Creating, enriching and valorising treebanks of Ancient Greek: the ongoing Pedalion-project

Alek Keersmaekers
Research Foundation Flanders & KU Leuven
alek.keersmaekers@kuleuven.be

Wouter Mercelis
KU Leuven
wouter.mercelis@student.kuleuven.be

Colin Swaelens
KU Leuven
colin.swaelens@student.kuleuven.be

Toon Van Hal
KU Leuven
toon.vanhal@kuleuven.be

Abstract

This paper shows the extent to which treebanks of Ancient Greek play a central role in the ongoing Pedalion project at the University of Leuven. Building on diverse treebanks readily available today, the project aims to make progress in the automated parsing of classical and post-classical Greek texts. Rather than developing new technology as such, our project endeavours to make deliberate and methodical use of the technology that already exists, essentially by *combining* and *adapting* both technology and data. This contribution offers a 'roadmap' of our project, surveying (a) the existing work on which we can rely, (b) the strategies which we adopt to reach better results in the automated processing of Ancient Greek and (c) the deliverables that have already been realised or are forthcoming.

1 Introduction

Although corpus-based methods are becoming increasingly central to present-day research in historical linguistics, the possibilities for conducting corpus-based linguistic research in Ancient Greek are still restricted, despite a range of recent international research initiatives (see Haug, 2014). Our ongoing project aims to make some progress in the automated language progressing of Ancient Greek. It starts from the basic assumption that promising results with wide-ranging applicability can be achieved by relying on the invaluable work already undertaken in a wide range of Ancient Greek dependency treebank projects. The specifics of our approach can be characterised as follows:

- Rather than developing new technology as such, our project endeavours to make deliberate and methodical use of the technology that already exists, essentially by *combining* and *adapting* both technology and data. In doing so, special attention is paid to the specifics of the Greek language.
- Instead of aiming solely at reaching better parsing accuracy, the project also aims to offer a number of tangible deliverables.
- Such deliverables should not be limited to specialised instruments tailored to the needs of researchers and linguists: there are also didactic applications in development that can assist a larger audience in mastering Ancient Greek.

In what follows, we offer a 'roadmap' of our project by succinctly outlining (a) the work on which we are gratefully building (section 2), (b) the strategies we adopt to achieve better results in the automated processing of Ancient Greek (section 3) and (c) the deliverables that have already been realised or are in progress (section 4). In a recent paper, Simon Mahony highlighted the importance of "joining together and sharing resources", particularly "[i]n the case of ancient languages, just as with other vulnerable subject areas" (Mahony, 2016, 44). We hope that our ongoing project in some measure meets Mahony's concern.

2 Elaborating on existing treebanks

Two projects are well-known and are prominently present in the yearly CONLL shared task (Zeman et al., 2018), the Perseus Ancient Greek and Latin Dependency Treebanks and the PROIEL Treebank. But there are many additional undertakings that deserve special mention. Hence, we offer a succinct survey

of dependency treebanks of Ancient Greek (for a survey which includes constituency treebanks as well, see Robie, 2017):

- Perseus *Ancient Greek Dependency Treebanks* (AGDT); ca. 550K tokens. Encompasses Archaic poetry, Classical poetry and prose. Offers lemma, morphological, syntactic and (in a few cases) semantic information. Own annotation style. See (Bamman and Crane, 2011).
- PROIEL treebanks; ca. 248K tokens. Encompasses prose texts. Offers lemma, morphological, syntactic and pragmatic information. Own annotation style. See (Haug and Jøhndal, 2008).
- Sematia; ca. 6K tokens. Documentary papyri. Offers lemma, morphological and syntactic information, following the AGDT annotation scheme (with some minor modifications). See (Henriksson and Vierros, 2017).
- Gorman treebanks; ca. 240K tokens. Encompasses prose texts. Offers lemma, morphological and syntactic information. Complies with the AGDT annotation scheme. See (Gorman, 2016).
- Harrington treebanks; ca. 18K tokens. Encompasses prose texts. Offers lemma, morphological, syntactic and semantic information, following a modified version of the AGDT annotation scheme. See (Harrington, 2018).
- Pedalion treebanks; ca. 119K tokens. Offers lemma, morphological and syntactic information, currently experimenting with semantic information, following the AGDT annotation scheme. See below.
- Aphthonius; ca. 7K tokens. Encompasses prose texts. Offers lemma, morphological, syntactic and semantic information, following the AGDT annotation scheme. See (Yordanova, 2018).

In order to be able to join forces with the data sets outlined above and to enable communication between them, we imported the XML-files into a relational FileMaker Database, which serves as the back-office of our undertaking. The annotation styles of both the PROIEL treebank—whose set of syntactic labels is slightly more extensive than the set used in the Perseus treebanks, given that, for instance, special labels are for instance assigned to 'agent' and indirect objects (see Haug, 2010)—and the Harrington treebank were automatically converted to the Perseus standards on the basis of a rule-based method.

3 NLP technology and strategies used

The project's current focus lies on making progress in automated syntactic analysis. Scholars active in the field of stochastic natural language processing approaches to Ancient Greek have so far focused mainly on morphological analysis (see, for instance, Dik, 2018 and Celano et al., 2016). Keersmaekers (2019) recently succeeded in obtaining very promising results for morphology (ca. 95% accuracy) based on a text corpus focusing on the Greek papyri, while also including tokenisation and lemmatisation (the latter with about 99% accuracy) in a pipeline model. This offered a good starting point for further progress in automated syntactic analysis.

Due to the free constituent order and the highly inflected nature of Ancient Greek, progress in automatically analysing Ancient Greeks texts is rather slow. Techniques successfully applied to English texts do not guarantee the same level of performance when applied to an Ancient Greek corpus. Lee et al. (2011) achieve an Unlabelled Attachment Score of 70.5% with a joint tagging/parsing model, while the highest Labelled Attachment Score (LAS) Mambrini and Passarotti (2012) report is 71.7%, trained and tested on Homeric Greek. In the most recent CONLL shared task on multilingual syntactic parsing, the highest achieved LAS (with the HIT-SCIR parsing system) is 79.4% for the Perseus treebanks and 79.3% for the PROIEL treebanks (Zeman et al., 2018).

In order to achieve better results in the automatic analysis of Greek sentences, we have developed multiple strategies. Considerations of space prevent us from fleshing out the strategies which have so far been implemented in order to obtain better results (a more extensive overview of the strategies implemented is in preparation). We will limit ourselves to a succinct survey:

- **Expanding** the training data. The fact that the results of machine learning strongly depend on the extent of the available data is sometimes substantiated by referring to a quote by Peter Norvig, Google's director of Research, who once said that his company did not have "better algorithms, we just have more data" (see, e.g., Rosenfeld and Kraus, 2018: 41). We will discuss this in section 4.1.

- **Homogenising** the training data. It is not only the extent of the data that matters, quality is also key (see, e.g, Schluter and Van Genabith, 2007). An important goal of our research is to make the existing treebank data available more homogeneous, since the number of different annotators and standards has led to a large number of inconsistencies. This will improve the 'learnability' of the data for a syntactic parser, as well as create a better standard against which the test data can be evaluated (while also enhancing the possibilities for corpus linguistic research). The complex FileMaker database, containing all tokens of all available Ancient Greek dependency treebanks, has proven to be an invaluable tool in detecting inconsistencies. See section 4.2 for more information.
- **Adapting** the annotation format: the annotation style of the Perseus treebanks is inspired by the one used by the Prague Dependency Treebank and is easily human-readable. However, this does not guarantee that it is also easy to learn for an automatic parsing system. Therefore, we tested which annotation styles are the easiest to learn for specific structures which the parser typically struggles with, including elliptic and coordination structures. We did so by automatically transforming the trees on the basis of a number of rules and testing the accuracy on a test set. For coordination structures, for instance, we found that it is possible to increase parsing accuracy by 5-6% points overall (and 25-30% points for the nodes involved in these structures) if the data are presented in the right format—this involved attaching nodes involved in a coordination structure directly to one of the previous coordinated nodes with the generation relation 'CO' (coordinate), in a way comparable to the style of annotation of the Universal Dependencies project.
- **Enriching** the annotation format: we experimented with several features, including enriched part-of-speech tags and semantic information, to further improve parsing accuracy (see also section 4.4).
- **Testing** different parsers (see Mercelis, 2019): our earliest parsing experiments all made use of MaltParser (Nivre et al., 2007). In addition, the integration of MaltOptimizer (Ballesteros and Nivre, 2012) allowed the parser to select the most optimal features for the analysis of Ancient Greek. However, since the results of MaltParser were relatively modest (a LAS of about 0.734 on our test data, cf. section 4.1), we also tested some more recently developed parsers, which use neural networks, i.e. ComboParser (Rybak and Wroblewska, 2018) and the Turku Neural Parser (Kanerva et al., 2018). With the latter in particular we were able to make major improvements, reaching an LAS of up to 90 per cent. However, this number is based on manually annotated morphology, while the numbers are probably lower for automatically morphologically annotated texts.

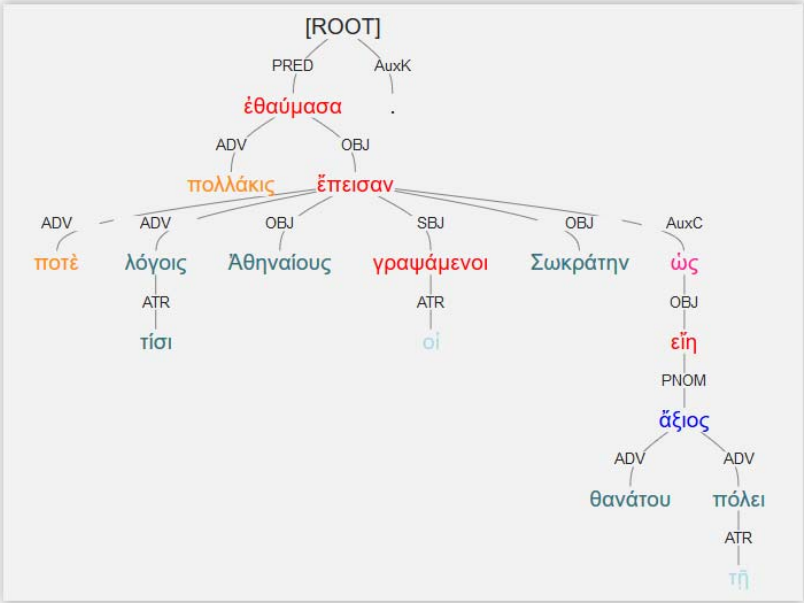

Figure 1: Example of an automatically annotated sentence
["πολλάκις ἐθαύμασα τίσι ποτὲ λόγοις Ἀθηναίους
ἔπεισαν οἱ γραψάμενοι Σωκράτην ὡς ἄξιος εἴη θανάτου τῇ πόλει"]

Nevertheless, it is still relatively difficult to evaluate and compare parsing results, because the data we use still contains several inconsistencies (see above)—let alone the fact that in some cases multiple analyses of the Greek can be defended. Moreover, in many cases having the right head is much more important than having the correct relation (e.g. the distinction between argument and adjunct, which is often fluid), whereas in other cases the reverse is true (e.g. particle attachment). Therefore we combine an automatic evaluation with a close reading of fresh, pre-parsed texts, allowing us to assess the strengths and weaknesses of a new model from a frog's eye perspective. Figure 1 shows a representative example of an automatically annotated sentence (viz. the first sentence of Xenophon's *Memorabilia*), with relatively good results. We are also planning to develop new ways to evaluate the test data and the improvement in parsing accuracy in a more detailed manner.

4 Output and applications

4.1 Creating treebanks

Instead of solely aiming to achieve better parsing accuracy, we specifically wanted to offer some tangible deliverables. This is why we extended the rather limited set of morphologically and syntactically annotated prose texts currently available in quantitative terms as well as in terms of its genre diversity, thus significantly increasing the quantity and quality of the training data. By making new trees (see Table 1 for an overview), developed in keeping with the guidelines of the Perseus Dependency Treebanks, we were able to make a detailed analysis of the strengths and weaknesses of the subsequent versions of the Ancient Greek parser under development. The detection of enduring shortcomings and parsing problems also reveals which issues should be prioritised in order to obtain better results. In addition, it allows us to uncover some inconsistencies present in the existing treebanks. The Leuven treebanks were not built from scratch, but on the basis of a pre-tagged and pre-parsed version, which considerably improved and accelerated the treebanking process. Part of these trees (all of which are beta versions) have served as test data in comparing the different parsers (Mercelis, 2019: see *supra*).

Author	**Details**	**Prose/Poetry**	**#Tokens**
Aesop	Select fables	Prose	7,5K
Anon.	Batrachomyomachia	Poetry	2,2K
Aristophanes	Thesmophoriazusae	Poetry	9K
Diverse authors	Papyrus texts	Prose	12K
Diverse authors	Pedalion example sentences	Prose & Poetry	20K
Euripides	Medea	Poetry	10K
Lucian	Prometheus, Symposion, Lis vocalium, Philopseudes 33-36, The Mule	Prose	21K
Lyric Poetry	Mimnermus, Theocritus, Semonides	Poetry	1,5K
Lysias	On the Pension (Or. 24)	Prose	1,5K
Menander	Dyskolos	Poetry	8K
Paeanius	Breviarium (parts of chapter 1)	Prose	6K
Prose authors	Longus, Isocrates, Hippocrates (Fragments)	Prose	1,5K
(Pseudo-)Plato	Cleitophon and Crito	Prose	5,8K
(Septuagint)	Parts of Genesis [For the part-of-speech annotation, we made thankful use of (Kraft, 1988).]	Prose	14K
Total			119K

Table 1: Overview of the recently produced treebanks (with approximate numbers of tokens).

We make use of the very user-friendly open-source Arethusa treebank editor, which is an intuitive tool for building and reviewing treebanks (see Figure 1 for an example). In future versions, we will have to pay particular attention to the metadata of our trees, which are currently rather poor.

Apart from offering manually checked treebanks, our project also encompasses automatically parsed data of ca. 37 million tokens. Given that these data could, despite all the errors inherent to the process, be of immediate interest to linguists of Ancient Greek and represent a syntactic 'sister' to *Perseus under*

PhiloLogic (see Dik, 2018), we will make the majority of these data available (copyright issues related to a number of texts included prevent us from publishing the corpus in its entirety). In this stage, we are happy to provide future annotators with pre-parsed versions of specific texts, so as to speed up the annotation process.

4.2 Correcting and modifying treebanks

By creating new treebanks on the one hand and by systematically assembling the data of existing treebanks on the other, we were able to trace inconsistencies and errors in existing treebanks of Ancient Greek. A survey of these modifications is published on our GitHub page, where the Readme file offers more information (http://github.com/pedalion/treebanks). The modifications are of various kinds. The number of what we believe are clear errors represent only a minor—although not unsubstantial—part of the file: most suggestions are made for purposes of homogenisation. As it is a work in progress, it is safe to say that this file might also contain a number of changes for the worse. The current release version contains modifications of ca. 120K tokens. These modifications have already been implemented in our own treebank query tool, DendroSearch, of which the functionality is outlined in the following section.

4.3 Querying treebanks: DendroSearch

Despite the abundance of treebank initiatives today, there are hardly any tools available which enable users to perform detailed queries in the treebanks. The Iliados tool (briefly mentioned in Mahony, 2016: 42) is restricted to a relatively small selection of poetic texts in the Perseus' Ancient Greek Dependency Treebanks. Annis, a tool that can query the Perseus Latin and Ancient Greek Treebank, has been offline since 2013, but recently a graph-based version was developed (see Krause, 2019). The PROIEL treebank can be queried through the INESS-tool (Rosén et al., 2012).

To encourage corpus-based research in the existing treebanks we developed DendroSearch, a stand-alone tool that is explicitly designed to query Greek treebanks in a user-friendly way. Through a series of panels, users can build complex queries and send them to a search system which goes through all available treebank material and presents the results (see Figure 2). For this tool we integrated all the corrections we made, all the conversions between annotation formats we implemented, as well as the treebanks that were produced by our research group, into the existing treebanks. We hope that the tool as well as the source code, which will be made available on GitHub, will be useful to other researchers currently developing treebank query initiatives. In future versions, visualisation capabilities could be improved so as to make querying the treebanks even more intuitive, and a number of basic statistical analytics (e.g. collocation and collostructional analysis) could be introduced. Additionally, a new version will encompass the possibility of performing semantic queries, which is the topic of the next section.

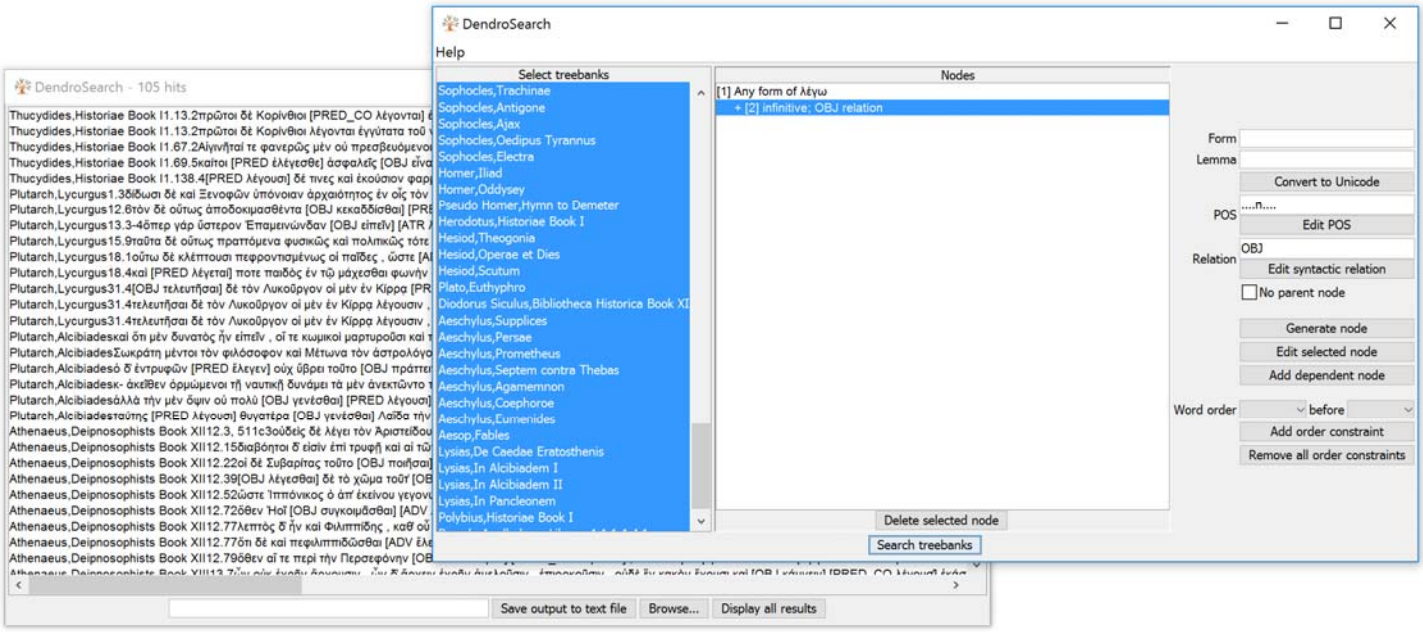

Figure 2: Screenshot of the DendroSearch stand-alone tool

4.4 Enriching treebanks: the role of semantics

We are currently experimenting with adding a semantic layer to the morphological and syntactic annotation (see Swaelens, 2019). The approach is twofold. On the one hand, we aim to assign a semantic hypernym to the lemmas of each noun (e.g. 'person', 'animal', 'non-concrete' etc.), verb (e.g. 'emotion', 'perception', 'stative' etc.) and adjective ('quantifier' vs. 'qualifier'). It is likely that this will further improve parsing accuracy (as some tests have indicated) and also enhance searching possibilities. In addition, we are also experimenting with distributional vectors of Greek lemmas, based on a large automatically annotated corpus of approximately 37 million tokens (see 4.1).

Alongside this lemma-based approach, we try to define the semantic role of adverbials and attributes. The underlying hypothesis is that for certain parts of a sentence the semantic role (e.g. standard of comparison, agent, possessor, direction, etc.) is more significant than the syntactic function (it is often very difficult to make a consistent distinction between adverbials and objects, for instance). Most of the semantic roles were added by student annotators, but we are also developing approaches which will do this automatically or semi-automatically. Semantic role labelling is present in a number of treebank initiatives (viz. the Perseus Ancient Greek Dependency Treebanks, see Celano and Crane, 2015, and the Harrington treebanks). Table 2 displays the semantic roles currently distinguished in the Pedalion project (mainly based on Crespo et al., 2003). Swaelens (2019: 32-34) includes a comparative table contrasting the use of semantic roles in the different Greek Treebank initiatives.

AGENT	DURATION	LOCATION	RESULT
BENEFICIARY	EXPERIENCER	MANNER	SOURCE
CAUSE	EXPLANATION	MATERIAL	TIME
COMPANION	EXTENT OF SPACE	PATIENT	TIME FRAME
CONCESSION	GOAL	POSSESSOR	TOTALITY
CONDITION	IDENTITY	PROPERTY	VALUE/PRICE
DEGREE/MEASURE	INSTRUMENT	RECIPIENT	
DIRECTION	INTERMEDIARY	RESPECT	

Table 2: Overview of the semantic roles currently distinguished in Pedalion.

4.5 Valorising treebanks work in a didactic context

In some recent papers the pedagogical value of making Ancient Greek and Latin treebanks has been highlighted (see e.g. Mambrini, 2016). So far, the focus has been on the educational benefits of treebank *creation*. Annotating a treebank implies close reading and making detailed morphological and syntactic analyses, which will considerably increase a student's awareness of the complexities and difficulties inherent in Ancient Greek syntax. While subscribing to this view, we also argue that treebanked texts can, and should, play a significant role as *products* and *tools* for receptive language learners as well. Our ongoing project implements three ways of valorising existing treebanks for educational purposes.

Needless to say, a first obvious application consists in offering reading support for treebanked texts or text fragments. The Perseids and Arethusa initiatives, already mentioned above, enable users to create treebanks with beautifully visualised trees of analysed sentences (see Figure 1 for an example). Through collaboration with Perseids and Arethusa staff members we were able to make use of their recently generated "Treebank Template" (https://github.com/perseids-publications/treebank-template), which also allows our users to browse through the trees in a convenient and user-friendly way (http://en.pedalion.org/reading).

A second application concerns vocabulary. *Chilia*, building on a frequency-based vocabulary of Ancient Greek (Van Hal, 2013), contains the 1000 most frequent lemmas found in Classical Ancient Greek texts. To some extent, its development be seen against the precarious backdrop of Ancient Greek studies in high schools in the Low Countries. Although Ancient Greek is still relatively well-represented in gymnasia programs in Flanders and the Netherlands, the number of pupils is too limited to attract much attention on the part of educational publishers. This explains why teachers are forced to make use of somewhat dated learning tools, which might contribute to a further decline of pupil numbers. On the other hand, this situation also creates the possibility to take the lead in creating Open Educational Resources tailored to the needs of high school pupils. *Chilia* is conceived of as a modest contribution in this direction. The novelty of *Chilia* consists in the fact that every single entry is accompanied by a short

real-life sentence (some of which are slightly abridged) which contextualises the lemma in question. Furthermore, all the sentences included contain only words which occur in the *Chilia* word list (with the exception of proper names). So, for instance, the lemma *pote* is accompanied by the following example sentence from the Athenian orator Andocides (c. 440–c. 390 BC): "ἦν γάρ **ποτε** χρόνος, ὦ Ἀθηναῖοι, ὅτε τείχη καὶ ναῦς οὐκ ἐκεκτήμεθα" ["**Once** there was a time, Athenians, when we had neither walls nor a fleet"]. Given that the other words in this sentence belong to the 1000 most frequent words as well, learners of Ancient Greek are enabled to study vocabulary in context and in a self-reinforcing way. Most sentences were selected by relying on Ancient Greek treebanks that exist today or by specifically searching—in a semi-automatic way—for sentences that meet the required conditions. *Chilia* will be published both as a stable e-publication (which can be downloaded in pdf-format) and in a dynamic online-environment, which will enable users to visualise the syntactic trees of the example sentences and to establish links to other online initiatives, such as *Logeion* (see Dik, 2019).

Treebanks will also play a role as an enhancement of an already existing Open Educational Resource, viz. the online modular grammar of Ancient Greek (Van Hal and Anné, 2017), the English version of which is still partly under construction. This grammar aims to overcome the static nature of traditional grammars by granting users the possibility to switch from the language's formal level to its semantic, syntactic, or pragmatic level and vice versa through principles of faceted search. The syntax encompasses a large number of original example sentences (many of which stem from post-classical authors, active in the Hellenistic or Roman period), the majority of which have been treebanked. By clicking on a specific example sentence, users can consult the syntactic tree.

5 Conclusions

This paper has presented ongoing work for the Pedalion-project at the University of Leuven. Reasons of space have prevented us from fully substantiating our methods and strategies, but we plan to do so in following publications. An important pillar of this work is the fruitful combination of several existing resources in order to (a) create new linguistically annotated data, (b) improve the quality of the existing data, (c) make the existing data easier to query for users with limited programming skills, (d) expand on the existing data and (e) valorise the data for pedagogical purposes. As for (a), we make use of state-of-the-art NLP technology to quickly create large amounts of new data. The advantage of this strategy is that is much faster to correct pre-tagged, pre-lemmatised and pre-parsed data than it is to create new treebanks from scratch. As for (b), the homogenisation of existing projects has numerous benefits, including improving the performance of the NLP technology and making it easier to compare its results, as well as making it easier to query these projects for linguistic information and to do so more reliably. As for (c), we have created a user-friendly tool to query the treebanks, DendroSearch, which will allow a broader audience to make use of the various research possibilities that the existing treebanks are already offering. As for (d), we have shown how we plan to add semantic information (at the lemma level as well as in terms of semantic roles) as a valuable supplementary layer for linguistic enquiries. Finally, as for (e), we have created and will continue creating a set of tangible deliverables with pedagogical purposes. As researchers involved in a project that gratefully makes use of the painstaking work done by other people in the scientific community, we also present this paper as a call to invite others to expand on our work (which will be made publicly available on GitHub) as well as discuss new future possibilities of collaboration.

Acknowledgements

We make grateful use of the large number of treebanks readily available. Our treebank data was created and edited with the help of the Arethusa application (https://github.com/alpheios-project/arethusa) as provided by the Perseids Project at Tufts University (https://perseids.org). (Arethusa is now being jointly maintained by the Perseids Project and The Alpheios Project, Ltd.) Since January 2019, our work is also partly funded through an FWO research grant (Research Foundation Flanders). We are especially indebted to the students who enthusiastically participated in our treebanking classes. Numerous colleagues were willing to assist us in our experiments, to comment on earlier drafts of this paper or to exchange ideas prior to the writing of this short paper. We would therefore like to thank Bridget Almas, Giuseppe Celano, Harry Diakoff, Zach Fletcher, Bob & Vanessa Gorman, Dag Haug, Francesco Mambrini, Merisa Martinez, Reuben Pitts, James Tauber, Demmy Verbeke, Marja Vierros and two anonymous reviewers.

References

Miguel Ballesteros and Joakim Nivre. 2012. “MaltOptimizer: An optimization tool for MaltParser.” In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 58–62. Association for Computational Linguistics, Stroudsburg. <https://www.aclweb.org/anthology/E12-2>

David Bamman and Gregory Crane. 2011. “The Ancient Greek and Latin Dependency Treebanks.” In *Language Technology for Cultural Heritage*, edited by Caroline Sporleder, Antal van den Bosch and Kalliopi Zervanou, 79–98. Springer, Berlin & Heidelberg.

Giuseppe G. A. Celano and Gregory Crane. 2015. “Semantic Role Annotation in the Ancient Greek Dependency Treebank.” In *Proceedings of the Fourteenth International Workshop on Treebanks and linguistic Theories (TLT14)*, edited by Markus Dickinson et al., 26–34. Institute of Computer Science, Warsaw. <http://tlt14.ipipan.waw.pl/files/4614/5063/3858/TLT14_proceedings.pdf>

Giuseppe G. A. Celano, Gregory Crane, and Saeed Majidi. 2016. “Part of Speech Tagging for Ancient Greek.” *Open Linguistics* 2(1):393–399. doi:10.1515/opli-2016-0020.

Emilio Crespo, Luz Conti, and Helena Maquieira. 2003. *Sintaxis del griego clásico*. Gredos, Madrid.

Helma Dik. 2018. *Perseus under PhiloLogic*. <http://perseus.uchicago.edu/>

Helma Dik. 2019. *Logeion.* <https://logeion.uchicago.edu/>

Vanessa Gorman. 2016. *Gorman Treebanks*. <https://history.unl.edu/vanessa-b-gorman> and <https://github.com/rgorman/author_attribution/tree/master/vg_combined_trees2>

Matthew Harrington. 2018. *Perseids Project - Treebanked Commentaries at Tufts University*. <https://perseids-project.github.io/harrington_trees/>

Dag T. T. Haug. 2010. “PROIEL Guidelines for Annotation” <http://folk.uio.no/daghaug/syntactic_guidelines.pdf>.

Dag T. T. Haug. 2014. “Computational Linguistics and Greek.” In *Encyclopedia of Ancient Greek Language and Linguistics*, edited by Georgios K. Giannakis, 1:354–356. Brill, Leiden & Boston.

Dag T. T. Haug and Marius Jøhndal. 2008. “Creating a Parallel Treebank of the Old Indo-European Bible Translations.” In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, edited by Caroline Sporleder and Kiril Ribarov, 27–34. <https://www.hf.uio.no/if-ikk/english/research/projects/proiel/Activities/proiel/publications/marrakech.pdf>

Erik Henriksson and Marja Vierros. 2017. “Preprocessing Greek Papyri for Linguistic Annotation.” *Journal of Data Mining & Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages.* < https://hal.archives-ouvertes.fr/hal-01279493v1/document>

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. “Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task.” In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, 133–142. Association for Computational Linguistics, Brussels. <http://www.aclweb.org/anthology/K18-2013>

Alek Keersmaekers. 2019. “Creating a Richly Annotated Corpus of Papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language.” *Digital Scholarship in the Humanities* (online): 1–16. doi:10.1093/llc/fqz004.

R. Kraft. 1988. *Morphologically Analyzed Septuagint (version 1.0)*. Computer-Assisted Tools for Septuagint Studies (CATSS). University of Pennsylvania. <http://ccat.sas.upenn.edu/gopher/>

Thomas Krause. 2019. *ANNIS: A graph-based query system for deeply annotated text corpora*. PhD-Dissertation Humboldt Universität zu Berlin. <https://edoc.hu-berlin.de/handle/18452/20436>.

John Lee, Jason Naradowsky, and David A. Smith. 2011. “A Discriminative Model for Joint Morphological Disambiguation and Dependency Parsing.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human language technologies*, 1:885–894. Association for Computational Linguistics, Stroudsburg. < https://www.aclweb.org/anthology/P11-1089>

Simon Mahony. 2016. “Open Education and Open Educational Resources for the Teaching of Classics in the UK.” In *Digital Classics Outside the Echo-Chamber: Teaching, knowledge exchange & public engagement*, edited by Gabriel Bodard and Matteo Romanello, 33–50. Ubiquity Press, London. <https://oapen.org/search?identifier=649985>

Francesco Mambrini. 2016. “The Ancient Greek Dependency Treebank: Linguistic annotation in a teaching environment.” In *Digital Classics Outside the Echo-Chamber: Teaching, knowledge exchange & public engagement*, edited by Gabriel Bodard and Matteo Romanello, 83–99. Ubiquity Press, London. doi:10.5334/bat.f.

Francesco Mambrini and Carlo Passarotti. 2012. “Will a Parser Overtake Achilles? First experiments on parsing the Ancient Greek Dependency Treebank.” In *Eleventh International Workshop on Treebanks and Linguistic Theories*, 133–144. Edições Colibri. <https://pdfs.semanticscholar.org/b5b5/4154385697b29fe3d9f0fce080c7d34525cd.pdf>

Wouter Mercelis. 2019. *Syntactisch parsen van Oudgriekse teksten: een vergelijkende studie.* Unpublished master thesis KU Leuven.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. "MaltParser: A language-independent system for data-driven dependency parsing." *Natural Language Engineering* 13(2):95–135. doi:10.1017/S1351324906004505

Jonathan Robie. 2017. "Nine Kinds of Ancient Greek Treebanks." *Open Data for Digital Biblical Humanities*. <http://jonathanrobie.biblicalhumanities.org/blog/2017/12/20/treebanks-for-ancient-greek/>

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. "An open infrastructure for advanced treebanking." In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012, Istanbul, Turkey, May 2012*, edited by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, 22–29. European Language Resources Association (ELRA), Istanbul. <https://ling.w.uib.no/files/2013/02/lcrec2012at-iness-paper-published.pdf>

Ariel Rosenfeld and Sarit Kraus. 2018. *Predicting Human Decision-making: From prediction to action*. Morgan & Claypool, s.l.

Piotr Rybak and Alina Wroblewska. 2018. "Semi-Supervised Neural System for Tagging, Parsing and Lemmatization." *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*: 45–54. Brussels: Association for Computational Linguistics, Brussels. <http://www.aclweb.org/anthology/K18-2004>

Natalie Schluter and Josef van Genabith. 2007. "Preparing, Restructuring, and Augmenting a French Treebank: Lexicalised parsers or coherent treebanks?" In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 200–209. <http://doras.dcu.ie/15265/1/78_Paper_meta.pdf>

Colin Swaelens. 2019. *De rol van semantiek bij de automatische zinsanalyse van het Grieks*. Unpublished master thesis KU Leuven.

Toon Van Hal. 2013. *Ankura. Basiswoordenlijst Oudgrieks*. Garant, Antwerpen & Apeldoorn.

Toon Van Hal and Yannick Anné. 2017. "Reconciling the Dynamics of Language with a Grammar Handbook. On Pedalion, an ongoing Greek grammar project." *Digital Scholarship in the Humanities* 32(2):448–454.

Polina Yordanova. 2018. *Treebank of Aphtonius, Progymnasmata*. <https://github.com/polinayordanova/Treebank-of-Aphtonius-Progymnasmata>

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. "CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies." In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, 1–21. Association for Computational Linguistics, Brussels. <https://www.aclweb.org/anthology/K18-2001>