* relation stands for attributive and in this example applies to both the relative clause 'which you wrote' as a whole since its function is to further describe 'the book' and to the article 'the.'

This dependency graph does not preserve the order of the words in the sentence but only shows their hierarchical relationships. In contrast, a phrase-structure tree (2)[2] of the same sentence both preserves the order of the words in the sentence and indicates the hierarchical relationships between words.

(2)



Since dependency annotation does not directly encode the relationship between word order and syntactic function, it is perhaps not the ideal annotation for use in linguistic research that is concerned with phenomena relating to word order. Thus, in my project of building a parsed corpus of Ancient Greek, I have chosen to use a phrase-structure annotation of the same basic type as the Penn Treebank (Marcus et al., 1994) and the Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000; Kroch et al., 2004, 2010). In this paper, I discuss the modifications of the basic phrase-structure annotation system found in the Penn-style corpora that I have implemented in order to represent various aspects of Ancient Greek that are different from modern and historical English in the first book of the *Histories* of Herodotus and the book of Matthew in the Greek New Testament, the two samples from my parsed corpus that I used to test my annotation system. I emphasize the utility of the various annotation choices I have made for answering questions of particular interest to scholars of Ancient Greek.

There are three central areas in which I have modified the Penn Treebank-style annotation in order to better suit Ancient Greek. Proceeding from the word-level to the sentence-level, these are verbal part-of-speech (POS) tags, types of NP objects, and the position of clitics. In section 2, I briefly introduce the Penn Treebank style of annotation, which is a type of phrase-structure (as opposed to dependency) annotation. In section 3, I give an overview of the differences between the verbal systems of English and Ancient Greek and present my solutions for representing aspect and tense as well as the complex Ancient Greek voice system and the “sequence of moods” phenomenon. In section 4, I argue that a more finely grained set of function tags is required to distinguish between the different types of noun phrase objects in Greek. Finally, section 5 addresses the question of how to represent clitic position in a phrase-structure annotation system. Section 6 concludes.

[2] The (*trace*) element used in this example to indicate the dual role of the wh- pronoun ‘which’ will be discussed in more detail in section 2 below.

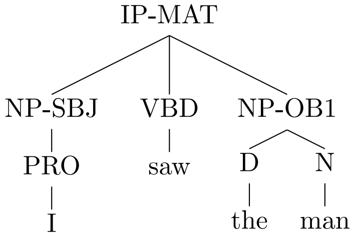
2 Introduction to Penn Treebank-style Annotation

Penn Treebank-style annotation operates on two basic principles: it arranges words hierarchically in terms of their syntactic relationships, and it preserves the original ordering of words[3]. For example, the simple sentence ‘I saw the man’ would be represented as follows in (3), where a pair of parentheses delineates each level, and each level contains two components: a label on the left (a phrase label, a part-of-speech (POS) tag, etc.) and content on the right (phrase(s), a word, etc.). The top-level label in this example, IP-MAT, stands for an inflectional phrase at the matrix or main clause level—in more traditional grammatical terms, this corresponds to an independent clause or sentence. Likewise, IP-SUB stands for inflectional phrase, subordinate or, in other words, a dependent clause.

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

The Penn Treebank-style tree in (3) is equivalent to the graphical tree representation in (4).

(4)



Discontinuities are represented by means of placeholders—traces—in the structure showing the origin of a displaced[4] element and indicating its relationship to the displaced element via numerical co-indexation. There are two types of traces used in the Penn Treebank-style annotation system that will be relevant for my discussion of Ancient Greek in this paper. First, traces of the form `*T*` are used to represent the displacement of question words to the left edge of the sentence in languages such as English and Ancient Greek. For example, in (5), the wh- noun phrase ‘what’ is co-indexed with a trace labeled `NP-OB1`, showing that this wh- phrase is to be interpreted as the direct object (i.e., the “first” object, hence `-OB1`) of the verb ‘see.’

```
(5)  (CP-QUE (WNP-1 (WPRO What))          << displaced element
                    (C 0)
             (IP-SUB (VBD did)
                     (NP-SBJ (PRO you))
                     (VB see)
                     (NP-OB1 *T*-1)))     << co-indexed trace indicating functional position
```

The second type of trace is essentially a “catch-all” category to represent any kind of displacement not described by the usage of the `*T*` trace or the bare `*` trace (not described above because it rarely occurs in Ancient Greek). The third trace is designated with `*ICH*`, which is mnemonic for “interpret constituent here.” In the English corpora annotated in the Penn Treebank-style, the `*ICH*` is used for phenomena such as the extraposition of prepositional phrases (Santorini, 2006):

[3] For those readers familiar with computer programming languages, the canonical format for Penn Treebank-style annotation is modeled on LISP.

[4] Many syntactic theories describe the displacements discussed here as movement operations. I use the term displacement to remain more theory-neutral.

(6)
```
(IP-MAT (NP-SBJ (PRO We))
        (VBD ate)
        (NP-OB1 (ADJR better) (N pumpkin) (N pie)
                (PP *ICH*-1))     << trace of extraposed PP
        (ADVP-TMP (ADV yesterday))
        (PP-1 (P than)            << extraposed PP
              (CP-CMP (C 0)
                      (IP-SUB (NP-SBJ (PRO we))
                              (ADVP-TMP (ADV ever))
                              (HVD had)
                              (VBN *)
                              (PP (P in)
                                  (NP (PRO$ our) (NS lives))))))
        (. .))
```

Although Penn Treebank-style annotation is hierarchical, it is not strictly binary branching. For example, in example (3) above, there is no verb phrase (VP) boundary marked although such a boundary is predicted in most (if not all) modern generative syntactic theories. The basic philosophy of Penn Treebank-style annotation is centered around two related principles: the annotation system is designed to maximize consistency while at the same time reducing as much as possible controversial or extremely subjective annotation decisions. Accordingly, VP boundaries are systematically unmarked since they are very difficult to diagnose consistently.

Finally, there is a third motivating principle behind Penn Treebank-style annotation that is important to understand. The primary goal of Penn Treebank-style corpora is the facilitation of automated search[5], not linguistically-accurate markup.

> Our primary goal has been to create an annotation system that facilitates automated searches, not to give a correct linguistic analysis of each sentence. For instance, if a construction can be found unambiguously through a combination of properties of a bracketed sentence, our annotation may not contain all of the structure that a full phrase structure diagram of the sentence would have. (Santorini, 2006)

A corollary to this is that the labels used in the annotation system should not be taken as descriptive claims about the language under consideration. Rather, the labels should be viewed as atheoretical tools to aid in the automatic classification of sentences according to their syntactic features. For example, the construction shown in (7a) in Ancient Greek is annotated as if the infinitival clause is the subject of the sentence (shown in (7b) below) since Ancient Greek, unlike English, does not show an overt expletive subject—i.e., the 'It' in 'It is not appropriate...' in the English translation.

(7) a. ho de apokrithēs ēpen uk estin kalon labēn ton arton tōn teknōn kai balēn
he but answering said not is good to.take the bread the-GEN children-GEN and to.throw
tois kunariois
the-DAT dogs-DAT
'But he answered, "It is not appropriate to take the children's bread and throw it to the dogs."'
(Matthew 15.26)

b.
```
( (IP-MAT (NP-SBJ (D ho))
          (CLPRT de)
          (IP-PPL (VPRP-AOR apokrithes))
          (VBD-AOR epen)
          (IP-MAT-SPE (NEG uk)
                      (BEP-IMPF estin)
                      (ADJP-PRD (ADJ kalon))
```

[5] In particular using the search functions of the open source software CorpusSearch2.

```
                    (IP-INF-SBJ (IP-INF-1 (VBN-AOR laben)
                                          (NP-OB1 (DA ton)
                                                  (NA arton)
                                                  (NP-ATR (DS$ ton) (NS$ teknon))))
                                (CONJP (CONJ kai)
                                       (IP-INF=1 (VBN-AOR balen)
                                                 (NP-OB2 (DSD tois) (NSD kunariois))))))
       (. .))
(ID GreekNT,Matthew.855))
```

It has been argued for other languages (e.g., German, cf. Safir (1985), among others) that a null (unpronounced) expletive subject occurs in the same syntactic position as does the overt expletive in English, and of course the same may be true of Ancient Greek. However, for the sake of simplicity in the annotation system—that is, to avoid introducing an additional null category `(NP-SBJ *exp*)` for null expletive subjects—I annotate such examples as containing infinitival subjects.

3 Verbal POS Tags

English expresses tense and aspect in its verbal forms for the most part via auxiliary verbs in combination with different participial forms of the main verb.

- simple past: I wrote.
- present progressive: I am writing.
- present perfect: I have written.

In contrast, Ancient Greek expresses all differences of tense and aspect within the morphological form of the main verb itself.

- *egrapsa* 'I wrote'
- *grafo* 'I write/I am writing'
- *gegrammai* 'I have written'

Thus, while the POS tags in the Penn Treebank-style Penn Parsed Corpora of Historical English are limited to just seven tags for main verbs (VAG = present participle, VAN = passive participle, VB = infinitive, VBD = past, VBI = imperative, VBN = perfect participle, VBP = present), the combinatorial possibilities in Greek would number well over 100 if I were to create a separate tag for each tense, aspect, mood, and finiteness combination. Instead of creating many such tags, I have chosen to add information to verbal forms via "dash" tags. Using this strategy, I have managed to keep the number of verbal tags in my corpus of Ancient Greek relatively small—there are only 27 tags—without suffering loss of information.

The verbal tags in my corpus of Ancient Greek are based on seven basic tags:

- 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

This basic tag set is extended into 14 tags by the addition of a P on all middle or passive voice verbs (see subsection 3.1 below for more on the three voice distinctions made in Ancient Greek)—thus, VBPP, VBDP, VBNP, VBIP, VBSP, VBOP, VPRP (see Figure 1 below).

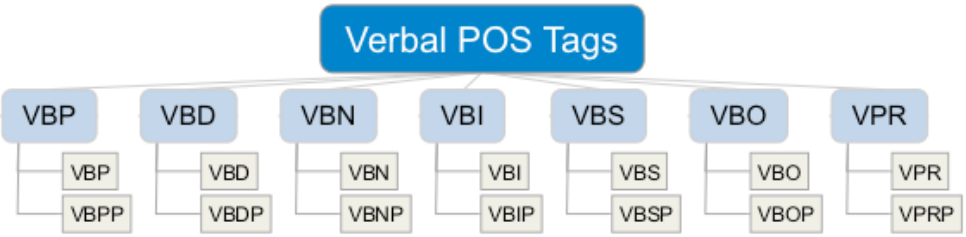


Figure 1: The 7 basic verbal POS tags plus their middle/passive voice extensions with -P.

In addition, the participle tags—active or passive—are extended with $ for genitives[6], A for accusatives, and D for datives. So, for example, the complete basic tag set for active participles is VPR, VPR$, VPRA, VPRD, resulting in a running total of 20 verbal tags. Additionally, all verbal forms can be extended with one or more verbal "dash" tags representing certain aspect, tense, and voice distinctions: -IMPF (imperfective), -AOR (perfective), -PRF (perfect), -FUT (future), and -PASS (on *syntactic* passives, see subsection 3.1 below).

Finally, two special dash tags occur only on optatives to mark optatives that are "standing in" for either indicatives or subjunctives due to the "sequence of moods" phenomenon in Ancient Greek (Smyth, 1956, §2615), illustrated in (9) in comparison to the roughly similar English "sequence of tense" phenomenon illustrated in (8).

(8) a. John says he will come to the party.
b. John said he would come to the party.
(subordinate verb changes from 'will' to 'would' under the matrix past tense verb 'said')

(9) a. ...*egnōsan* hoi stratiōtai hoti kenos ho fobos **ēē**...
realize-3PL.AOR.IND the soldiers that groundless the fear be-3SG.IMPF.OPT
'...the soldiers realized that their fear was groundless...' (Xen. Anab. 2.2.21)
b. hōs *ēpen* ho Saturos hoti **oimōksoito**, ē mē **siōpēsēen**...
when say-3SG.AOR.IND the Satyrus that suffer-3SG.FUT.OPT, if not keep.quiet-3SG.AOR.OPT
'When Satyrus said that he would suffer if he did not keep quiet...' (Xen. Hell. 2.3.56)

The -IND dash tag occurs on optatives (indicated with bold in (9) above) that are standing in for indicative verbs when the matrix verb of the clause (indicated with italics in (9) above) is a secondary sequence verb. The -KJV[7] dash tag occurs on optatives that are standing in for subjunctive verbs under a secondary sequence matrix verb.

3.1 The -PASS Dash Tag

The -PASS dash tag is used to indicate syntactic passives exclusively, where a "syntactic passive" is defined as a construction involving the promotion of a typical object in an active construction to the subject of the

[6] $ is used to agree with the standard in the Penn Parsed Corpora of Historical English.

[7] Mnemonic for the German word for subjunctive—Konjunktiv—to avoid confusion with the -SBJ dash tag for subjects.

sentence. The voice system in Ancient Greek is more complex than in many other languages, with a threefold division between active, middle[8], and passive voice. Syntactic passives can have morphological forms that are either ambiguous between middle and passive voice (10a, 10b) 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. For example, the aorist "passive" *efanē* of the verb *fainō* 'cause to appear' has the simple intransitive meaning '(he/she/it) appeared' (10c).

(10) a. middle/passive morphology, active meaning:
. . . hama de kithōni ekduomenō *sunekduetai* kai tēn
at.the.same.time but tunic taking.off take.off.with-3SG.PRS.**mid/pass** also the-ACC
aidō gunē
modesty-ACC woman-NOM
'. . . but at the same time as she removes her tunic, a woman dispenses with her modesty too.' (Hdt. 1.8.3)

b. middle/passive morphology, passive meaning:
. . . 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)

c. passive morphology, intransitive meaning:
. . . 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)

The benefit of distinguishing between morphological and syntactic passives in a corpus of Ancient Greek—and particularly in the syntactically parsed Greek New Testament—is that this will allow for comparison between the Greek text and translations of the text with respect to how often the active versus passive syntax of the original is implemented in the translation.

For example, in the first chapter of Matthew, comparing the Greek New Testament to the Middle English translation of the New Testament by William Tyndale (Light, 2011), there are some examples where a passive in the English corresponds to a passive in the Greek (11) and some where it does not (12).

(11) a. . . . *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)

b. . . . she *was foude*-3SG.**pass** with chylde by ye holy goost (Tyndale Matthew 1.18)

(12) a. tuto de holon *gegonen* hina *plērōthē* to rhēthen hupo
this but all happen-3SG.PRF.IND.**act** that fulfill-3SG.AOR.SBJV.**pass** the thing-spoken by
Kuriu dia tu profētu. . .
God through the prophet
'All this has happened in order that it might be fulfilled what was spoken by God through the prophet. . . ' (Matthew 1.22)

b. All this *was done*-3SG.**pass** *to fulfill*-INF.**act** yt which was spoken of the Lorde by the Prophet. . . ' (Tyndale Matthew 1.22)

[8] The semantic contribution of the middle voice versus the active or the passive is variable: the middle voice can mark that the subject of the verb has some stake in the action described (most often benefiting from it), reflexivity, causativity, etc.

4 Types of Noun Phrase Objects

In the Penn Treebank-style corpora of English, there are only two function tags to distinguish between direct (`NP-OB1`) and indirect (`NP-OB2`) noun phrase objects of verbs. The syntax of Greek, however, includes at least two more types of objects that are worth distinguishing from direct and indirect objects. First, there are objects that occur in a "quirky" case—essentially, these are direct objects that appear in the genitive or dative case instead of the accusative. For example, the Ancient Greek verb *mimnēskō* 'remember' takes a genitive object (Smyth, 1956, §1356), as do compounds built from this verb:

(13) . . . epimnēsomai amfoterōn homoiōs.
mention-1SG.FUT.MID both-**gen** alike
'. . . [I] will mention both alike.' (Hdt. 1.5.4)

Second, there are objects that derive their case from a prepositional prefix on the verb itself. For example, the verb *sunanakēmai* 'sit down with' takes a dative object, just as the preposition *sun* 'with' does (Smyth, 1956, §1545):

(14) kai idu polloi telōnai kai hamartōloi elthontes *sun*anekēnto tō
and behold many tax.collectors and sinners having.come sit.down.*with*-3PL.IMPF.MID the-**dat**
iēsu kai tois mathētais autu
Jesus-**dat** and the-**dat** disciples-**dat** his
'And behold, many tax collectors and sinners, having come, sat down with Jesus and his disciples.' (Matthew 9.10)

In order to represent these two additional types of objects in my corpus of Ancient Greek, I have added two additional NP function tags: `-OBQ` for objects in a quirky case and `-OBP` for objects deriving their case from a prepositional prefix on the verb. These tags, in combination with the lemmatization of the corpus, will make it possible for scholars studying linguistic change in Ancient Greek to quantify the variation in the types of objects that occur with certain verbs.

5 Clitic Position

Ancient Greek contains clitic pronouns, particles, and even verbs—these are prosodically weak (unstressed) elements whose position in a clause is highly constrained. Clitics form a phonological unit with some neighboring word on the right or left, but they can't be considered affixes because they also exhibit syntactic independence. An example of a clitic from English is the possessive *'s*. That possessive *'s* is a clitic is best shown by the positions it occurs in. The possessive *'s* always attaches to the edge of an entire noun phrase—a syntactic distribution—in contrast to the English plural affix *-s*, which attaches to the *head* of a noun phrase:

	Morpheme attaches to head noun	**Morpheme attaches at phrase edge**
Plural	[The [boy**s**]$_N$ I met]$_{NP}$ waved to me.	*[The [boy]$_N$ I met]$_{NP}$**s** waved to me.
Possessive	*[The [boy**'s**]$_N$ I met]$_{NP}$ bike. . .	[The [boy]$_{NP}$ I met]$_{NP}$**'s** bike. . .

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

My strategy towards representing the position of clitics varies depending on the type of clitic. For example, the most common type of clitic—the sentential particles *de*, *gar*, etc. serving as a sort of connective "glue" between sentences—are given their own POS tag: CLPRT. Other clitic particles are given POS tags starting with CL so that all (non-verbal) clitics can be found by searching for all tags beginning with CL. For example, the emphatic clitic particle *ge* bears the POS tag CLGE. A more complicated case involves the clitic forms of the verbs *eimi* 'to be' and *fēmi* 'say.' These are given an additional -CL dash tag on their verbal POS tag, and in addition, when a clitic verbal form breaks up a phrase, a special type of trace is employed to show the clitic's true hierarchical position as the main verb of the clause instead of as a daughter of the phrase it breaks up. These various strategies for representing clitic position in the corpus are designed to keep the corpus as human-readable as possible while still including as much information as is necessary for linguistic research on the syntax of Greek clitics to be carried out.

Even within syntactic theories that represent displacement as movement, the displacement of clitics is regarded as a different process from other types of syntactic displacement (such as the so-called "wh-movement" operation that displaces question words such as 'what' to the left edge of the sentence in languages like English) and has been argued to occur at a later stage in the derivation of a sentence than syntactic displacement operations—namely, during the process that takes the output of syntactic operations as its input, linearizes this input, and maps it onto a phonological form (Embick and Noyer, 1999, 2001; Embick, 2003). In this section I argue that because clitic displacement is a different type of phenomenon from the other displacement operations, it makes sense to represent it differently from other displacement operations—either via POS tags alone or via a new kind of trace—in the syntactic annotation scheme for my corpus of Ancient Greek.

5.1 Second-position sentential particles

First, the very common second-position sentential particles (see (15) with the particle *de* 'but' represented in italics, as all clitic elements in this section are) are represented with the special POS tag CLPRT alone (16).

(15)	**tōn**	*de*	**mantēiōn**	**amfoterōn**	es	tōuto	hai	gnōmai	sunedramon
	the	but	oracles	both	to	the.same	the	judgments	concurred

'But the judgments of both the oracles came to the same thing.' (Hdt. 1.53.3)

No trace of displacement is employed in order to facilitate readability, since these clitics occur in almost every sentence and since the trace would in almost every case be in the same position, as the first element in the clause (cf. also the English translation of (15)—in particular, how the particle is translated as the first word in the English). In order to determine whether the clitic intervenes in a phrase or is located after the first phrase in the sentence, it is enough to query for whether or not the sentential particle is dominated by an `IP` or `CP` node or some other phrasal category and to count the words preceding the particle. For example, in (15) the sentential particle *de* 'but' is immediately dominated by an NP node, with one word, the article *tōn* 'the (gen. pl.)' preceding it; this is therefore an example where, as expected, *de* 'but' occurs in its canonical second position, and it intervenes in the noun phrase *tōn mantēiōn* 'the oracles (gen. pl.),' causing a discontinuity in that noun phrase.

(16)
```
( (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))
```

At present, the two other clitic particles in Ancient Greek—the clitic conjunction *te* 'and' and the emphatic clitic particle *ge*—are treated identically to the second-position sentential particles. That is, they receive dash tags beginning with CL—CLTE and CLGE, respectively—and no traces of displacement are employed to represent their position.

5.2 Clitic pronouns and verbs

Discontinuities also occur when the intervention of clitic pronouns and clitic forms of the verbs *eimi* 'be' and *fēmi* 'say'—results in a discontinuous phrase. For example, in (17a), the placement of the clitic pronoun *me* 'me' breaks up the phrase *amfoterōn tutōn* 'both these.' Similarly, in (17b), a clitic verb form breaks up a noun phrase.

(17) a. 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)

b. **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)

The clitic pronouns and verbal forms have a more complex distribution, and so I represent these with a dash tag -CL and a special type of trace *CL* at the proper hierarchical level where the clitic element is interpreted. There are no restrictions on whether this trace occurs to the left or to the right of the displaced element, but an effort is made to put the trace in the likely position where the clitic originated in the syntax given that clitics often appear in second position, inverted in linear order with either the first or the last element of some phrase. For example, (17a) is represented as in (18), under the hypothesis that the clitic pronoun *me* 'me' underwent linear inversion with *amfoteron* 'both.'

(18)
```
( (IP-MAT-SPE (ADVP-TMP (ADV nun))
              (CLPRT de)
              (NP-OB1 *CL*-1)              << trace of clitic pronoun
              (NP-OBP (Q$ amfoteron)
                      (NP-CL-1 (CLPROA me)) << intervening clitic pronoun
                      (DS$ touton))
              (VPR-AOR apokleisas)
              (NP-SBJ *pro*)
              (VBP-IMPF eches)
              (, ,))
  (ID Herodotus,Histories.370))
```

The redundant marking of the -CL dash tag and the *CL* trace is intentional, as this allows for marking verbal forms as clitic (there are no distinct POS tags in my annotation system for clitic verbal forms) so that both the clitic forms of verbs can be found by searching as well as the instances in which these clitic forms intervene in an otherwise continuous phrase, making the trace necessary.

6 Conclusion

In this paper, I have introduced the Penn Treebank style of phrase-structure annotation and described three main areas in which I modified this annotation system in planning and beginning the construction of a parsed corpus of historical Greek. First, I described changes to the set of verbal POS tags used to apply labels at the word level in the corpus. I argued that since Greek expresses concepts of tense and aspect synthetically within a single verb form, the set of verbal POS tags should be expanded to include seven basic tags, case suffixes for participles, a -P suffix for verbs in the middle or passive voice, and a set of functional dash tags covering a wide variety of concepts, most importantly identifying syntactic passives and optatives representing indicatives or subjunctives in secondary sequence. Secondly, I argued that more than two types of objects—that is, more than just direct and indirect objects—should be identified in the Greek corpus via two additional "dash" tags. The tag `-OBQ` identifies objects that occur in a "quirky" case—essentially, these are direct objects that are not in the accusative case, but rather in the genitive or dative. The tag `-OBP` similarly identifies objects that derive their case from a prepositional prefix on the verb, taking the case that the preposition requires. Finally, I discussed various strategies for representing the position of clitic elements in a phrase-structure annotation system. These strategies ranged from just labeling the clitic elements with special POS tags in the case of second-position sentential particles, the conjunctive particle *te*, and the emphatic particle *ge*, to marking the interpreted position of a clitic with a co-indexed trace in the case of clitic pronouns and verbal forms.

References

Bamman, D. and G. Crane (2008). Guidelines for the syntactic annotation of the Ancient Greek dependency treebank. Technical report, The Perseus Project, Tufts University.

Embick, D. (2003). Linearization and local dislocation: Derivational mechanics and interactions. *Linguistic Analysis 33*(3-4), 303–336.

Embick, D. and R. Noyer (1999). Locality in post-syntactic operations. In *Papers in Morphology and Syntax, Cycle Two*, pp. 265–317. Cambridge: MIT Working Papers in Linguistics.

Embick, D. and R. Noyer (2001). Movement operations after syntax. *Linguistic Inquiry 32*(4), 555–595.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prague dependency treebank. In E. Hajičová (Ed.), *Issues of Valency and Meaning: Studies in Honor of Jarmila Panevová*, pp. 106–132. Prague: Karolinum.

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

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

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

Light, C. (2011). Excerpts from the Tyndale New Testament. Unpublished parsed corpus.

Marcus, M., B. Santorini, and M. A. Marcinkiewicz (1994). Building a large annotated corpus of English: the Penn treebank. In S. Armstrong (Ed.), *Using Large Corpora*, pp. 273–290. Cambridge: MIT Press.

Safir, K. (1985). Missing subjects in German. In J. Toman (Ed.), *Studies in German grammar*, Number 21 in Studies in generative grammar, pp. 193–229. Foris.

Santorini, B. (2006). Annotation manual for the Penn historical corpora and the PCEEC. Online.

Smyth, H. W. (1956). *Greek Grammar.* Harvard University Press. Revised by Gordon M. Messing.