The Ancient Greek and Latin Dependency Treebanks

David Bamman and Gregory Crane

Abstract This paper describes the development, composition, and several uses of the Ancient Greek and Latin Dependency Treebanks, large collections of Classical texts in which the syntactic, morphological and lexical information for each word is made explicit. To date, over 200 individuals from around the world have collaborated to annotate over 350,000 words, including the entirety of Homer’s Iliad and Odyssey, Sophocles’ Ajax, all of the extant works of Hesiod and Aeschylus, and selections from Caesar, Cicero, Jerome, Ovid, Petronius, Propertius, Sallust and Vergil. While perhaps the most straightforward value of such an annotated corpus for Classical philology is the morphosyntactic searching it makes possible, it also enables a large number of downstream tasks as well, such as inducing the syntactic behavior of lexemes and automatically identifying similar passages between texts.

Key words: treebanks, dependency grammar, digital libraries, Ancient Greek, Latin

1 Introduction

The definitive Classical reference grammars of the 19th and 20th centuries, such as Herbert Smyth’s Greek Grammar [44] and Raphael Kühner’s Ausführliche Grammatik der lateinischen Sprache [29], are monuments of scholarship that distill lifetimes of reading and linguistic observation into succinct aphorisms such as the following:

“Apodotic ὅτε is very common in Homer and Herodotus, not rare in Attic poetry, but infrequent in Attic prose.” (Smyth 2837).
On occasion these works offer a window into the traditional philological practice that lies behind them, as in Kühner’s comparison of the Latin *accusativus cum infinitivo* construction with subordinate clauses containing an overt complementizer (e.g., *quod*):

“So hat nach meiner Zählung bei *doleo* 57 Stellen mit Acc. c. Inf. gegen 4 *quod*, bei *miror* 110 gegen 8, bei *glorior* 19 gegen 2, bei *queror* 71 gegen 15, bei *gaudeo* 84 gegen 9 usw.”

(1914:77)

In its most basic form, classical philology of this sort is by definition a data-driven science: it relies on a fixed dataset (the extant corpus of Ancient Greek and Latin) and builds larger arguments by the simple act of counting. Kühner here publishes his tally of ACI vs. *quod*-clauses in order to advance the argument that the ACI is more frequent in indirect discourse than subordinate clauses are, and one can assume that either such an explicit tally or an implicit one (collected over a lifetime of reading) is what drives Smyth’s observations on relative frequency as well.

Where classical philology has so far diverged from data-driven science, however, is in its reliance on the authority of the editor rather than on the data itself. As much as the judgment of Kühner and Smyth may far exceed our own, the cornerstone of the scientific method is the reproducibility of experiments, and as P. Cuzzolin [20] notes about this very passage of Kühner:

“...it is difficult to say what he meant by the word “Stelle” and impossible to say which texts his counting is based upon.”

Ideally, what we want to see is the evidence that drives such linguistic observations – not simply knowing that the ACI is used in some unknown sample of 57 sentences containing *doleo*, but exactly which sentences those are, which textual editions they come from, and how that small sample relates to the corpus at large (if only to measure its significance). While such a work may not have been possible in the print culture of the past, we are at a transformative moment now where we can begin leveraging the scientific method in the service of classical philology.

2 Treebanks

Our work in developing treebanks for Ancient Greek and Latin are our own efforts to help move classical philology into this scientific space. A treebank is a large collection of sentences in which the syntactic relation for every word is made explicit – where a human has encoded an interpretation of the sentence in the form of a linguistic annotation. While much of the research and labor in treebanks over the past twenty years has focused on modern languages such as English [33], Arabic [22, 32] and Czech [21], recent scholarship has seen the rise of a number

---

1 “And so, by my count, with *doleo* there are 57 sentences with the accusative + infinitive against 4 with *quod*, with *miror* 110 against 8, with *glorior* 19 against 2, with *queror* 71 against 15, with *gaudeo* 84 against 9 etc.”