Journal of Open
Humanities Data

Gorman, V B 2020 Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data,* 6: 1. DOI: https://doi.org/10.5334/johd.13

DATA PAPER

Dependency Treebanks of Ancient Greek Prose

Vanessa B. Gorman
University of Nebraska-Lincoln, US
vgorman1@unl.edu

This dataset is a collection of dependency syntax trees of representative texts from ancient Greek prose authors (Aeschines, Antiphon, Appian, Athenaeus, Demosthenes, Dionysius of Halicarnassus, Herodotus, Josephus, Lysias, Plutarch, Polybius, Thucydides, and Xenophon), totaling to date 550,000+ tokens. It is hand-annotated by one person, using the Arethusa program on the Perseids website. Original texts were obtained from the Perseus Digital Library, and some (as indicated) were computer pre-parsed at the Pedalion Project. The database is stored in a stable form (2019-12-31) on Zenodo (DOI: 10.5281/zenodo.3596076) and in a continuously updated form on GitHub in .xml format (https://vgorman1.github.io/). The repository can be used for pedagogical purposes and for research in linguistics analysis and corpus linguistics, stylistics, natural language processing, classification, and literary and historical analysis.

Keywords: Linguistic analysis; Linguistics; Language development; Historiography

(1) Overview

Repository location

https://zenodo.org/record/3596076#.XlZ7CxP7Su4

Context

(2) Methods

Steps

I made the trees using the Arethusa software on the Perseids website [13]. Original text files were obtained from the Perseus Project [14] (Tufts Univ.) and from the Pedalion Project (UK Leuven). I followed the rules of dependency syntax, employing the standard AGDT 1.1 tagset [2] and refining them according to the discussion of dependency syntax offed by Pinkster [15]. I have not used the 2.0 tagset based on Smyth developed by Celano [4]: the level of specificity increases the subjectivity of the annotation decisions exponentially, often relying more on semantics than syntax (what is the difference between a partitive genitive and a genitive of material in the phrase 'piece of pie'?), and the tagset is specific to Greek, making a linguistic comparison between languages more difficult.

Sampling strategy

While no formal statistical sampling methods were used, I chose to annotate at least 20,000 tokens each from a variety of Greek prose authors. As the size of an average 'book' by many authors, it represents a dataset large enough to use for significant sampling algorithms. I have included works from the Classical, Hellenistic, and Roman periods: Aeschines, Antiphon, Appian, Athenaeus, Demosthenes, Dionysius of Halicarnassus, Herodotus, Josephus, Lysias, Plutarch, Polybius, Thucydides, and Xenophon.

Quality Control

The relation labeling follows the general instructions for the AGDT 1.1 tagset given in Bamman and Crane [2]. I have created more detailed instructions for annotating major linguistic phenomena not covered in Bamman and Crane [2] in the 'Treebanking Tips' file within this dataset, relying heavily on the parallel interpretation of dependency syntax offered for Latin by Pinkster [15].

(3) Dataset description

Object name

https://github.com/vgorman1/Greek-Dependency-Trees

Format names and versions

.xml

Creation dates

2014-03-01 to 2019-12-31

Dataset Creators

Vanessa Gorman is the manual annotator of these trees. Original Greek texts came from the Perseus Project [14] and are pre-processed within the Arethusa program at the Perseids Project [13]. Arethusa offers possible lemmas and morphology options, from which the proper form must be selected and, if necessary, corrected or created. The syntactic analysis (relation labeling) is performed manually, except that the files for Demosthenes 1 and 59 are hand-corrected from a computer version pre-parsed by the Pedalion Project at UK Leuven (main supervisor Toon Van Hal and developer Alek Keersmaekers) [9].

Language

English. Ancient Greek.

License

CC0 1.0

Repository name

Zenodo for the stable DOI. Github for the continuously updated version.

Publication date

2019-12-31 [Zenodo].

(4) Reuse potential

Linguistic analysis

Many recent advances in linguistic knowledge are due to the development of the methods of corpus linguistics. Treebanks such as the ones presented here are a resource for the application of these methods. For example, Greek copular verbs and subject-verb agreement have recently been studied on the basis of annotated dependency data [10, 11]. McGillivray and Vatri [12] use treebanks to examine the relationship of acoustic and syntactic information. I intend future work on a valency dictionary of Greek verbs based on this dataset.

Stylistics

Ancient Greek allowed a relatively free word order, and the rules that govern it are not easy to discern. Treebanks offer a powerful tool for discovering those rules. Syntactically- and morphologically-annotated data allow for word order to be studied in a controlled fashion. For example, the frequency of the relative order of a participial indirect object and a nominal direct object can be easily determined and all examples quickly identified. The advantages offered by such specificity is apparent in the recent literature [3, 4, 8].

Natural language processing

Accurate machine parsing of natural language syntax is a high priority among computer scientists. In order to achieve success in this area, it is crucial to have a sufficient corpus of accurately annotated texts to provide both training and testing data. This dataset is rare in consisting of a corpus of texts, manually annotated by one person and representing primarily Attic and Atticizing Greek, rather than the non-standard poetic or dialectic Greek of other collections (AGDT or PROIEL). Thus it offers the opportunity to develop and evaluate algorithms of a standard dialect with relatively complex morphology and frequent discontinuous syntactic structures [9].

Classification

Categorizing uncertain texts is one of the original concerns of the digital humanities. Most studies in this area have relied on various measures of vocabulary richness for their criteria of analysis, while the value of syntactic information has been discounted. In contrast, recent work shows that the morpho-syntactic data provided by the present database may significantly improve the results in some classification problems, such as author attribution [5, 6]. I am pursuing future studies on issues surrounding dubious passages and the level of stylometric variation within any one author.

Literary and historiographical analysis

Applied to single texts, classification methods using morpho-syntactic annotation can reveal divisions and segmentation invisible to the unaided eye. Investigation of these units may lead to a deeper understanding of, *inter alia*, the compositional structure of the work.

Pedagogy

The availability of a large database of syntactically-annotated sentences is an important asset for students of the ancient Greek language. The structures posited by dependency grammar are close to the kinds of grammar analyses traditional in Greek pedagogy. This similarity makes them more helpful to students than, e.g., phrase structure trees would be. In addition, recent software, such as the Alpheios browser extension, allows the trees in this corpus to be combined with vocabulary glosses and links to a standard reference grammar [1]. The result is an on-line reading environment that guides students word-by-word through definitions, morphology, and syntax (e.g., https://vgorman1.github.io/Greek-Language-Class/ [7, in progress]).

Limitations

The principal limitation of this repository lies in the human element. Just as different people make different decisions in annotating specific structures, so also the same human annotator may change her mind over time. We lack detailed instructions on specific annotation choices. I am compiling just such documentation, a very preliminary version of which can be viewed in the repository ('Treebanking Tips').

Acknowledgements

I would like to thank the instruction, guidance, and technical support I have received from Bridget Almas, Zachary Fletcher, Marie-Claire Beaulieu, and, especially, Robert Gorman.

Competing Interests

Vanessa Gorman is a voluntary consultant for the Alpheios Project [1], but otherwise declares that she has no conflicting interests.

References

1. **Alpheios Project.** Available at: https://alpheios.net/ (accessed 2019-12-31).
2. **Bamman, D** and **Crane, G.** Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1); 2008. https://static.perseids.org/guidelines-syntactic-annotation-greek-1-1.pdf (accessed 2019-12-31).

3. **Beschi, F.** The Ancient Greek Sentence Left Periphery. *Journal of Greek Linguistics*. 2018; 18: 172–210. DOI: https://doi.org/10.1163/15699846-01802003
4. **Celano, G G A.** A Computational Study on Preverbal and Postverbal Accusative Object Nouns and Pronouns in Ancient Greek. *The Prague Bulletin of Mathematical Linguistics No. 101*. 2014; 97–110. DOI: https://doi.org/10.2478/pralin-2014-0006
5. **Gorman, R J.** Author Identification of Short Texts Using Dependency Treebanks without Vocabulary. *Digital Scholarship in the Humanities*. 2019. DOI: https://doi.org/10.1093/llc/fqz070
6. **Gorman, V B** and **Gorman, R J.** Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry. *Open Linguistics*. 2016; 2: 500–510. DOI: https://doi.org/10.1515/opli-2016-0026
7. **Gorman, V B.** Reading Ancient Greek in the Digital Age. An on-line, open access course in Attic Greek. Available at: https://vgorman1.github.io/Greek-Language-Class/ (accessed 2020-1-23).
8. **Gulordava, K.** Word Order Variation and Dependency Length Minimisation: A Cross-Linguistic Computational Approach. Thèse de doctorat: Univ. Genève, 2018, no. L. 920. DOI: https://doi.org/10.13097/archive-ouverte/unige:106855
9. **Keersmaekers, A, Mercelis, W, Swaelens, C** and **Van Hal, T.** Creating, Enriching and Valorising Treebanks of Ancient Greek: the Ongoing Pedalion-project. *Semantic Scholar*. 2019; https://www.semanticscholar.org/paper/Creating-%2C-enriching-and-valorising-treebanks-of-%3A-Keersmaekers/8776d8a0ca80d1c947276cca289a0fa7d16b6671 (accessed 2019-12-31).
10. **Mambrini, F.** Nominal vs Copular Clauses in a Diachronic Corpus of Ancient Greek Historians. *Journal of Greek Linguistics*. 2019; 19: 90–113. DOI: https://doi.org/10.1163/15699846-01901003
11. **Mambrini, F** and **Passarotti, M.** Subject-Verb Agreement with Coordinated Subjects in Ancient Greek. A Treebank-Based Study. *Journal of Greek Linguistics*, 2016; 16: 87–116. DOI: https://doi.org/10.1163/15699846-01601003
12. **McGillivray, B** and **Vatri, A.** Computational Valency Lexica for Latin And Greek in Use: a Case Study of Syntactic Ambiguity. *Journal of Latin Linguistics*. 2015; 14: 101–126. DOI: https://doi.org/10.1515/joll-2015-0005
13. **Perseids Project.** Available at: https://www.perseids.org/ (accessed 2019-12-31).
14. **Perseus Digital Library.** Available at: http://www.perseus.tufts.edu/hopper/ (accessed 2019-12-31).
15. **Pinkster, H.** *Oxford Latin Syntax, Vol. 1: The Simple Clause*. Oxford: Oxford University Press; 2015. DOI: https://doi.org/10.1093/acprof:oso/9780199283613.003.0001

How to cite this article: Gorman, V B 2020 Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data* 6: 1. DOI: https://doi.org/10.5334/johd.13

Published: 26 March 2020

]u[*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS