Journal of Open Humanities Data

PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri

DATA PAPER

MARJA VIERROS
ERIK HENRIKSSON
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The PapyGreek Treebanks dataset contains documentary texts written in Postclassical Greek (ca. 300 BCE–700 CE), morphosyntactically annotated according to Dependency Grammar. The source of the texts is the Duke Databank of Documentary Papyri (DDbDP), which preserves the modern editorial treatment of the documents in TEI Epidoc XML encoding. Aiming to expose linguistic variation in the DDbDP, we have annotated two versions of a selection of documents: the plain transcription and an editorially corrected version. The dataset also comprises metadata about the documents' dating and provenance, text type, and the persons involved. Furthermore, it facilitates linguistic research on these texts.

CORRESPONDING AUTHOR:
Marja Vierros
Department of Languages, University of Helsinki, Helsinki, Finland
marja.vierros@helsinki.fi

KEYWORDS:
linguistic analysis; European languages; language variations; language development

TO CITE THIS ARTICLE:
Vierros, M., & Henriksson, E. (2021). PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri. *Journal of Open Humanities Data*, 7: 26, pp. 1–6. DOI: *https://doi.org/10.5334/johd.55*

1 OVERVIEW

REPOSITORY LOCATION

doi: *https://doi.org/10.5281/zenodo.5074307*

CONTEXT

The corpus was produced and is used by the project Digital Grammar of Greek Documentary Papyri (PapyGreek). Parts of the data were used in Vierros and Yordanova (in press).

2 METHOD

STEPS

We have obtained our source files from the Duke Databank of Documentary Papyri (DDbDP),[1] a corpus of non-literary Greek and Latin texts written on papyri, ostraca, or wooden tablets, encoded in TEI Epidoc XML (Elliott, Bodard, Cayless, & al., 2006). Based on the encoded information about the texts' modern editorial treatment, we have split the files in two, one version containing the plain transcribed text and the other the editorial corrections and regularizations (for the details, see Vierros, 2018; Vierros & Henriksson, 2017). We have then annotated both versions separately. For example, if the original document has the Greek word *moi* 'me' (dative) and the editor has corrected it to *mou* 'me' (genitive), we have annotated both forms. The two versions are encoded in our dataset using XML attributes of <word> elements suffixed with _orig and _reg (e.g., postag_orig and postag_reg).

The annotation has been done at the PapyGreek website[2] using an embedded Arethusa Treebank editor (Perseids Project and Alpheios Project, Ltd.).[3] We have used the morphological analyses provided by Morpheus (Crane, 1991), an automatic morphological tagger which ships with Arethusa, as well as those produced by Alek Keersmaekers[4] using machine learning techniques (Keersmaekers & Depauw, in press), making corrections as necessary. Syntactic annotation has been done manually.

The dataset also includes various kinds of metadata. Dating and location metadata has been retrieved from the Heidelberger Gesamtverzeichnis der Griechischen Papyrusurkunden Ägyptens, or HGV (part of the idp.data repository). References to the source HGV and DDbDP files can be found in the `<document_meta>` element of each file. In addition, we have described the persons associated with producing each text: authors (e.g., the sender of a letter), writers (who penned the text), addressees, and external scribal officials. We use our own designated person IDs for those persons, but when available, we have added the Trismegistos Person identifiers as well (e.g., Depauw & Van Beek, 2009). TM identifiers are not available for the anonymous scribes, whose handwriting may be recognised across different documents, and we consider marking them worthwhile. We have also described the documents' text types (not indicated in the source files). The document types are listed in three levels: hypercategory (e.g., law), category (e.g., contract) and subcategory (e.g., marriage). This typology is planned to be compatible with the data of the project Everyday Writing in Graeco-Roman and Late Antique Egypt: A Socio-Semiotic Study of Communicative Variation (EVWRIT).[5]

Although the dataset is in XML, we use a MySQL database to store and process our data in our server, with separate tables for documents, sentences and words. There are two main advantages of working with a relational database vs. XML files: the former is easier to update incrementally; and indexed SQL tables are arguably better suited for data queries, the development of which is central to the PapyGreek project (see 4 below). The dataset's XML files have been generated using SQL table joins and the Python package lxml.[6]

1 *https://github.com/papyri/idp.data* (last accessed: 2021-08-06).

2 *https://papygreek.hum.helsinki.fi* (last accessed: 2021-10-04).

3 *https://github.com/alpheios-project/arethusa* (last accessed: 2021-08-06).

4 *https://github.com/alekkeersmaekers/duke-nlp* (last accessed: 2021-08-06).

5 *https://www.evwrit.ugent.be/* (last accessed: 2021-10-14).

6 *https://lxml.de* (last accessed: 2021-08-06).

DOI: 10.5334/johd.55

SAMPLING STRATEGY

The DDbDP contains many different types of documents, of which we have selected to annotate—in the first phase—mostly private and business letters (ca. 25K tokens), and petitions (ca. 11K tokens). We have selected the texts from certain archives in order to have a known context for the texts and the people within (e.g., the Zenon Archive, the Archive of Katochoi of the Sarapieion, the Athenodoros Archive, the ostraca from Mons Claudianus and selection of women's letters from different time periods). The present release (v1.01) focuses on the period BCE (ca. 32K tokens from the total of 44K). Later versions are planned to cover a larger time frame and range of text types.

QUALITY CONTROL

We have followed the Ancient Greek Dependency Treebank Guidelines 2.0 (Celano, 2014b), which builds on version 1.1 (Bamman & Crane, 2008) and originates in the Prague Dependency Treebank (see e.g., Celano, 2019; Hajič, 1998). We have not applied the advanced syntactic/semantic layer of AGDT Guidelines 2.0. Additional PapyGreek Guidelines can be found in a separate document in the data repository. Each text has gone through a human review process.[7]

3 DATASET DESCRIPTION

OBJECT NAME

PapyGreek Treebanks

FORMAT NAMES AND VERSIONS

XML

CREATION DATES

2019-09-06 to 2021-06-23.

DATASET CREATORS

Marja Vierros (reviewer, annotator); Erik Henriksson (developer); Polina Yordanova (reviewer, annotator); Arttu Alaranta (annotator); Petri Lahtinen (annotator); Lauri Marjamäki (annotator); Jamie Vesterinen (annotator); Iida Huitula (annotator); Sari Kock (annotator). Affiliation of all: University of Helsinki.

LANGUAGE

Original texts: Ancient Greek. Other: English.

LICENSE

CC BY-SA 4.0

REPOSITORY NAME

Zenodo; GitHub

PUBLICATION DATE

2021-06-23

4 REUSE POTENTIAL

LINGUISTIC ANALYSIS

Ancient Greek documentary sources have not been previously linguistically annotated with a review process (Vierros, 2018, pp. 105–106), making the PapyGreek treebanks a valuable

7 Current review board: Marja Vierros and Polina Yordanova.

DOI: 10.5334/johd.55

resource in the study of Postclassical Greek (e.g., historical morphology and syntax, linguistic variation, and historical sociolinguistics). The present release (v1.01) covers the CE centuries in smaller numbers, and is thus indicative of the Greek usage mostly in the early post-classical period. The genres of private and administrative letters and petitions represent well the everyday uses of the language, including both the formulaic and narrative parts in petitions and the private language use in letters.

An understudied topic in Greek syntax—word order—has gained fruitful research results with treebanked data (Mambrini & Passarotti, 2013); but more well-studied questions, too, may benefit from revisiting quantitatively using treebanks (e.g., Celano, 2014a; Mambrini, 2019). A PapyGreek search interface[8] developed by Erik Henriksson combines orthographic queries with morphosyntactic ones, making it a powerful tool to study morphosyntactic variation in tandem with phonology.

UNIVERSAL DEPENDENCIES AND NLP

Treebanks using the Ancient Greek Dependency Grammar specification can be converted into other formats, such as Universal Dependencies, and thus be used together with other languages or corpora. We decided to use the AGDT formalism instead of UD, because we wish our data to be directly comparable to other genres of Ancient Greek data, for which treebanked data existed only in AGDT when we started. Due to the simplicity of the schemata used, converting XML treebanks from one formalism to another is potentially a trivial task (Celano, 2019, p. 283); but see (Cecchini, Korkiakangas, & Passarotti, 2020) for an example of a more complicated transition.

Treebanks can also be used as training data for automatic lemmatizers and morphosyntactic parsers. We indeed hope that the PapyGreek treebanks are exploited for such purposes, in particular the development of automatic parsers of non-standard documentary Greek (see Keersmaekers, Mercelis, Swaelens, & Van Hal, 2019; Mambrini & Passarotti, 2012). Treebanked data can also serve as a basis for automatic semantic role labeling (Keersmaekers, 2020).

AUTHORSHIP ATTRIBUTION

Ancient Greek historical prose treebanks (V. B. Gorman, 2020) have been used in, e.g., stylometric and authorship attribution studies (R. Gorman, 2019; V. B. Gorman & R. J. Gorman, 2016). A paper on authorship attribution, which utilizes the PapyGreek treebanks' person metadata (see 2 above), is being prepared by the current authors. The results, we hope, will be useful in evaluating the practicality of authorship attribution based on short and fragmentary texts.

PEDAGOGY

Treebanking has proven to be an effective way to teach ancient languages (V. B. Gorman, 2021; Mambrini, 2016). The annotation software developed by the Perseids Project[9] and the Alpheios Project, Ltd.[10] are freely available, promoting the equality of learning Greek and Latin worldwide. The PapyGreek website is likewise open for all to learn papyrological Greek by using (and even creating new) treebanks. A pivotal part of the PapyGreek project is the development of an online Grammar of Postclassical Greek based on the treebanked material, which will serve as a valuable teaching resource.

LIMITATIONS/BARRIERS

Thus far, the size of the PapyGreek treebanks repository is limited due to the costly method of semi-manual annotation and review process. The token count is 44K, which makes 1.6% of the total number of tokens of texts that are "treebankable" in the DDbDP corpus (ca. 2.8M tokens; as "not treebankable" we have counted e.g., lists and labels). Even small-scale vetted data, however, are a valuable resource for improving the accuracy of automatic parsers.

8 *https://papygreek.hum.helsinki.fi/search* (last accessed: 2021-10-04).

9 *https://www.perseids.org* (last accessed: 2021-08-06).

10 *http://www.alpheios.net* (last accessed: 2021-08-06).

ACKNOWLEDGEMENTS

We would like to thank Bridget Almas and Zachary Fletcher from the Perseids project for help with Arethusa, and Alek Keersmaekers for automated morphological data.

FUNDING INFORMATION

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758481).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Marja Vierros: Conceptualization, Data curation; Funding acquisition, Investigation, Methodology, Resources, Supervision, Writing. **Erik Henriksson:** Software, Data curation, Investigation, Validation, Writing.

AUTHOR AFFILIATIONS

Marja Vierros *orcid.org/0000-0001-8531-7055*
Department of Languages, University of Helsinki, Helsinki, Finland
Erik Henriksson *orcid.org/0000-0003-0850-7554*
Department of Languages, University of Helsinki, Helsinki, Finland

REFERENCES

Bamman, D., & **Crane, G.** (2008). *Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank 1.1* [The Perseus Project, Tufts University]. Retrieved from *https://static.perseids.org/guidelines-syntactic-annotation-greek-1-1.pdf* (last accessed: 6 July 2021).

Cecchini, F. M., Korkiakangas, T., & **Passarotti, M.** (2020). A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the 12th language resources and evaluation conference* (pp. 933–942). Marseille, France: European Language Resources Association. Retrieved from *https://aclanthology.org/2020.lrec-1.117* (last accessed: 4 October 2021).

Celano, G. G. A. (2014a). A Computational Study on Preverbal and Postverbal Accusative Object Nouns and Pronouns in Ancient Greek. *The Prague Bulletin of Mathematical Linguistics, 101* , 97–110. DOI: *https://doi.org/10.2478/pralin-2014-0006*

Celano, G. G. A. (2014b). *Guidelines for the Annotation of the Ancient Greek Dependency Treebank.* Retrieved from *https://github.com/PerseusDL/treebankdata/tree/master/AGDT2/guidelines* (last accessed: 6 July 2021).

Celano, G. G. A. (2019). The Dependency Treebanks for Ancient Greek and Latin. In M. Berti (Ed.), *Digital Classical Philology* (pp. 279–298). Berlin: De Gruyter. DOI: *https://doi.org/10.1515/9783110599572-016*

Crane, G. R. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing, 6*, 243–245. DOI: *https://doi.org/10.1093/llc/6.4.243*

Depauw, M., & **Van Beek, B.** (2009). People in Greek Documentary Papyri. First Results of a Research Project. *Journal of Juristic Papyrology, 39*, 31–47.

Elliott, T., Bodard, G., Cayless, H., et al. (2006). *EpiDoc: Epigraphic Documents in TEI XML*. Retrieved from *http://epidoc.stoa.org* (last accessed: 6 July 2021).

Gorman, R. (2019). Author identification of short texts using dependency treebanks without vocabulary. *Digital Scholarship in the Humanities, 35*(4), 812–825. DOI: *https://doi.org/10.1093/llc/fqz070*

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*

Gorman, V. B. (2021). *Reading Ancient Greek in the Digital Age.* Retrieved from *https://vgorman.com* (last accessed: 6 August 2021).

Gorman, V. B., & **Gorman, R. J.** (2016). Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry. *Open Linguistics, 2*(1). DOI: *https://doi.org/10.1515/opli-2016-0026*

Hajič, J. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová* (pp. 12–19). Prague: Charles University Press.

Keersmaekers, A. (2020). Automatic semantic role labeling in Ancient Greek using distributional semantic modeling. In *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 59–67). Marseille: European Language Resources Association (ELRA).

Keersmaekers, A., & **Depauw, M.** (in press). Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri. In *Classics@*. Center for Hellenic Studies.

Keersmaekers, A., Mercelis, W., Swaelens, C., & **Van Hal, T.** (2019). Creating, Enriching and Valorizing Treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (pp. 109–117). DOI: *https://doi.org/10.18653/v1/W19-7812*

Mambrini, F. (2016). The Ancient Greek Dependency Treebank: Linguistic Annotation in a Teaching Environment. In G. Bodard & M. Romanello (Eds.), *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement* (pp. 83–99). London: Ubiquity Press. DOI: *https://doi.org/10.5334/bat.f*

Mambrini, F. (2019). Nominal vs copular clauses in a diachronic corpus of Ancient Greek historians: A treebank-based analysis. *Journal of Greek Linguistics, 19*(1), 90–113. DOI: *https://doi.org/10.1163/15699846-01901003*

Mambrini, F., & **Passarotti, M.** (2012). Will a Parser Overtake Achilles? First experiments on parsing the Ancient Greek Dependency Treebank. In *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11). 30 November–1 December 2012* (pp. 133–144). Lisbon, Portugal.

Mambrini, F., & **Passarotti, M.** (2013). Non-projectivity in the Ancient Greek dependency treebank. In *Proceedings of the second international conference on dependency linguistics (Depling 2013)* (pp. 177–186). Prague: Matfyzpress. Retrieved from *https://aclanthology.org/W13-3720/* (last accessed: 4 October 2021).

Vierros, M. (2018). Linguistic Annotation of the Digital Papyrological Corpus. In N. Reggiani (Ed.), *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri* (pp. 105–118). Berlin: De Gruyter. DOI: *https://doi.org/10.1515/9783110547450-006*

Vierros, M., & **Henriksson, E.** (2017). Preprocessing Greek Papyri for linguistic annotation. *Journal of Data Mining & Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*. DOI: *https://doi.org/10.46298/jdmdh.1385*

Vierros, M., & **Yordanova, P.** (in press). Querying syntactic constructions in Ancient Greek parsed corpora: A case study on the genitive absolute in literature and documentary papyri. In A. Novokhatko, S. Chronopoulos, & F. K. Maier (Eds.), *Classics@*. Center for Hellenic Studies.

TO CITE THIS ARTICLE:
Vierros, M., & Henriksson, E. (2021). PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri. *Journal of Open Humanities Data*, 7: 26, pp. 1–6. DOI: *https://doi.org/10.5334/johd.55*

Published: 05 November 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.