The GLAUx corpus: methodological issues in designing a long-term, diverse, multilayered corpus of Ancient Greek

Alek Keersmaekers
KU Leuven
alek.keersmaekers@kuleuven.be

Abstract

This paper describes the GLAUx project ("the Greek Language Automated"), an ongoing effort to develop a large long-term diachronic corpus of Greek, covering sixteen centuries of literary and non-literary material annotated with NLP methods. After providing an overview of related corpus projects and discussing the general architecture of the corpus, it zooms in on a number of larger methodological issues in the design of historical corpora. These include the encoding of textual variants, handling extralinguistic variation and annotating linguistic ambiguity. Finally, the long- and short-term perspectives of this project are discussed.

1 Introduction

The increasing availability of large-scale corpus resources has had a lasting impact on the field of linguistics. In the field of corpus linguistics, large quantities of data have made it possible to precisely model complex multifactorial processes of linguistic change (e.g. Perek and Hilpert, 2017; Gries et al., 2018). Modern methods in natural language processing also increasingly make use of word embeddings, which encode rich information about the use of a word learned from large datasets (Collobert et al., 2011; see Kutuzov et al., 2018 for diachronic word embeddings).

From a diachronic perspective, the Greek language corpus is an ideal candidate for a large-scale corpus-linguistic approach: it is not only one of the longest preserved languages (with a large body of text already in the 8th century BC, and continuing up until the present day), but it also is extremely well-documented: the *Thesaurus Linguae Graecae* library of Ancient Greek literary texts, for example, contains more than 110 million words (Pantelia, 2021). To make such an approach possible, this paper will describe *GLAUx* ("the Greek Language Automated"), a project aiming to collect a large corpus (spanning sixteen centuries) of Ancient Greek texts from various sources and to automatically annotate this corpus for rich linguistic information.

The construction of such a long-term historical corpus is obviously not a trivial task. The goal of this paper is therefore to describe the central problems encountered during this endeavor and the approaches currently adopted to tackle these problems. This will be discussed in section 3, after giving an overview of the data and annotation layers in section 2. Finally, section 4 will give an outlook of future work for this (long-term) project.

2 The corpus

2.1 Text types

Greek texts are usually classified into three categories: literary, papyrological and epigraphical texts. 'Literary' texts are typically transmitted to us through the manuscript tradition. Papyrological and epigraphical texts are written on soft materials such as papyrus and hard materials such as stone respectively, and preserved in their original state. This material dimension also correlates with a genre dimension: 'literary' texts were considered important enough by medieval monks to be copied, and include a wide range of texts (usually, but not always, in a high register), i.e. not only creative text genres such as poems and narrative prose, but also e.g. scientific texts, oratory, philosophical and religious texts. 'Papyrological' and 'epigraphical' texts include all sorts of everyday writing, including letters, receipts and petitions (typically written on papyrus) or texts that are meant to be

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 39–50
August 6, 2021. ©2021 Association for Computational Linguistics

more durable, e.g. decrees, epitaphs and honorary inscriptions (typically written on stone).[1]

2.2 Related work

Corpus work for Greek started with the development of the *Thesaurus Linguae Graecae* (TLG) in 1972, a full text library of literary texts, currently spanning the 8th century BC to the 15th century AD (Pantelia, 2021). While this project undoubtedly contains the largest collection of Greek text to date (more than 110 million words), and also includes high-accuracy lemmatization, its texts are not publicly available but can only be accessed through a search engine, heavily restricting any possibilities for serious corpus linguistic research as a consequence. An open source alternative is the *Perseus Digital Library* (Crane, 2021) and the *First One Thousand Years of Greek* project (First1K; Crane et al., 2021), both now included in the international *Open Greek and Latin* project. In comparison with the *TLG*, however, their coverage is more restricted (in total about 31 million Greek words, and most texts are situated before the 4th century AD) and the texts are often based on older editions. Non-literary Greek texts are made available by the *Packard Humanities Institute* [2] (epigraphy; see Iversen, 2007) and the Integrating Digital Papyrology project (papyri; Cayless et al., 2021).

While the projects mentioned above only include the full text, there have also been some efforts to add linguistic annotation. A wide variety of treebanking projects have manually annotated Greek texts for morphology, lemmas, (dependency) syntax and sometimes semantics, most prominently the *PROIEL* project (Haug and Jøhndal, 2008; 277,000 tokens), the *Ancient Greek Dependency Treebanks* (AGDT; Bamman et al., 2009; 560,000 tokens), the *Gorman* trees (Gorman, 2020; 324,000 tokens) and the *Pedalion* project (Keersmaekers et al., 2019; 320,000 tokens), as well as some smaller projects (in total, the manually annotated work includes about 1.5 million tokens). The former two projects are also included in the *Universal Dependencies* (UD) project (Nivre et al., 2016).

There have also been some efforts to annotate even larger amounts of data automatically. Celano (2017) lemmatized and tagged the data of *Perseus* and *First1K* with *Mate* tagger (Bohnet and Nivre, 2012), achieving an accuracy of 88%. The *Diorisis* corpus (Vatri and McGillivray, 2018), including texts from the *Perseus* project and some online sources (about 10 million tokens), was lemmatized and tagged with *TreeTagger* (Schmid 1994), with an accuracy of 91%. There have also been attempts to automatically analyze the papyrus corpus: Celano (2018) achieved a tagging accuracy of 62% and a lemmatization accuracy of 47%, while Keersmaekers (2020b) achieved a morphological tagging accuracy of 95%, 99% for lemmatization, 85% for syntactic parsing and 81% for semantic role labeling. All these automatically analyzed datasets are openly available online.

2.3 Source texts

The source texts for the papyrological part of *GLAUx* and the (planned) epigraphical part are both collected in a single repository (see the previous section). The literary texts, in contrast, are more scattered: while the *TLG* has the most exhaustive collection, its source materials are not publicly available. A large part of the literary corpus has been made available by the *Open Greek and Latin* project (see the previous section), while additional texts can be found on a number of web sources[3] (e.g.). Table 1 gives an overview of the source texts included in the *GLAUx* corpus.

	Tokens	**Source**
Literary	23.2 million; more to be added	*Open Greek and Latin*; Web
Papyrological	4.5 million	*Integrating Digital Papyrology*
Epigraphical	3.2 million (to be added)	*Packard Humanities Institute*

Table 1: Text types of the *GLAUx* corpus.

[1] To refer to texts written on papyrus but that thematically fit better in the literary corpus, the term 'literary papyri' is typically used (while the everyday texts in the papyrus corpus are often called 'documentary papyri'). This paper will use the terms 'literary', 'papyrological' and 'epigraphical' as a genre indicator, i.e. all 'literary' texts, whether transmitted through the manuscript tradition or written on papyrus are called 'literary', while the term 'papyri' is reserved for the documentary papyri.

[2] https://inscriptions.packhum.org

[3] E.g. https://el.wikisource.org; https://www.hs-augsburg.de/~harsch/augustana.html; https://penelope.uchicago.edu/Thayer/E/Roman/Texts/home.html.

In terms of chronology, almost all texts of the papyrological corpus are from the third century BC to the eighth century AD (this is related to the Greek rule of Egypt, where most papyri are found). The epigraphical corpus can generally be dated from the fourteenth century BC to the seventh century AD. The boundaries of the literary corpus are more difficult to define: while it starts with the Homeric poems in the eighth century BC, an end date is more difficult to settle on, as Greek was still widely used until the fall of the Byzantine empire – and continued to be used afterwards (obviously, Greek is still a living language). For the *GLAUx* corpus, we set its boundaries at the eighth century BC to the eighth century AD, so that literary and non-literary texts would be attested in the whole period, while it still contains sixteen centuries of Greek.[4] At the moment of writing, a first version of the full papyrological data as well as an experimental version of the literary data up to the second century AD has been released on GitHub;[5] since the epigraphical corpus has some unique challenges (in particular the high degree of dialectal variation and the lack of epigraphical training data: see 3.1 and 3.2), we plan to add it to *GLAUx* in the long term (see also Dell'Oro and Celano, 2019 for a discussion of some specific issues involved with these texts).

The literary texts have a wide range of text genres, including poetic texts (epic poetry, lyric poetry, tragedy and comedy), philosophic and scientific prose (e.g. medicine, mathematics, geography), historical texts, rhetorical texts, commentaries, religious texts, biographies, narrative fiction, and various other smaller genres. *GLAUx* generally follows the genre classification of the *TLG* in a simplified format (i.e. I assigned just one genre to each text, instead of multiple genres as is sometimes the case with the *TLG* texts), although this classification will be revised in the future to maximize its usefulness for automated processing purposes (see 3.2) and its interoperability with other resources (e.g. the genre classification of the *Diorisis* corpus).[6] For the papyri, the *GLAUx* corpus follows the classification of the *Trismegistos* project (Depauw and Gheldof, 2014), developed by Joanna Stolk, which includes letters, petitions, contracts, lists, receipts, labels, pronouncements, declarations, reports, accounts, other administrative texts, judicial texts and paraliterary texts as the main text genres.

2.4 Annotation

Due to the size of the *GLAUx* corpus (currently more than 27 million tokens), all annotation was necessarily carried out automatically, building on methods developed by Keersmaekers (2020b) for the papyrus corpus. The treebank data discussed in section 2 was used as training data, which include a wide range of periods and text genres (although most of the data is literary).

Part-of-speech and morphology: For morphological and part-of-speech tagging *RFTagger* was used (Schmid and Laws, 2008), a HMM-based tagger using decision trees to estimate contextual probabilities (as well as suffix features to decide on the lexical probabilities of unknown words). Its output can be constrained by a lexicon that provides possible morphological analyses for each word form, in which case only the morphological analyses present in the lexicon are considered as possible part-of-speech/morphological tags – for this a morphological lexicon generated by the rule-based *Morpheus* morphological analysis tool (Crane, 1991) was used. Prediction accuracy (for the full tag combining part-of-speech and morphological information) ranged from 0.908 (philosophical treatises) to 0.961 (biblical texts), with an average prediction accuracy of 0.945. In terms of text genre, orations, papyri and epic poems also have high accuracy rates, next to biblical texts, while, next to philosophical treatises, comedies, lyric poems and tragedies also have a low accuracy rate (see Table 2).

The morphological annotation is consistent with the (2.0 version) tag set of the *AGDT* (see 2.2). The morphological categories are person, number, tense/aspect, mood, voice, gender, case and degree. Of these, gender, case, tense and mood have the

[4] In terms of important linguistic developments, this data includes the Archaic period (8th-6th century BC), which mainly encompasses poetic texts written in a variety of dialects; the Classical period (5th-4th century BC), in which the Attic dialect spoken in Athens became the prestige language of literary texts; the Koine period (3rd century BC-4th century AD), in which Greek became a standardized language since the conquests of Alexander the Great; and the early Medieval period (from the 5th century AD onwards), when Greek was mainly used in the Byzantine empire.

[5] https://github.com/perseids-publications/glaux-trees

[6] See https://perseids-publications.github.io/glaux-trees/ for the current *GLAUx* texts classified by genre.

	Accuracy (N)
Biblical	0.961 (33,994)
Military	0.959 (3,234)
Oratory	0.952 (22,699)
Papyri	0.951 (8,166)
Epic Poetry	0.951 (49,694)
Biography	0.948 (12,265)
History	0.946 (81,560)
Philosophical Dialogue	0.944 (4,146)
Dialogue	0.943 (1,132)
Epistolography	0.941 (1,261)
Narrative Fiction	0.939 (9,883)
Rhetoric	0.937 (3,768)
Polyhistory	0.929 (9,154)
Tragedy	0.924 (21,421)
Lyric Poetry	0.921 (1,084)
Comedy	0.920 (5,640)
Philosophical Treatise	0.908 (9,239)

Table 2: Tagging accuracy by genre.

	Accuracy (N)
Degree	0.995 (49,374)
Number	0.990 (164,492)
Voice	0.987 (48,913)
Part-of-speech	0.985 (278,344)
Person	0.977 (27,728)
Mood	0.970 (27,728)
Tense	0.968 (48,913)
Case	0.959 (136,764)
Gender	0.958 (136,764)

Table 3: Tagging accuracy by morphological attribute.

lowest prediction accuracy (see Table 3), since they include many ambiguous forms (in particular between neuter and masculine, between nominative and accusative and between indicative and subjunctive). Part-of-speech classes are divided into the traditional classes of nouns, adjectives, verbs, adverbs, pronouns, conjunctions, prepositions, numerals, articles and interjections. Since Greek makes a morphological distinction between verbs, nouns, adjectives and uninflected words, these categories are also relatively easy to handle for the tagger (with a 0.985 accuracy for part-of-speech only).

Currently I am also expanding the morphological annotation with a derivational annotation layer, linking complex morphological derivations (e.g. παιδίον paidíon "little child") to a stem or root (e.g. παιδ- *paid-*, used in the word παῖς *pais* "child") and morphological pattern (e.g. *-ion* diminutives), which will further expand linguistic research possibilities for end users: see Litta et al. (2019) for comparable work for the Latin language.

Lemmas: The data was lemmatized with *Lemming* (Müller et al., 2015), a log-linear model of lemmatization making use of formal (edit trees between form and lemma, as well as affixes), lemma, part-of-speech and morphology and dictionary features (i.e. whether the lemma occurs in a list of pre-defined lemmas: for this I used the *Liddell-Scott-Jones* (LSJ) lexicon of Greek; Jones et al., 1996). Lemmatization accuracy was 0.969 initially; I was able to increase this to 0.980 by again using a *Morpheus* lexicon as a constraint, i.e. by restricting the output of *Lemming* to lemmas recognized by *Morpheus* as a valid lemma for the given Greek form/morphology-combination (if the form was recognized by *Morpheus*: otherwise, *Lemming* could freely decide upon a possible lemma). These results are higher than the state-of-the-art reported in Vatri and McGillivray (2020),[7] but the high accuracy is not completely unexpected, since in most cases only one option is possible due to the morphological complexity of Greek words. Accordingly, words that are not recognized by *Morpheus* have a significantly lower lemmatization accuracy (0.812). For the poetic data, lemmatization accuracy is a little lower than the prose data: accuracy ranges from 0.965 (comedies) to 0.975 (epic poetry) for the poetic data, while most prose genres have an accuracy of more than 0.980 (with oratory and biblical texts on the high end): see Table 4. The lemmas are generally consistent with the *LSJ* lexicon as well as

[7] They report lemmatization accuracies of 0.91 for a part of book 1 of the *Iliad* (with the CLTK backoff lemmatizer) and of 0.97 for *Lysias*, speech 7 (with the Diorisis lemmatizer). While the test set is different, *Lemming*'s lemmatization accuracy is 0.974 for the whole of the Iliad and 0.990 for all the Lysias data included in our treebank material. The results are not entirely comparable, however: our training set is different than the data that the tools used by Vatri and McGillivray (2020) are trained on, and we used the treebank material rather than our own annotation (as Vatri and McGillivray did) as a gold standard.

	Accuracy (N)
Biblical	0.989 (29,713)
Oratory	0.987 (19,876)
Dialogue	0.986 (998)
Biography	0.985 (10,655)
Military	0.985 (2,898)
Philosophical Dialogue	0.982 (3,576)
History	0.982 (73,278)
Rhetoric	0.981 (3,276)
Epistolography	0.980 (1,101)
Philosophical Treatise	0.980 (8,132)
Narrative Fiction	0.979 (8,392)
Epic Poetry	0.975 (42,836)
Papyri	0.972 (7,268)
Tragedy	0.972 (18,027)
Polyhistory	0.969 (8,095)
Lyric Poetry	0.967 (928)
Comedy	0.965 (4,650)

Table 4: Lemmatization accuracy by genre.

the lemmas included in the *Morpheus* codebase (which is largely based on *LSJ*).

Syntax: The *GLAUx* corpus was also annotated with dependency information consistent with the *AGDT* (2.0) guidelines, which are based on the annotation format of the *Prague Dependency Treebanks* (Böhmová et al., 2003). For this task the *Stanford Graph-Based Dependency Parser* (Dozat et al., 2017) proved suitable, a biaffine neural (LSTM) graph-based parser making use of character, token and part-of-speech embeddings. This parser was able to achieve a 0.845 labeled attachment score (LAS) for the papyri and a LAS ranging from 0.751 (philosophical and scientific prose) to 0.881 (biblical texts) for literary texts depending on text genre. Several remaining problems are caused by inconsistencies in the training and/or test data, which may be resolved by homogenization efforts (which we have already carried out in the past, and which we will also further carry out in the future). While the *AGDT* annotation format was used for historical reasons (most treebank projects of Greek are based on this format), in the future we plan to move to *UD* (Nivre et al., 2016), which is the annotation standard that is currently widely supported by the broader NLP community.

Semantics: Finally, *GLAUx* also includes semantic role annotation. For this task we had to develop our own annotation standard and training data, since there was relatively little semantically annotated data available, and the tag set of the *AGDT* for semantic annotation (Celano and Crane, 2015) was too fine-grained for automatic prediction and based on an old reference grammar that is not up-to-date with modern linguistic theory.[8] As the annotation was mainly done by job students, the semantic roles were based on the roles they were accustomed to, i.e. the ones developed for the pedagogical Pedalion project (Van Hal and Anné, 2017). However, this role set was expanded and revised to be compatible with a number of frameworks used for other languages as well (the description of arguments in particular remains rather underdeveloped in the Pedalion grammar), most importantly *VerbNet* (Kipper-Schuler, 2005) and *LIRICS* (Petukhova and Bunt, 2008). Currently 34 roles are distinguished (agent, beneficiary, cause, companion, comparison, concession, condition, degree, direction, duration, experiencer, extent of space, frequency, goal, identity, instrument, intermediary, location, maleficiary, manner, material, patient, possessor, property, recipient, respect, result, source, stimulus, theme, time, time frame, totality, value). For this purpose the semantic role labeler developed by Keersmaekers (2020a) was used, which makes use of a Random Forest classifier over a wide range of features (most importantly formal characteristics of the target word, its syntactic label, and lemma vectors of the target word and its head). This method was able to achieve an accuracy ranging from 0.687 for poetic texts to 0.838 for religious texts, with a relatively low number of training examples (about 12,500).

3 Problems

3.1 Text preservation and encoding textual variants

Many Greek texts have an intricate transmission history. Literary texts are typically transmitted through centuries of copying by medieval monks.

[8] For Latin, the *Index Thomisticus Treebank* also includes semantic role annotation based on the tectogrammatical layer of the Prague Dependency Treebanks (Passarotti, 2014), which is considerably more detailed than the role set used here (distinguishing 67 ‘functors’).

Consequently, these texts do not have one version but multiple ones, as presented in the critical apparatus of the texts. Ideally, this critical apparatus would be directly encoded in the corpus, i.e. multiple versions of the same text would be aligned and each of these versions would be linguistically analyzed. In doing so, researchers will immediately know when encountering an unusual syntactic pattern whether there are any alternative readings or not (and, for example, be able to check whether the frequency of specific patterns remains the same if only words without alternative readings are taken into account). Unfortunately, the texts included in the *GLAUx* corpus are from a variety of sources that rarely include the critical apparatus. If more digital editions of critical apparatuses become available in the future, the quality of the *GLAUx* data will certainly be improved, but in the meantime *GLAUx* users should be aware that the underlying data is not always perfect (and might include some medieval alterations rather than actual language use in some cases).

The situation is different for the papyrological and epigraphical corpus, for which we have the original text as it was written in antiquity. This is not to say that no textual criticism is involved: firstly, some parts of the text may be harder to interpret or be entirely lost due to physical damage to the text material, in which case the interpretations of the editors of what this missing text should be (if such an interpretation is possible) can be considered a suggestion with which not everyone may agree. Secondly, the papyrological and epigraphic corpora have considerable spelling variation. For the papyrus corpus, editors usually standardize the spelling of papyrus texts, and these standardizations are included with the original forms in the XML version of the digital edition. For *GLAUx* we preserved both the 'original' and the 'standard' version for each word in the corpus (i.e. for a word like ἔχι which is an irregular spelling of ἔχει, both the forms 'ἔχι' and 'ἔχει' are included in the corpus). We based our automatic analysis on the standard version (in this case ἔχει), as the NLP tools we used were able to handle this version better (see also Keersmaekers 2020b: 12-14).

In addition, editors also often standardize morphology based on a classical norm, in which case performing the automatic analysis on the standard version is not advisable. In (1), for example, Μάρων (*Márōn*) is standardized by the editor to Μάρωνος (*Márōnos*). This is not based on phonological criteria, as there are no phonological reasons to omit the syllable -*os* at the ending of a word: rather the editor standardized the nominative Μάρων to the genitive Μάρωνος, as this case is normally expected after the preposition παρά (*pará*) "from". Labeling this word as a genitive based on the standard version would therefore misrepresent the case as it is actually used by the writer (which might be interesting from a diachronic perspective). Based on the lemma and morphological classification of the standard version, we therefore developed a rule-based system to generate this 'original' morphological information (e.g. when the standard version is a genitive on -ωνος and the original version is on -ων, and we know that the lemma belongs to the paradigm of words on -ων that have their genitive on -ωνος, we know that the correct case for the original version is a nominative).[9]

(1) ἀπέσταλκα δέ σοι τὸ δεῖγμα τοῦ παρὰ **Μάρων** (standardized to **Μάρωνος**) (P. Col. 3 51)
apéstalka dé soi tó deîgma toû pará **Márōn** (standardized to **Márōnos**)
I've sent you the sample from **Maron**.

Nevertheless, in some cases it is more difficult to decide whether we are dealing with phonological or morphological standardization: in (2), the use of the genitive σου (sou) where the editor expects the dative σοι (soi) – the standard expression of the recipient in Classical Greek – might be related to changes in case usage, but a phonological reason for the use of σου can also not be excluded, since the sounds of σου (/su/) and σοι (/sy/) are phonetically close to each other. For the current version of *GLAUx* we decided to include both a morphological analysis based on the original version (e.g. genitive in this example) and standard version (e.g. dative in this example), and leave a further classification which of these 'problems' are related to phonology and which ones to morphology for future research.

[9] This system builds on the work of Depauw and Stolk (2014), who have classified editorial regularizations for the papyri into broader categories (e.g. "ων instead of ωνος").

(2) δὸς τῷ κομείζοντί **σου** (standardized to **σοι**) τὴν ἐπιστολὴν (P. Oxy. 2 96)
dós tô komeízontí **sou** (standardized to **soi**) tḕn epistolḗn
Give to the person who has brought **you** the letter (…)

3.2 Extralinguistic variation

The Greek corpus is extremely diverse genre-wise, covers an extremely long time span, and the epigraphic corpus in particular also has considerable dialectal variation. This is, in the first place, a problem for automatic annotation: it is well known in NLP that accuracy drops when trying to analyze out-of-domain data, i.e. data that differs considerably from the training data. Not all these factors might be equally problematic: for the computational modelling of Greek lexical meaning, for example, McGillivray et al. (2019) found that genre is a more important factor than time, and argue that "literary Classical Greek is conservative when it comes to lexical semantics" (I also found similar results in my own experiments with meaning processing: see Keersmaekers 2020b: 119). As a complicating factor, there is a complex interplay between genre, diachrony and dialectal variation in literary Greek: some examples include Atticistic tendencies in post-classical Greek texts (i.e. imitating the prestige Athenian language variant of the fifth century BC) or the use of regional coloring tied to specific text types (e.g. the use of the Doric dialect in the chorus of tragedies, or an imitation of the Homeric dialect, which is already a mix of different dialects itself, in late epic poems).

There are several possible solutions to deal with this problem. One obvious solution is diversifying the training data. It has been shown by experiments on morphological tagging (Dik and Whaling, 2008) and syntactic parsing (Mambrini and Passarotti, 2012) of Ancient Greek that the quality of automatic annotation will significantly improve using in-domain data – similarly, I found that even a very small amount of papyrological training data could significantly improve the results for the automated analysis of these texts (Keersmaekers 2020b: 33). For the *Pedalion* treebanks which were included in the training data, we therefore aimed to include a variety of text types which are less well represented by the major treebanking projects (especially post-classical material), ranging from mathematical texts to private letters to horror stories.

Additionally, standardizing the training and/or test material during automatic analysis may also often lead to better results (see Piotrowski, 2012: 87): we have also taken some steps in this direction (see the use of standardized spelling as discussed in the previous section). Finally, in NLP several techniques have been developed to deal with out-of-domain labeling (e.g. Blitzer et al., 2006, Schnabel and Schütze, 2014). For syntactic parsing, I will experiment with the use of treebank embeddings (Stymne et al., 2018) in the future, which have shown to handle heterogeneous data well. An open question with the use of these techniques is which texts constitute the given domain that our NLP models should be adapted to (i.e. given a certain text type such as papyrus letters, which training data should be considered 'in-domain' and which 'out-of-domain'), given the complex interactions between genre, diachrony and dialect outlined above. Possibly text similarity measures (see Turney and Pantel, 2010) may provide valuable insights in this respect.

A more fundamental question is whether it is advisable to use a single annotation format for such a diverse corpus. On the one hand, several NLP projects such as *UD* (Nivre et al., 2016) have developed an annotation format for even broader purposes (covering all natural languages), and one could argue that the categories used in part-of-speech tagging and syntactic parsing are broad enough not to be affected by language variation too much (while semantic annotation should, ideally, be universal). On the other hand, the *GLAUx* corpus includes a large number of 'languages' as 'Greek', which may in some cases very strongly differ from each other (e.g. the language of mathematical texts vs. epic poems): researchers such as Haspelmath (2010) and Croft (2013) have also argued against the generalizability of linguistic categories. In a practical sense, this issue might be resolved by detailed annotation documentation of constructions that are highly peculiar to a particular text genre: expanding the manually annotated treebank data to more 'unusual' text genres, as discussed above, is obviously highly beneficial for identifying such constructions.

3.3 Linguistic ambiguity and historical change

It is well known that linguistic ambiguity is an important factor in diachronical change: change often happens in ‘bridging contexts’, i.e. contexts that are ambiguous between two constructions (Heine, 2002; Eckardt, 2006; Traugott, 2012). For example, the Greek word ἵνα (*hina*) develops from a conjunction introducing a purpose clause, as in (3), to a complementizer, as in (4). Ambiguous examples such as (5), in which the ἵνα-clause could either be interpreted as a complement clause or a purpose clause, may have caused this change. At any rate, such examples are highly problematic for the annotation format of the *AGDT*, in which a strict distinction is made between complement clauses and adverbial clauses.

(3) ἐντεῖλαι περὶ τούτου Κράτωνι **ἵνα** μὴ πάλιν σκυλῇτε με ἀναβῆναι πρὸς ὑμᾶς. (P. Strasb. 5 346)
enteîlai perí toútou Krátōni **hína** mḗ pálin skulête me anabênai prós humâs.
Give orders for this to Kraton, **so that** you do not force (?) me again to come to you.

(4) Ὠφελίωνι ἐνετειλάμην **ἵνα** καὶ αὐτὸς δοῖ ἑτέραν καὶ τοὺς ἄρτους μοι πέμψηι. (P. Ryl. 2 229)
ōphelíōni eneteilámēn **hína** kaí autos doî hetéran kaí toús ártous moi pémpsēi.
I have ordered Ophelion **to** give you another one and to send me the loaves of bread.

(5) ἔντειλαι τῶι παρά σου, **ἵνα** τὸ τάχος γέ[νη]τ̣αι. (PSI 4 326)
énteilai tōi pará sou, **hína** tó tákhos génētai.
Give commands to your messager “**in order that** there will be haste” or “**that** there should be haste”

When performing automatic annotation, such ambiguities may be reflected in the underlying probabilities of the natural language processing model: example (5) shares features both of a prototypical adverbial clause (e.g. unlike in (4), the subject of the ἵνα-clause and the recipient of the command are different entities) and a prototypical complement clause (the verb ἐντέλλω *entéllō* “command” typically requires an argument expressing the command), which should in principle be learnable by a NLP system if the relevant features are annotated. Hence when automatically labelling clauses for the adverbial/complement distinction, I found that clauses with high predicted probabilities of being a complement showed very prototypical features of complement clauses and vice versa for adverbial clauses (although the cases with ‘in-between’ probabilities showed a mix of complement, adverbial and ambiguous examples: see Keersmaekers 2020b: 158-174 for more detail). While corpus projects often simply only include the most probable label in their annotation, this underlying probability distribution may offer valuable information to detect such ‘less prototypical’ cases (although the output probabilities are obviously highly dependent on the quality of the automatic technique and the feature set it is provided with). For reasons of transparency I will therefore make as much information about the automatic prediction publicly available as possible.[10]

4 Conclusion and outlook

This paper has described *GLAUx*, an ongoing project aiming to compile a large and diverse corpus of historical Greek. A test version of this corpus has already been released on GitHub:[11] we aim to release a first version including all the papyrus data and the literary data until the fourth century AD in the course of 2021. I identified some important issues in constructing this corpus, and suggested a number of possible solutions: these include the encoding of textual variants, dealing with a high degree of extralinguistic variation and annotating ‘ambiguous’ constructions. These issues should be highly relevant for other researchers working with historical corpora, and I hope that this discussion may inspire further research.

The annotation of this corpus will be continuously improved in the coming years, as it is put to work in several research projects at the KU Leuven. It plays a key role in the pedagogical Pedalion project[12] and in a recently approved

[10] While this section mainly discussed label ambiguities, syntactic head attachment may also be ambiguous: see e.g. McGillivray and Vatri, 2015 for a discussion on how to resolve such ambiguities. Again, automatic methods could be suitable to detect such ambiguities, if the right features (e.g. valency and prosodical information, as discussed by McGillivray and Vatri) are provided.

[11] https://github.com/perseids-publications/glaux-trees

[12] http://www.pedalion.be

research project entitled *Language and Ideas: Towards a New Computational and Corpus-Based Approach to Ancient Greek Semantics and the History of Ideas* (FWO, Research Foundation – Flanders, grant number 3H200733). In this project we will examine how the *GLAUx* corpus can be applied to the study of language-related ideas expressed in Ancient Greek. The underlying hypothesis is that applying well-informed corpus-based methods, going beyond the level of the individual word or term, enables us to study (intellectual and conceptual) history from a wider perspective. It goes without saying that the applications for other domains and projects are manifold.

Some short-term enhancements we are planning include improving the underlying NLP work (in particular, we are currently exploring the possibilities of training an *ELECTRA* transformer model: see Clark et al., 2020), the addition of a derivational annotation layer and changing the syntactic annotation format to *Universal Dependencies*. In the long term, we will also expand *GLAUx* with the epigraphical data and develop techniques to handle the peculiarities of these texts, and expand the literary data up until the eighth century AD. To improve the accessibility of the data, we are currently designing detailed documentation about the different annotation layers of *GLAUx*, and will also provide a user interface to query the data.

All the data provided for *GLAUx* will be openly released online. We are currently discussing collaboration opportunities with other major digital projects of Greek, including the *Open Greek and Latin* project[13] and Trismegistos[14], so as to expand the possibilities for digital approaches to Ancient Greek as much as possible in the near future.

Acknowledgments

The work described in this paper has its basis in the Natural Language Processing and corpus design work I carried out during my PhD (funded by FWO, Research Foundation – Flanders, grant number 1162017N), which was supervised by Dirk Speelman, Toon Van Hal and Mark Depauw, as well as a one-year project funded by KU Leuven.[15] Mark Depauw has also considerably assisted me with the papyrus part of this corpus, and Toon Van Hal with the literary part. Currently the underlying NLP work is also further being improved through the help of Wouter Mercelis. I would also like to thank the three anonymous reviewers and the workshop organizers for their feedback, which has greatly helped to improve the quality of this paper. Lastly, the work described in this paper would not have been possible without the immense effort of various corpus annotators (including our own job students at KU Leuven) to provide a large amount of manually annotated data for Greek which this project has used as training data.

References

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, pages 5–15, Milan. EDUCatt: Ente per il Diritto allo Studio Universitario dell'Università Cattolic. https://convegni.unicatt.it/meetings_Proceedings_TLT8.pdf.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney. Association for Computational Linguistics. https://www.aclweb.org/anthology/W06-1615.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, Text, Speech and Language Technology 20, pages 103–127. Springer, New York. https://doi.org/10.1007/978-94-010-0201-1_7.

Bernd Bohnet and Joakim Nivre. 2012. A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea. Association for Computational Linguistics. https://www.aclweb.org/anthology/D12-1133.

Hugh A. Cayless, James M.S. Cowey, Ryan Baumman, and Timothy David Hill. 2021. Papyri.info IDP (Integrating Digital Papyrology) Data. https://github.com/papyri/idp.data.

[13] https://www.opengreekandlatin.org

[14] https://www.trismegistos.org

[15] https://www.kuleuven.be/onderzoek/portaal/#/projecten/3H200333

Giuseppe G. A. Celano. 2017. Lemmatized Ancient Greek Texts. https://github.com/gcelano/LemmatizedAncientGreekXML.

Giuseppe G. A. Celano. 2018. An Automatic Morphological Annotation and Lemmatization for the IDP Papyri. In Nicola Reggiani, editor, *Digital Papyrology II: Case Studies on the Digital Edition of Ancient Greek Papyri*, pages 139–147. De Gruyter Open Access Books, Berlin, Boston. https://doi.org/10.1515/9783110547450-008.

Giuseppe G. A. Celano and Gregory Crane. 2015. Semantic role annotation in the ancient greek dependency treebank. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 26–34, Warsaw. Polish Academy of Sciences, Institute of Computer Science. http://tlt14.ipipan.waw.pl/files/4614/5063/3858/TLT14_proceedings.pdf.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. arXiv:2003.10555.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537. https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf.

Gregory Crane. 1991. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6(4):243–245. https://doi.org/10.1093/llc/6.4.243.

Gregory Crane. 2021. Perseus Digital Library. https://github.com/PerseusDL/canonical-greekLit.

Gregory Crane, Lenny Muellner, Bruce Robertson, Alison Babeu, Lisa Cerrato, Thomas Koentges, Rhea Lesage, Lucie Stylianopoulos, and James Tauber. 2021. First1KGreek. https://opengreekandlatin.github.io/First1KGreek.

William Croft. 2013. Radical Construction Grammar. In Thomas Hoffmann and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*, pages 211–232. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780195396683.013.0012.

Francesca Dell'Oro and Giuseppe GA Celano. 2019. Epigraphic Treebanks: Some Considerations from a Work in Progress. *Classics*@(First Drafts@). https://chs.harvard.edu/wp-content/uploads/2020/11/DellOroCelano_4.pdf.

Mark Depauw and Tom Gheldof. 2014. Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. In Łukasz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, and Jochen Schirrwagen, editors, *Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops*, pages 40–52, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-08425-1_5.

Mark Depauw and Joanne Stolk. 2014. Linguistic variation in Greek papyri: Towards a new tool for quantitative study. *Greek, Roman, and Byzantine Studies*, 55(1):196–220. https://grbs.library.duke.edu/article/view/15245/6561.

Helma Dik and Richard Whaling. 2008. Bootstrapping Classical Greek Morphology. In *Digital Humanities 2008*, pages 105–106, Oulu. Association for Literary and Linguistic Computing, Association for Computers and the Humanities and Society for Digital Humanities. http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3002.

Regine Eckardt. 2006. *Meaning Change in Grammaticalization: An Enquiry into Semantic Reanalysis*. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199262601.001.0001.

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

Stefan Th. Gries, Tobias Bernaisch, and Benedikt Heller. 2018. A corpus-linguistic account of the history of the genitive alternation in Singapore English. In Sandra C. Deshors, editor, *Modeling World Englishes: Assessing the interplay of emancipation and globalization of ESL varieties*, Varieties of English Around the World, pages 245–280. John Benjamins Publishing Company, Amsterdam; Philadelphia. https://doi.org/10.1075/veaw.g61.10gri.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687. https://doi.org/10.1353/lan.2010.0021.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In Caroline Sporleder and Kiril Ribarov, editors, *Proceedings of the second* *workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34, Marrakech. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf.

Bernd Heine. 2002. On the Role of Context in Grammaticalization. In Ilse Wischer and Gabriele Diewald, editors, *New Reflections on Grammaticalization*, Typological Studies in Language 49, pages 83–101. Benjamins, Amsterdam. https://doi.org/10.1075/tsl.49.08hei.

Paul A. Iversen. 2007. The Packard Humanities Institute (PHI) Greek Epigraphy Project and the Revolution in Greek Epigraphy. Abgadiyat, 2(1):51–55.

Henry Stuart Jones, Henry George Liddell, Roderick MacKenzie, Robert Scott, and A. A. Thompson. 1996. *A Greek-English Lexicon*. Clarendon, Oxford, New ed. with new supplement edition.

Alek Keersmaekers. 2020a. Automatic semantic role labeling in Ancient Greek using distributional semantic modeling. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 59–67, Marseille. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.lt4hala-1.9.

Alek Keersmaekers. 2020b. *A computational approach to the Greek papyri: developing a corpus to study variation and change in the post-classical Greek complementation system*. Ph.D. thesis, KU Leuven. https://lirias.kuleuven.be/3084305.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, Enriching and Valorizing Treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, August. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/W19-7812.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania. https://repository.upenn.edu/dissertations/AAI3179808.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico. Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1117.

Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. https://www.aclweb.org/anthology/W19-8505.

Francesco Mambrini and Marco Carlo Passarotti. 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*, pages 133–144. Edições Colibri. https://publicatt.unicatt.it/handle/10807/37956.

Barbara McGillivray, Simon Hengchen, Viivi Lähteenoja, Marco Palma, and Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907. https://doi.org/10.1093/llc/fqz036.

Barbara McGillivray and Alessandro Vatri. 2015. Computational valency lexica for Latin and Greek in use: a case study of syntactic ambiguity. *Journal of Latin Linguistics*, 14(1):101–126. https://doi.org/10.1515/joll-2015-0005.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1272.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. https://www.aclweb.org/anthology/L16-1262.

Maria C. Pantelia. 2021. Thesaurus Linguae Graecae® Digital Library. http://www.tlg.uci.edu.

Marco Passarotti. 2014. From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 100–109, Gothenburg. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0615.

Florent Perek and Martin Hilpert. 2017. A distributional semantic approach to the periodization of change in the productivity of constructions. *International journal of corpus linguistics*, 22(4):490–520. https://doi.org/10.1075/ijcl.16128.per.

Volha Petukhova and Harry Bunt. 2008. LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L08-1428.

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael, California. https://doi.org/10.2200/S00436ED1V01Y201207HLT017.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester.

Helmut Schmid and Florian Laws. 2008. Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester. Coling 2008 Organizing Committee. https://www.aclweb.org/anthology/C08-1098.

Tobias Schnabel and Hinrich Schütze. 2014. FLORS: Fast and Simple Domain Adaptation for Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, 2:15–26. https://doi.org/10.1162/tacl_a_00162.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser Training with Heterogeneous Treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2098.

Elizabeth Closs Traugott. 2012. The Status of Onset Contexts in Analysis of Micro-Changes. In Merja Kytö, editor, *English Corpus Linguistics: Crossing Paths*, Language and Computers, pages 221–255. Rodopi, Amsterdam. https://doi.org/10.1163/9789401207935_012.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188. https://doi.org/10.1613/jair.2934.

Toon Van Hal and Yannick Anné. 2017. Reconciling the dynamics of language with a grammar handbook: The ongoing Pedalion grammar project. *Digital Scholarship in the Humanities*, 32(2):448–454. https://doi.org/10.1093/llc/fqv068.

Alessandro Vatri and Barbara McGillivray. 2018. The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65. https://doi.org/10.1163/24523666-01000013.

Alessandro Vatri and Barbara McGillivray. 2020. Lemmatization for Ancient Greek: An experimental assessment of the state of the art. *Journal of Greek Linguistics*, 20(2):179–196. https://doi.org/10.1163/15699846-02002001.