Ancient Greek corpora and treebanks

Christianna Antonopoulou[1] & Stavros Skopeteas[2]

University of Athens[1] and University of Göttingen[2]

October 2023

The following summary contains notes that we created for our research and is certainly not exhaustive. We are happy to share it online with interested students and researchers.

1 Corpora 2

1.1 Treebanks 2

1.2 Further corpora online 4

1.3 Lists of collections 5

1.4 Further information ..... **Fehler! Textmarke nicht definiert.**

2 Tools 6

2.1 DendroSearch 6

2.2 Arethousa 6

2.3 Pedalion 7

3 Studies 8

1 Corpora

1.1 Treebanks

1.1.1 Pedalion

Editors: Toon Van Hal & Alek Keersmaekers

Contents: Aeneas, Aesopus, Septuaginta, Aristophanes, Chariton, Chion, Epictetus, Epicurus, Euripides, Heron, Julian, Pseudo-Homer, Isocrates, Longus, Lucian, Lysias, Menander, Paeanius, Phlegon, Plato, Procopius, Sappho, Sextus Empiricus, Theophrastus, Xenophon, diverse authors

Format:

Queries/Visualization: https://perseids-publications.github.io/pedalion-trees/

Download: https://github.com/perseids-publications/pedalion-trees/tree/master/public/xml

1.1.2 PapyGreek 3.0

Citation: Vierros, Marja, Henriksson, Erik, Yordanova, Polina, Alaranta, Arttu, Lahtinen, Petri, Vesterinen, Jamie, Huitula, Iida, & Kock, Sari. (2021). PapyGreek Treebanks (v1.01) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5074307

Contents: Literary and documentary Papyri

Comment: contains the results of the project *Linguistic Annotation of the Greek Documentary Papyri – Detecting and Determining Contact-Induced, Dialectal and Stylistic Variation* (M. Vierros)

Format: XML

Project website: https://www.helsinki.fi/en/researchgroups/digital-grammar-of-greek-documentary-papyri

Queries: https://papygreek.com/search

Download: https://zenodo.org/record/7157692

1.1.3 PROIEL

Contents: New Testament (parts), Herodot (parts), Chronicle Sphrantzes (15 BCE)

Comment: test data manually annotated, annotations automatically propagated in development data.

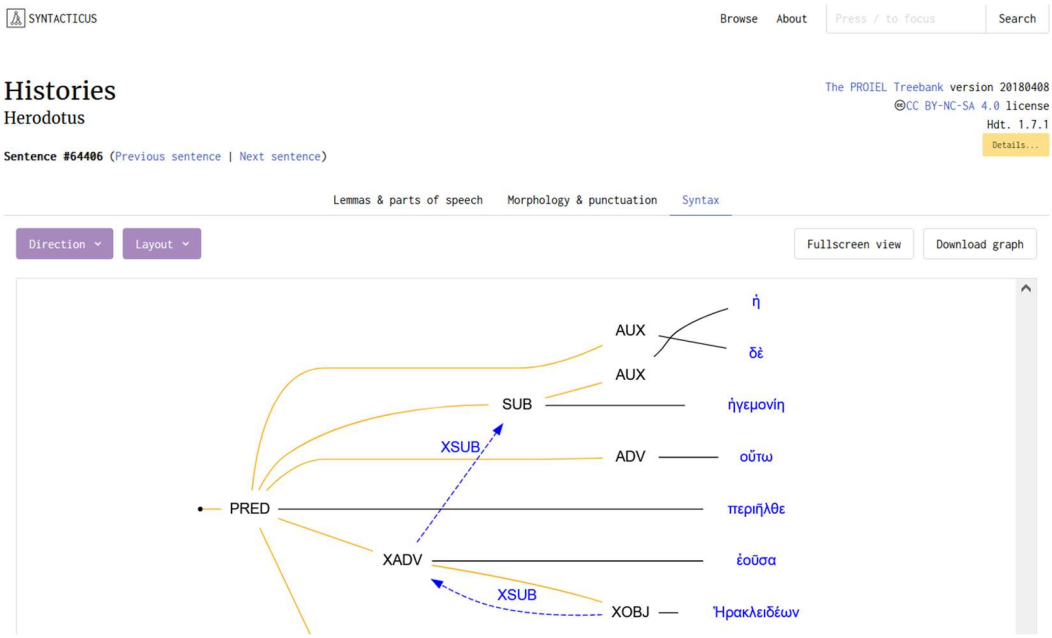
Format: CONLL

Queries/visualizations:

available in Dendrosearch

Weblicht, https://weblicht.sfs.uni-tuebingen.de/Tundra/UD_Ancient_Greek-PROIEL_v2.4/

PROIEL-Reader, hosted by syntacticus.org: https://syntacticus.org/browse



Annotation tool: https://github.com/mlj/proiel-webapp

Download: https://github.com/UniversalDependencies/UD_Ancient_Greek-PROIEL

Guidelines. https://dev.syntacticus.org/annotation-guide/

see also tagset for POS in Eckhoff et al. 2018, 10.1007/s10579-017-9388-5

and the summary in the UD page: https://universaldependencies.org/treebanks/grc_proiel/index.html

1.1.4 AGDT

Contents: Aeschylus, Aesop, Atheneus, Diodorus Sicilius, Herodotus, Hesiod, Homer, Lysias, Plato, Plutarch, Polybius, Pseudo-Apollodorus, Pseudo Homer, Sophocles, Thucydides

Comment: small-size treebank

Description: https://github.com/PerseusDL/treebank_data/tree/master/v2.0/Greek

https://perseusdl.github.io/treebank_data/

Download: https://perseusdl.github.io/treebank_data/

Guidelines: https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md

see also http://www.perseus.tufts.edu/~ababeu/tlt8.pdf

1.1.5 Papyri.info

https://papyri.info/

1.1.6 Duke-nlp

Citation: Keersmaekers, Alek, and Mark Depauw. Forthcoming. “Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri.” *Classics@*.

Contents: automatically parsed corpus of documentary papyri—Duke-nlp—consisting of most papyri available in the **papyri.info**. The files can be downloaded and queried with DendroSearch—the Duke-nlp data is divided into several smaller subsections by text types.

Format: XML

Download: https://github.com/alekkeersmaekers/duke-nlp/tree/master/xml

1.1.7 GLAUx

Contents: large corpus with many authors, Classical, Postclassical

Comment: automated/unsupervised,

Format: XML

Visualization: https://perseids-publications.github.io/glaux-trees/

Download: https://github.com/perseids-publications/glaux-trees/tree/master/public/xml

1.1.8 Gorman treebanks

Contents: Aeschines, Antiphon, Appian, Athenaeus, Demosthenes, Dionysius of Halicarnassus, Herodotus, Josephus, Lysias, Plutarch, Polybius, Thucydides, Xenophon and further authors (but parts of books)

Format: XML

Queries: available in Dendrosearch

Download: https://github.com/vgorman1/Greek-Dependency-Trees/tree/master/xml%20versions

Description: 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

1.1.9 Harrington's Treebanks

small collection for pedagogical purposes:

Queries: available in Dendrosearch

Guidelines, Description: https://perseids-project.github.io/harrington_trees/

Download:

1.2 Further corpora online

DAMOS	database of Mycenaean at Oslo, not downloadable	https://damos.hf.uio.no/
Dodona	oracular questions from Dodona, not downloadable	https://dodonaonline.com/
Papyri.info	documentary and literary papyri, searchable corpus	www.papyri.info
TLG	Thesaurus, texts	https://stephanus.tlg.uci.edu/
Perseus	downloadable texts	https://www.perseus.tufts.edu/hopper/opensource/download
Inscriptions	searchable inscriptions, not downloadable	https://inscriptions.packhum.org/

Alpheios	Aeschylos, Hesiod, Homer, Longus, Lysias, Plato, Sophocles, Xenophon tools for understanding Greek texts, accompanied with lexical/grammatical resources that can be used for language learning.	https://alpheios.net/

1.3 Lists of collections

UChicago	https://guides.lib.uchicago.edu/efts/greek
Lexicity	http://lexicity.com/resources/greek/texts/
Digital classicist	https://wiki.digitalclassicist.org/Category:Corpora
Trismegistos	https://www.trismegistos.org/

2 Tools

2.1 DendroSearch

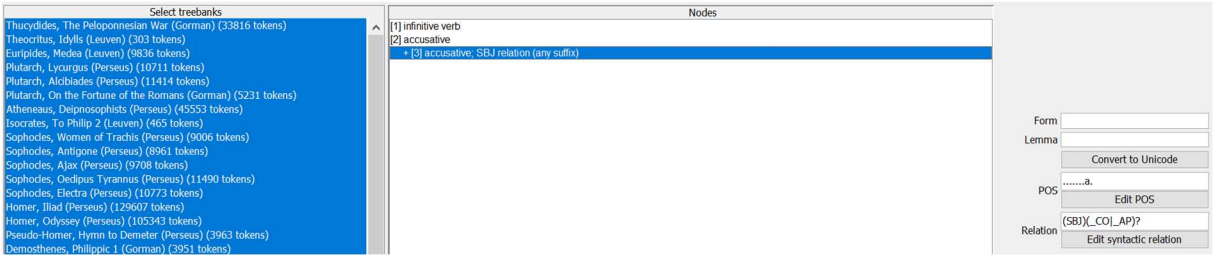
Tool for querying Ancient Greek treebanks

Download: https://github.com/alekkeersmaekers/dendrosearch (click on "Code" > download zip)

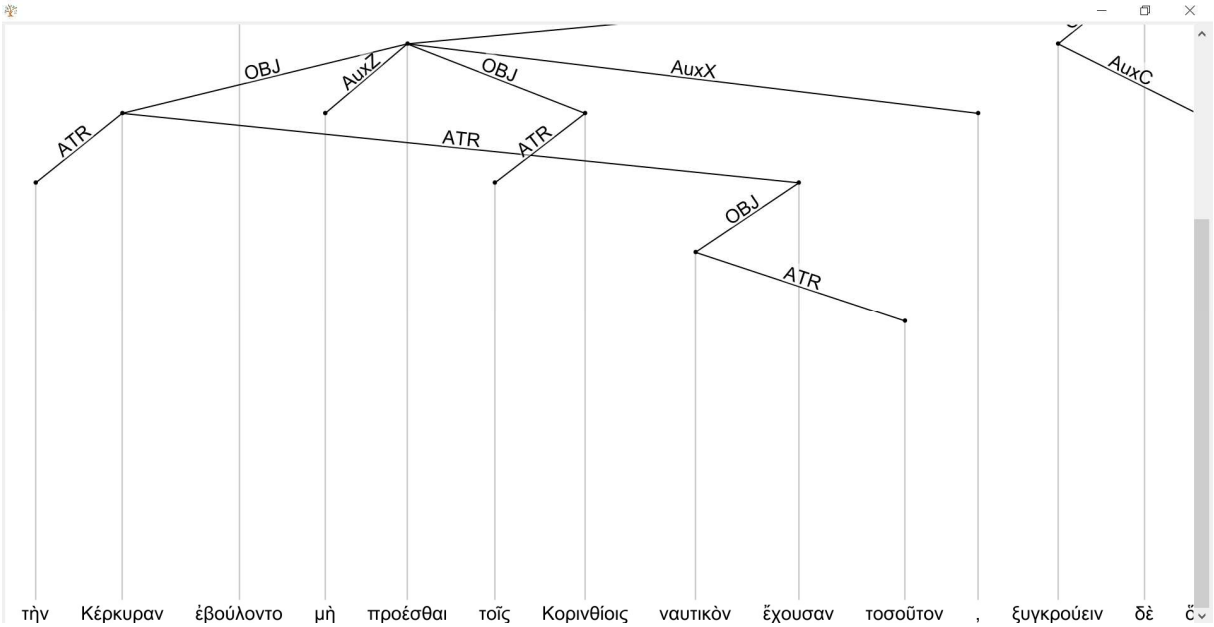
Contents: contains several databases of the perseus project (PROIEL, Gorman, AGDT, Leuven, Harrington), syntax searchable.

Guidelines: guideline html file is found in the distribution, illustrative examples with syntactic annotations

Examples: infinitives with accusative subject (user guidelines)



Clicking on the result sentences opens the dependency relations in an editor:



The findings can be saved in a .txt file and further processed in EXCEL.

2.2 Arethousa

Tool of the Perseus project for data annotations. Using the Morpheus tool for morphological annotation, provides facilities for the annotation of dependencies.

Description/Guidelines: https://www.perseids.org/tools/arethusa/app/#/

Download: https://github.com/alpheios-project/arethusa

2.3 Pedalion

Some query functions are online implemented in Pedalion.

https://en.pedalion.org/node/205069

3 Studies

Bamman, D. and Crane, G. 2008. Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1).

static.perseids.org/guidelines-syntactic-annotation-greek-1-1.pdf

Beschi, F. The Ancient Greek Sentence Left Periphery. Journal of Greek Linguistics. 2018; 18: 172–210.

DOI: doi.org/10.1163/15699846-01802003

Baumgardt, Frederik, Monica Berti, Giuseppe G. A. Celano, Gregory R. Crane, Stella Dee, Maryam Foradi, Emily Franzini, Greta Franzini, Simona Stoyanova 2014. The Open Philology Project at the University of Leipzig. *DH 2014*

Open Philology Project: open, extensible; dynamic textbooks that use annotated corpora; workflows for integrating machine-annotated data.

lrec-conf.org/proceedings/lrec2014/pdf/940_Paper.pdf

Celano, Giuseppe G.A. 2019. The Dependency Treebanks for Ancient Greek and Latin. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, Berlin, Boston: De Gruyter Saur, pp. 279-298.

dependency treebanks for Ancient Greek and Latin: Ancient Greek and Latin Dependency Treebank (AGLDT), Index Thomisticus Treebank (IT-TB), PROIEL Treebank, and SEMATIA Treebank.

doi.org/10.1515/9783110599572-016

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: doi.org/10.2478/pralin-2014-0006

Eckhoff, Hanne M., Kristin Bech, G. Bouma, K. Eide, D. Haug, O. E. Haugen, Marius L. Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*.

dependency treebanks of early attestations of Indo-European languages; web annotation interface, set of annotation schemes, guidelines. Dependency grammar scheme complemented by detailed morphological tags.

doi.org/10.1007/s10579-017-9388-5

Gorman, R J. 2019. Author Identification of Short Texts Using Dependency Treebanks without Vocabulary. Digital Scholarship in the Humanities.

DOI: 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 and Gorman, R J. 2016. Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry. Open Linguistics; 2: 500–510.

DOI: doi.org/10.1515/opli-2016-0026

Gulordava, K. 2018. Word Order Variation and Dependency Length Minimisation: A Cross-Linguistic Computational Approach. Thèse de doctorat: Univ. Genève, no. L. 920.

DOI: doi.org/10.13097/archive-ouverte/unige:106855

dependency distances and syntactic structure cross-linguistic treebanks

Haug, D. Marius L. Jøhndal 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations.

syntactic annotation scheme, comparative study of the oldest extant versions of the New Testament in Indo-European languages: Greek original, translations into Latin, Gothic, Armenian and Church Slavonic. Tag set for syntactic variation, pragmatics, and information structure.

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

Lavidas, Nikolaos and Haug, Dag Trygve Truslew. 2020. Postclassical Greek and Treebanks for a Diachronic Analysis. *Postclassical Greek: Contemporary Approaches to Philology and Linguistics*, edited by Dariya Rafiyenko and Ilja A. Seržant, Berlin, Boston: De Gruyter Mouton. 163-202.

diachronic analysis based on an annotated corpus, various Postclassical stages of Greek; diachronic examination of backward control in Sphrantzes as well as in Herodotus and the Gospels

doi.org/10.1515/9783110677522-008

Mambrini, F. 2019. Nominal vs Copular Clauses in a Diachronic Corpus of Ancient Greek Historians. Journal of Greek Linguistics; 19: 90–113.

DOI: doi.org/10.1163/15699846-01901003

nominal and copular construction of predicate nominal; Ancient Greek Dependency Treebank (AGDT); historians Herodotus, Thucydides, Polybius.

Mambrini, F and Passarotti, M. 2016. Subject-Verb Agreement with Coordinated Subjects in Ancient Greek. A Treebank-Based Study. Journal of Greek Linguistics; 16: 87–116.

DOI: doi.org/10.1163/15699846-01601003

agreement can be controlled by the coordinated phrase as a whole, or it can be triggered by just one of the coordinated words. subject and verb agreement, one morphological feature that is expected to covary (number).

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: doi.org/10.1515/joll-2015-0005

Vierros, M, P Valentinova Yordanova. 2022. Querying syntactic constructions in Ancient Greek parsed corpora: A case study on the genitive absolute in literature and documentary papyri. Classics@

the genitive absolute construction; comparing results obtained from querying (a) two corpora of literary texts (AGDT and Gorman), (b) PapyGreek corpus of Greek documentary papyri through XSLT-based queries.

Vierros, MK, EI Henriksson. 2017. Preprocessing Greek Papyri for linguistic annotation. Journal of Data Mining and Digital Humanities,

Greek documentary papyri. Digital platform “Sematia”, transforming the digital texts available in TEI EpiDoc XML format to a format which can be morphologically and syntactically annotated (treebanked); metadata concerning the text type, writer and handwriting of each act of writing.

hdl.handle.net/10138/313221

Vierros, MK, EI Henriksson. 2021. PpyGreek Treebanks: A dataset of Linguistically Annotated Greek Documentary Papyri.