Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5334–5344
Marseille, 20-25 June 2022

AGILe: The First Lemmatizer for Ancient Greek Inscriptions

Evelien de Graaf*, Silvia Stopponi*, Jasper Bos, Saskia Peels-Matthey and Malvina Nissim
Centre for Language and Cognition Groningen, University of Groningen, The Netherlands
{e.de.graaf.6,j.k.bos.1}@student.rug.nl
{s.stopponi,s.peels,m.nissim}@rug.nl

Abstract

To facilitate corpus searches by classicists as well as to reduce data sparsity when training models, we focus on the automatic lemmatization of ancient Greek inscriptions, which have not received as much attention in this sense as literary text data has. We show that existing lemmatizers for ancient Greek, trained on literary data, are not performant on epigraphic data, due to major language differences between the two types of texts. We thus train the first inscription-specific lemmatizer achieving above 80% accuracy, and make both the models and the lemmatized data available to the community. We also provide a detailed error analysis highlighting peculiarities of inscriptions which again highlights the importance of a lemmatizer dedicated to inscriptions.

Keywords: Digital Classics, lemmatization, Ancient Greek inscriptions

1 Introduction

Lemmatization is a basic preprocessing step for automatic linguistic analysis. It is particularly helpful in the case of morphologically complex languages: by reconducting a large variety of wordforms to a single lemma, one can enable meaningful generalizations, especially at the semantic level, which could otherwise remain opaque due to the sparsity of each single wordform. This is even truer in the case of low-resource languages or varieties, where the amount of available data is limited, and possibly more sophisticated processing tools that might not need to rely on lemmatization are not (yet) available.

We focus here on ancient Greek, a morphologically complex language, with rich nominal and verbal inflection. Ancient Greek is also relatively low-resource. While substantial effort has gone into collecting and digitizing all available data, such as for the big projects Thesaurus Linguae Graecae (Pantelia, 2001) and the Perseus Digital Library (Smith et al., 2000), the existing collection can be considered finite: apart of occasional discoveries of texts, often on stones and papyri, the corpus is not enlargeable. Another characteristic of the existing ancient Greek data is that the texts are written in different language varieties, sometimes very distant from each other, both synchronically and diachronically. For example, the language of ancient Greek inscriptions,[1] namely all the texts which were written on durable materials, such as stone, ceramic, metal and other materials, can differ substantially from that of literary texts, especially when it comes to orthography, morphology and dialectal variation. Most of the existing tools developed for ancient Greek at large are built on (and for) literary texts, and because of such differences they are not necessarily suitable for other kinds of ancient Greek texts.

For literary ancient Greek several lemmatizers have been developed and made available, and the most popular online corpora of literary texts, such as the *Thesaurus Linguae Graecae* (Pantelia, 2001), have been lemmatized. Crucially, this allows Classics scholars to perform lemma-based searches through web interfaces, enhancing the potential of automatic analysis also for the computationally non-savvy. Some lemmatized corpora of literary ancient Greek are also fully downloadable, such as the Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018). The situation for inscriptions is widely different, though: it is not possible to run searches by lemma on most Greek epigraphic corpora, since only a few of them have been manually lemmatized. For example, one of the most often used online corpora, provided by the Packard Humanities Institute,[2] is not lemmatized, and to search for all the possible wordforms of a lemma it is necessary to run many single searches, taking into account large dialectal and spelling variation; the amount of necessary queries increases even more if one wants to search for a *combination* of two lemmata. Moreover, in the few corpora which have been lemmatized (generally by hand) sometimes not all the tokens have been assigned a lemma, as in the case of the *Collection of Greek Ritual Norms* (CGRN), a manually-curated collection of epigraphic texts (see Section 2.1).

This situation clearly clashes with the importance of inscriptions for Classical studies; Bodel (2001) estimates that about 600,000 Greek and Latin inscriptions coming from the timespan c. 800 BCE - 700 CE have been uncovered, making them a goldmine for ancient Greek scholars. Creating tools to process ancient Greek inscriptions is therefore necessary, also considering the growing interest in applying distant reading methods to ancient texts, for example by training language models which might particularly benefit from reducing data sparsity via lemmatization.

With this situation in mind, we aimed at providing lemmatization for inscription data. First, we ran the existing lemmatizers specialized on literary ancient Greek texts, to find, unsurprisingly, that their performance on inscriptions is not particularly satisfactory, and largely behind their accuracy on literary ancient Greek (Section 3.2). As a consequence,

* Joint first authors

[1] In this article, we use the terms ancient Greek 'inscriptions' and 'epigraphic data' interchangeably.

[2] The PHI corpus is available at https://inscriptions.packhum.org/.

exploiting existing portions of (partially) manually annotated inscriptions, we trained an inscription-specific lemmatizer, **AGILe** (**A**ncient **G**reek **I**nscription **Le**mmatizer), achieving a much higher performance (Section 4.4). This paper describes the challenges of working with inscriptions, the performance of existing lemmatizers and the development of our specific tool, and eventually provides a detailed error analysis which links back to the specific characteristics of the different type of language used in literary and epigraphic data.

Contributions We develop and make available AGILe, the first lemmatizer specifically developed to work on ancient Greek inscriptions. We assess its performance also in comparison to existing lemmatizers, thereby providing the first evaluation of these systems on epigraphic data, and showing that, due to the peculiarities of these texts, such systems are not suitable and are outperformed by our lemmatizer by 20 to 40 percentage points. Lastly, we provide a rich qualitative analysis of errors made by our lemmatizer, thus offering a deep insight into the specificities of inscriptions and highlighting potential directions for improvement. All code and models are available: `https://github.com/agile-gronlp`.

2 Data

2.1 Inscription Data

Greek inscriptions from before the 4th century BCE are characterized by many different local alphabets, a large dialectal variation, and a lack of standardized spelling. For example, the consonant cluster /ks/, spelled with the orthographic sign ξ in the literary corpus, may be spelled as χ, ξ, χσ or κσ in the inscriptions, depending on the alphabet. The project *A Collection of Greek Ritual Norms* (CGRN) assembles a selection of 225 Greek inscriptions of a specific kind, usually referred to as "sacred laws" in current scholarship. This (misleading) rubric is in fact comprised of a substantial variety of epigraphic documents concerned with Greek cult, including decrees, calendars, boundary stones, testaments etc. The inscriptions in the CGRN, in particular, are normative texts concerning religious rituals, in varying degrees of formality. Given their long chronological range (from the 6th century BCE until the 1st century CE) and their topographical spread, covering all parts of the ancient Greek world,[3] this corpus is apt to exemplify the dialectal and spelling variations characteristic of Greek epigraphy.

The dataset includes texts that have been encoded in TEI XML format (according to the EpiDoc standard), for a total of 38,034 tokens, 25,229 of which have been lemmatized by hand. The other tokens are either not lemmatized at all (12,667) –mostly articles and prepositions, since they were uninteresting for the specific scholarly purposes of the CGRN– or they have a 'word'-tag, but the lemma-attribute of this tag is noted either as unclear or empty (in 138 cases),

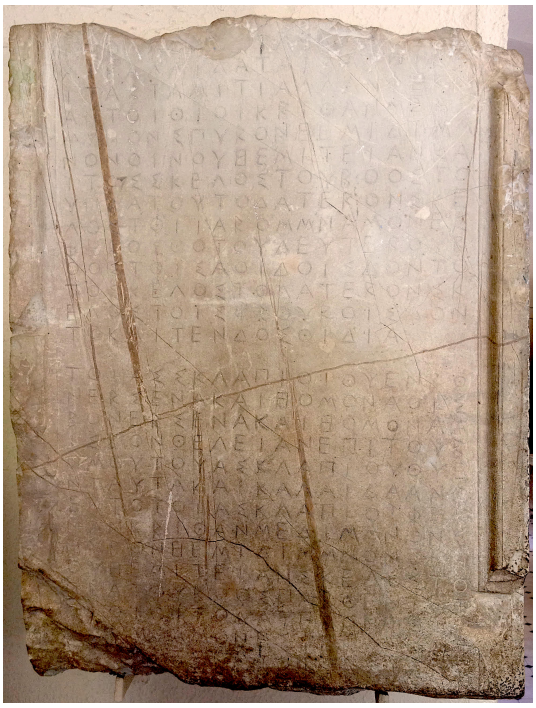

Figure 1: Stone with ancient Greek inscription (CGRN 34, end of 5th century BCE).

when wordforms are fragmentary because the stone is unreadable or broken. Figure 1 shows an example of a broken stone, a white marble stele, with two inscriptions dating from the 5th century BCE.[4] The CGRN follows the best available modern editions of the inscriptions included, and in a few cases provides a new edition. These editions usually follow the ancient Greek text as it was found on the stone, without regularizing spelling or grammar, but with the addition of word separations, punctuation, and proposals for restoration of parts of the text that are now missing or illegible due to damage of the text carrier. In a few cases, there are editorial corrections (of obvious spelling mistakes).

The language of Ancient Greek inscriptions, especially those preceding the 4th century BCE, has particular features, for example the use of the character *h* to indicate an initial aspiration and of the character ϝ (digamma) to represent the sound /w/. These two characters appear in inscriptions from various regions, but they are both generally absent from literary Greek, from the base forms in the dictionaries, and thus from the gold standard of the CGRN, where we find variants of the same wordforms, without *h* (substituted by the diacritic ʽ for aspiration) and without ϝ. The digamma disappeared without a trace, for example the word ταϝῦρος ('bull') became ταῦρος. An example of *h* substituted with the diacritic ʽ for aspiration is the word hέρος ('hero'), later written as ἥρως. In this last word we also see another characteristic of the language of early inscriptions, the fact that no difference is made between long and short vowels in the notation. In hέρος, a long /e/, later

[3] The inscriptions range from Attica to Asia Minor and Anatolia, from Sicily to Egypt; the corpus covers also the Peloponnese, Central and Northern Greece, the Aegean Islands, Crete, and the Black Sea region.

[4] This is CGRN 34, two sacrificial regulations of Apollo and Asclepius in Epidauros, found in a new wall of the temple of Asclepius there and still in place.

written as η, is noted in the same way as the short ε, and a long /o/, later written as ω, is noted as ο, in the same way as the short vowel (Colvin, 2007, 19). We will see in Section 5 how these and other typical features of inscriptions can make lemmatization more challenging.

The differences we encounter between the Greek of inscriptions and literary Greek are due to the different way of transmission of such texts. Inscriptions have been carved on stone (or other durable materials) by professional engravers and private individuals, permanently. We are therefore witness to great variation rising from epichoric (local) alphabets that were in use before the 5th century BCE, spelling idiosyncrasies and mistakes, and large dialectal variation. Literary works have often been transmitted to us in manuscripts copied by scribes throughout the centuries, an editorial practice that started with the Alexandrian scholars, was continued in the medieval textual tradition, and is still ongoing today. Scribes imposed a degree of uniformity on the texts, using only one alphabet, correcting mistakes, adding accents, transliterating into minuscule script, and modifying texts in other ways, for example to conform the language and style of a text to a specific dialect or literary model (Reynolds and Wilson, 1991). Some literary works are only known by evidence from one or more ancient papyri found in Egypt, but this is typically not the case of the corpora of ancient Greek literary texts used by scholars to train and evaluate existing lemmatizers of ancient Greek.

2.2 Other Data

Even if the current work focuses on the lemmatization of inscriptions, there are two other corpora of ancient Greek literary texts we will refer to because they were used by other scholars to train and evaluate the existing lemmatizers of ancient Greek, and, in the case of the PROIEL corpus (Haug and Jøhndal, 2008), also to boost the training of our new lemmatizer. The two corpora are two treebanks of ancient Greek texts, the PROIEL (Haug and Jøhndal, 2008) and the Ancient Greek Dependency Treebank (AGDT or 'Perseus' treebank). PROIEL[5] is a treebank of ancient Indo-European languages, the ancient Greek part of which is composed of the Greek New Testament, Herodotus' *Histories* and Sphrantzes' *Chronicles*. When referring to PROIEL in this article, we refer to this portion. The AGDT[6] is a portion of the ancient Greek texts provided by the Perseus Digital Library[7] that has been syntactically annotated. Both treebanks have also been released as part of the Universal Dependencies project (Nivre et al., 2018).[8] A comparison between the AGLDT, the PROIEL treebank and other available dependency treebanks for Greek and Latin is provided in Celano (2019).

3 Existing Lemmatizers for ancient Greek

As part of our goal to obtain a well-performing lemmatizer for inscriptions, we first assessed how existing lemmatizers perform on epigraphic data and compared their observed accuracy to that reported for literary texts, the genre on which they have mostly been trained. Indeed, several lemmatizers for ancient Greek exist, but they are only trained on literary texts. One exception is a lemmatizer specifically developed for Greek documentary papyri by Keersmaekers (2019), which achieved an accuracy of 98.5 on that genre. Keersmaekers (2019, 75) used Lemming (Müller et al., 2015) by retraining it on a morphologically annotated corpus of papyri and leveraging all lemmas included in the Greek-English Lexicon by Liddell and Scott (1940). Even if the language of documentary papyri is closer to literary texts than to inscriptions, the work done by Keersmaekers (2019; 2020) shows that there is interest in lemmatization of non-literary ancient Greek texts.[9]

3.1 Considered Lemmatizers

We test four different existing lemmatizers, namely GLEM (Bary et al., 2017), two lemmatizers provided within the Classical Language Toolkit (CLTK, Burns (2019), Johnson et al. (2021)), and the UDPipe model (Straka, 2018).

Other lemmatizers available for ancient Greek were not evaluated either because they were not freely accessible for testing or because they do not disambiguate between lemmas without the addition of a separate POS-tagger. For example: TreeTagger[10] was discarded because the parameter files provided by Alessandro Vatri and Barbara McGillivray do not contain any lemmas and it would have therefore been impossible to evaluate the performance of the tool. Bridge[11] only returns a lemma if there is a single possibility (in other words: it doesn't deal with ambiguities), which would have resulted in too small a portion of lemmatized wordforms, as multiple lemmas are often suitable for the same wordform, of course. Two more lemmatizers were discarded for the same reason, namely Morpheus[12] and Eulexis.[13] We also considered to use the Ancient Greek Wordnet,[14] but the API did not work at the time of our experiments (March 2021).

Both GLEM and the CLTK lemmatizers have been designed specifically for ancient Greek. **GLEM** (Bary et al., 2017) leverages POS information, thanks to a light version of the Frog POS tagger, an NLP module originally developed for Dutch (Hendrickx et al., 2016). This machine learning component is combined with a lexicon look up[15] and it is used to disambiguate between lemmas and to gen-

[5] https://github.com/proiel, more information at http://dev.syntacticus.org/proiel.html.

[6] https://github.com/UniversalDependencies/UD_Ancient_Greek-Perseus/, more information at http://www.dh.uni-leipzig.de/wo/projects/ancient-greek-and-latin-dependency-treebank-2-0/.

[7] http://www.perseus.tufts.edu/hopper/

[8] https://universaldependencies.org/

[9] This lemmatizer is currently not available online.

[10] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[11] https://bridge.haverford.edu/lemmatizer/

[12] http://www.perseus.tufts.edu/hopper/morph

[13] https://outils.biblissima.fr/en/eulexis-web/index.php

[14] https://greekwordnet.chs.harvard.edu/

[15] The lexicon was built by merging the Perseus lexicon (from

erate a lemma for forms not in the lexicon. The Frog POS-tagger was retrained on the PROIEL text of Herodotus' *Histories*, manually annotated for lemmas and parts-of-speech, while the Frog lemmatization module was trained on both Herodotus and the merged Perseus-PROIEL lexicon (Bary et al., 2017, 93).

The **CLTK** lemmatizer (Johnson et al., 2021) has been tested in this work both in its default[16] and in its backoff version, which combines several lemmatizers in sequence (Burns, 2020)[17]. The 'default' lemmatizer is part of a Stanza-based pipeline; to lemmatize ancient Greek the version trained on the PROIEL corpus was used.[18] The 'backoff' lemmatizer leverages the following resources, instead: a hand-written dictionary of 1,049 frequent, non-ambiguous lemmatized tokens, a list of 33,555 sentence-level token-lemma pairs from the AGDT and, finally, another token-lemma dictionary of 949,453 lemmatized tokens, already used in a previous version of the 'backoff' lemmatizer.[19]

The **UDPipe** lemmatizer is part of a pipeline that used as training data the Universal Dependencies treebanks (Nivre et al., 2018); for ancient Greek, in particular, the Universal Dependencies include two resources, the PROIEL and the Perseus treebank, which include lemma tags, and two different pipelines for ancient Greek were thus trained on these two treebanks. A joint model performs POS tagging, lemmatization and parsing, and the contextualized embeddings created by the POS-tagger are reused in lemmatization. This last task is performed as a classification task, through which wordforms are classified into lemma generation rules (Straka, 2018).

We checked previously reported accuracies for all of the considered lemmatizers, which we report in Table 1, though there is not a single evaluation benchmark in place. GLEM was evaluated in comparison to Frog and the CLTK lemmatizer on Herodotus, the annotated text from the PROIEL corpus (also used for training, though as a different portion), and on the first fifteen chapters of book one from Thucydides' *History of the Peloponnesian war*, annotated by this team; the results are in Table 1. Recently, Vatri and McGillivray (2020) assessed the accuracy of GLEM and of the two CLTK lemmatizers also on two other manually lemmatized texts, namely Homer, *Iliad* I 1-279 and Lysias, 7, *On the Olive Stump*. For the CLTK lemmatizer no other individual accuracy analysis exists. For UDPipe, the only lemmatizer evaluated in this work that was not specifically developed for ancient languages, we find self-reported accuracy on ancient Greek for the versions 2.0 (Straka et al., 2019a) and 2.3 (Straka et al., 2019b), trained and tested on the PROIEL and Perseus treebanks.

3.2 Performance on Inscriptions

Cleaned text files were created semi-automatically from all the 225 CGRN's XML files, without including punctuation,[20] and these text files were used as test data for the lemmatizers. From the XML files all the word forms were extracted, together with their lemma, in order to create gold standard CSV files. The lemmatizers were tested on these files, following the steps in Figure 2, and the results of the different lemmatizations were appended to the gold standard CSV files for comparison. At this point, several additional issues preventing automated comparison were discovered, and had to be fixed manually. The four lemmatizers were subsequently run on the CGRN texts, and their accuracy is summarized in Table 2.[21]

The lemmatizers face most issues with early and dialectal wordforms or names. In general, UDPipe seems to outperform the others regarding the lemmatization of names. CLTK is the closest to dealing correctly with cases of crasis[22] and with forms containing either the character ϝ or the character *h*, since in most cases the CLTK lemmatizer is the only one correctly deleting the first character and trying to replace the second. However, in general it is subsequently not able to correctly lemmatize the wordforms. GLEM is able to deal with some dialectal forms where we find an α instead of an η, but only in a minority of these cases. It seems that when only one lemmatizer is able to correctly lemmatize one of these 'difficult' forms, it is most often GLEM, closely followed by the CLTK default lemmatizer. UDPipe seems to correctly lemmatize in most cases wordforms that either GLEM or CLTK also correctly lemmatized, which we can thus consider 'easier' forms. The analysis also demonstrates that the backoff lemmatizer was rarely the only one to correctly lemmatize a wordform, suggesting that the CLTK backoff lemmatizer does not bring any relevant benefit. In 75.2% of the cases at least one lemmatizer was able to identify the correct lemma, a far better result than the performances of the individual lemmatizers.

the Perseus project) with the PROIEL lexicon.

[16]By 'default' lemmatizer we mean the tool behind the lemmatization performed in the pipeline for ancient Greek, available at: `https://docs.cltk.org/en/latest/languages.html`.

[17]Code available at `https://docs.cltk.org/en/latest/_modules/cltk/lemmatize/grc.html`.

[18]The Stanza pipeline is available at `http://stanza.run/`, the language model trained on PROIEL at `https://stanfordnlp.github.io/stanza/available_models.html`.

[19]Resources available at `https://github.com/cltk/grc_models_cltk/tree/master/lemmata/backoff`.

[20]During evaluation it became clear that the lemmatizers functioned better without any punctuation present, see also Vatri and McGillivray (2020, 183).

[21]For UDPipe the pipeline trained on the Perseus corpus was used. For CLTK both the standard ancient Greek Pipeline and the Backoff lemmatizer were tested. From the two pipelines (UDPipe and CLTK) the lemmas were obtained by extracting them from the files returned at the end of the process. The differences in the total amount of lemmatized wordforms derives from several issues, such as the wrong separation of the apostrophe from a wordform with an elided vowel. These forms were manually deleted from the CSV files, resulting in small differences between the total wordforms lemmatized.

[22]Crasis is the contraction of vowels by merging two adjacent words, e.g. καὶ ἐγώ to κἀγώ.

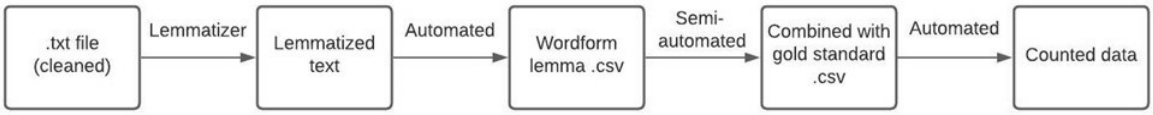

Figure 2: Representation of the steps taken in evaluation.

Lemmatizer ↓ / Test data →	Herodotus	Thucydides	Homer	Lysias	PROIEL	Perseus
GLEM punctuation (a,b)	95.7	93.0	72	81	-	-
GLEM no punctuation (b)	-	-	84	94	-	-
Frog (a)	87.1	75.6	-	-	-	-
CLTK (a)	78.7	76.6	-	-	-	-
CLTK backoff (b)	-	-	91	97	-	-
CLTK (b)	-	-	65	65	-	-
UDPipe 2.0 (c)	-	-	-	-	94.0	91.9
UDPipe 2.3 (c)	-	-	-	-	93.5	85.0

Table 1: Accuracy of all the lemmatizers on all the test data. The letters after the names of the lemmatizers point to the source(s) of the reported scores. All scores are rounded as reported in the original papers that we refer to.
(a) Bary et al. (2017, 94) report the accuracy of GLEM, Frog and the CLTK lemmatizer on Herodotus and Thucydides. Accuracies on Herodotus are obtained via 10-fold cross-validation.
(b) Vatri and McGillivray (2020, 191) report the accuracy of GLEM and of two CLTK lemmatizers (backoff and non-backoff) on Homer and Lysias. The evaluations of the CLTK lemmatizers from (a) and (b) are on different rows, since the authors used pre-release versions, probably at different stages of development.
(c) Straka et al. (2019a) and Straka et al. (2019b) report the accuracy of UDPipe on PROIEL and Perseus.

System	Acc	Wrong	Correct	Missed
UDPipe	46.3	13,474	11,606	149
CLTK	46.4	13,390	11,581	258
CLTKb	37.1	15,768	9,292	169
GLEM	62.5	9,379	15,650	200

Table 2: Accuracy of the lemmatizers on the CGRN. The total number of lemmas is 25,229 (see Section 2.1).

4 A Lemmatizer for Ancient Greek Inscriptions

After having evaluated the available lemmatizers on inscriptions, the need for a lemmatizer specifically trained on ancient Greek inscriptions became evident. The importance of training NLP tools on the target variety of language was also pointed out by Keersmaekers (2020, 33–34), who showed how morphological tagging of Greek papyri is more effective when the training data is closer to the language of the target texts. For example, removing prose authors from the training data resulted in most cases in a lower accuracy, since documentary papyri, the target texts in Keersmaekers (2020), are also written in prose, and in the same way adding poetry lowered the accuracy.

4.1 Data and Preprocessing

To build a specialized lemmatizer for ancient Greek inscriptions, two corpora have been combined for the training. The CGRN corpus forms the foundation with its texts separated in three representative splits, each including inscriptions from different periods, and preserving a 60-20-20 split for training, developing and testing. To improve the robustness of the model, a section of the PROIEL corpus (Haug and Jøhndal, 2008), consisting of the New Testament plus selections from Herodotus, has been included. The approximately 214,000 lemmas are divided over the original 88-6-6 split performed for the creation of the treebank,[23] and most punctuation marks present in the inscriptions have been removed for performance optimization.

4.2 Model

The neural lemmatizer of Stanza (Qi et al., 2020, 3), a model which is an ensemble of a dictionary-based lemmatizer and a neural sequence-to-sequence lemmatizer, was trained on the training set of our corpus. It ignores part-of-speech tags and saves the best model out of 50 epochs of fine-tuning. Otherwise, the default configuration is used. A detailed overview of the pre-processing and training process is available in our repository.

4.3 Inscription-specific Processing

To further optimize the lemmatizer, the model has been complemented with an optional lexicon lookup. The lexicon is composed of all entries in the *Liddell-Scott-Jones Greek-English Lexicon* (Liddell and Scott, 1940)[24] plus the gold lemmas from the training set. Each lemma predicted

[23] https://github.com/UniversalDependencies/UD_Ancient_Greek-PROIEL/tree/master.

[24] We used the Unicode version provided by Giuseppe Celano, derived from the betacode version provided by the Perseus project, https://github.com/gcelano/LSJ_GreekUnicode.)

by the model is checked against this lexicon; if the lemma does not occur in the lexicon, it is corrected to its first closest equivalent given its edit distance by assigning the cost of one to each operation needed to transform the predicted lemma into the closest word in the lexicon. The possible operations are substitution, insertion and deletion of a character, for example the cost of transforming the predicted lemma ἄνθροπος (non-existent) into the correct lemma ἄνθρωπος has the cost of 1, since only the substitution of ο with ω is needed, while transforming it into the other (wrong) lemma ἄνθραξ would have had the cost of 4, two substitution plus two deletions. In addition, the following custom rules are specified:

1. the digamma (ϝ) and the *h* are ignored, since the dictionary and gold lemmas are identical to the older lemmas without digamma, e.g. ϝοίκος("house") becomes later οἶκος, or, in the case of *h*, to the lemma without *h*, but with the diacritic ʽ;
2. the combinations of characters κ+σ/ς and χ+σ/ς have been converted to the character ξ, both capitalized and non-capitalized;
3. the combination φ+σ/ς, capitalized and non-capitalized, has been converted to ψ.

4.4 Results

Our best model uses all of the inscription-specific processing, and achieves an accuracy of 84.7% on the development set. We assessed the contribution of the lexicon look-up, which is the only external component of our lemmatizer, and observed a model which does not use it achieve an accuracy of 82.1%, still on development data. This suggests that while it helps to include this module, should one want to exclude it to make a lighter system, performance would not suffer substantially.

We evaluate our full model on the held out test data and obtain an accuracy of 85.1%, showing that the lemmatizer is robust on unseen data. As seen in Table 2, and discussed at length in Section 3, existing lemmatizers yielded a subpar performance on the entire CGRN. For a more direct comparison with AGILe, we ran them on the test portion only so that accuracies would be obtained and compared on exactly the same instances (5030 tokens). Table 3 reports the results, with the addition of AGILe's accuracy.[25] We can see that scores are very much in line with those observed on the whole dataset and that AGILe's performance goes well beyond that of lemmatizers that were not specifically developed for ancient Greek inscriptions.

In the next Section we provide a detailed error analysis of AGILe's decisions on test data to better understand its behaviour, especially in relation to the specific characteristics of epigraphic data.

Lemmatizer	Acc.
UDPipe	45.0
CLTK	41.6
CLTKb	34.8
GLEM	61.5
AGILe	**85.1**

Table 3: Accuracy of lemmatizers on the CGRN test set.

5 Error Analysis

5.1 General Analysis

We ran a manual analysis over a portion of the errors AGILe made on the test data. Out of all the wrongly lemmatized 750 tokens (14.9% of the total), 410 errors affect tokens that occur only once. For the tokens that occur multiple times, the lemmatizer seems quite consistent, meaning that all occurrences are either all wrong or all correct. We have observed that in the 14 cases in which different occurrences of the same wordform were inconsistently lemmatized (some wrongly and some correctly) the problem always resided with inconsistent gold annotation. In what follows, we analyze the nature of approximately 250 errors, by considering unique tokens (i.e. each wrongly lemmatized token was analyzed once, independently of its frequency) and including both tokens which occur just once and tokens occurring multiple times. First, we discuss the cases in which AGILe was unable to produce the correct lemma, then we move on to the cases in which AGILe actually turned out to be right after a manual evaluation.

AGILe sometimes had difficulties with spelling in early inscriptions, where the writing conventions to represent certain sounds were different from later Greek, on which the gold lemmas and the lexicon are based. For example, ω is spelled as ο and η as ε, such as in ἀρέν for ἀρήν or in ἀγõνα for ἀγῶνα, erroneously lemmatized by AGILe as ἀείρω and ἀγοίν.[26] A couple of other problems were indirectly caused by the ϝ. This character is removed via a rule before lemmatization so that for example the word form ϝοίνο is presented to AGILe as οίνο but not in all the cases this allowed the lemmatizer to arrive at a correct lemmatization. Even if οίνο was correctly lemmatized as a form of οἶνος, other wordforms such as for example ϝανακείοι and ϝρέξαντα, presented to the lemmatizer as ανακείοι and ρέξαντα, were wrongly predicted as ανάκειος and ρέγω (then corrected to φαρνάκειος and λέγω). A related issue is the fact that epigraphic data also contain spelling errors made by the inscribers and spelling irregularities. An example is the wrong form γίγιγνηται instead of the usual γίγνηται; in the wrong wordform the mistaken repetition of the characters γι (dittography) causes the incorrect prediction of the lemma as γιγνάω, subsequently corrected to γεννάω. A further case is the different spellings of the form ημιδιμμνον, a unit of volume which seems to have had lo-

[25] The scores are calculated on all tokens in the dataset for which the gold lemma is available, thereby excluding articles, particles, conjunctions and prepositions.

[26] The last was further corrected to the lemma ἀγορά. In other cases, the spelling did not hinder the correct lemmatization, for example of μέ as μή or of χρεμα as χρῆμα (after correction).

cal variation, ranging from ἡμίδιμμνον and ἡμέδιμνον to ἡμιμέδιμνον.[27] Another phenomenon difficult to handle also for AGILe, similarly to the other lemmatizers tested (see Section 3.2), was crasis, so that for example κἀπί, a contraction of καί and ἐπί, was wrongly predicted by AGILe as a form of (non-existent) κἀπός. In at least one case AGILe arrived at the correct lemma, but after correction; it is the case of τἐνδοσθίδια (crasis for τὰ ἐνδοσθίδια), wrongly predicted as τἐνδοσθίδια, but then corrected to the right ἐνδοσθίδια with the lexicon.

The lemmatizer also had difficulties with low-frequency forms. Due to the complex morphology of ancient Greek verb conjugation, various verb forms occur only once in the corpus or are more difficult to identify due to the archaic spelling of the inscriptions. Unique names (of locations, persons, months, festivals, deities, epithets, etc.), abounding in epigraphic data, were also often incorrectly lemmatized by AGILe. Examples of wrongly lemmatized wordforms of this kind are Προμεθίοις, a wordform of Προμήθεια, the festival of the Prometheia, wrongly predicted as the non-existent προμέθιον, and subsequently corrected to προμύθιον; Μαλεάτηι, a form of the lemma Μαλέατας, the god Maleatas, wrongly predicted as Μαλεάτης, and then corrected to γαλεώτης; and Γαμελιῶνος, the name of the month Gamelion, wrongly predicted as Γαμελιός instead of Γαμηλιών (then corrected to Γαμαλιήλ). Then there are cases in which it is not so easy to understand why and where the lemmatizer went wrong. For example when in seemingly similar situations, such as with similar but not identical wordforms of the same lemma, some forms are correctly lemmatized but others are not. One case is the token ἀνηλωμένα which is not recognized as a form of the verb ἀναλίσκω, while ἀναλώσει and ἀναλώσωσι, other forms of the same verb, are correctly lemmatized elsewhere. It is also not straightforward to understand why in 18 cases AGILe did not produce any prediction; 8 of them are occurrences of the wordform τεῖδε, which should be lemmatized as ὅδε; the remaining 10 are all proper names, a category that is particularly difficult for AGILe, but not all of them are rare or have an irregular inflection; for example, we would have expected to have a prediction at least for the genitive form Ἀπόλλονος (gold Ἀπόλλων) since instances of the alternative form Ἀπόλλωνος and of the two datives Ἀπόλλονι and Ἀπόλλωνι were correctly lemmatized. The same holds for the dative Ἀριστομάχωι (gold ἀριστόμαχος, incorrectly lower-cased) for which we would have expected a (capitalized) prediction, since 88 other datives in -ωι, included proper names, were correctly lemmatized by AGILe.

During the manual evaluation we discovered several issues that caused false negatives, for example some cases in which the 'gold' lemma was wrong (due to an incorrect input by the CGRN authors) but the lemma returned by AGILe was right. However, there were also 15 cases with errors in the gold lemma where AGILe returned the wrong lemma anyway. Moreover, there is a number of cases in which the output of AGILe is correct but not exactly identical to the gold lemma. For example, the form πρώτει was lemmatized as πρῶτος, a superlative, which is correct. But the gold lemma is the adjective in the comparative degree, πρότερος, since in the LSJ, the dictionary used as a reference by the authors of the CGRN to lemmatize the corpus, πρῶτος appears as a subheading under the entry πρότερος. In other cases, AGILe returns lemmas that are only spelling variants of the gold and thus technically correct: for example, it lemmatizes the token σκόρδων as σκόρδον, which is correct and presented by the LSJ as exactly equivalent to the variant σκόροδον (gold). Other false negatives are the frequent cases where AGILe produced as lemma a verb in the medio-passive voice, whereas the gold was in the active voice. This happened with the medio-passive form σπλαγχνίζεται, correctly lemmatized by AGILe as σπλαγχνίζομαι, a medio-passive lemma, but the gold lemma in the CGRN is the active form σπλαγχνίζω and this mismatch provoked a false negative. We also report some cases (about 12, including the instances occurring more than once) in which there is only an issue of capitalization, i.e. AGILe's answer is capitalized, whereas the gold is not or vice versa; for example, for the form Σκίροισι the lemma σκίρα was predicted, while the gold is the capitalized Σκίρα. There are also problems caused by accentuation, i.e. in around 30 cases (including repeated errors) AGILe produces the correct lemma but with a wrong accent yielding a non-existent form, which is then corrected, since AGILe does not find it in the lexicon. For example, for the form Φηραίωι the predicted lemma was the non-existent Φηραῖος, while the correct form would have been Φηραίος, with an acute accent instead of the circumflex. The predicted form was thus unfortunately corrected to ὡραῖος, since no lemma Φηραῖος was found in the lexicon. There also are cases in which the form to lemmatize is ambiguous between more than one lemma due to the complex morphology of ancient Greek and AGILe produced one of the possible correct answers but it was the wrong choice in that context. For example, the token ἀρὴν, occurring in one text, can be a form of both the lemma ἀρά ('prayer', 'curse') and the lemma ἀρήν ('lamb, sheep in general'); the last was the correct answer in that context but AGILe incorrectly selected the first. Similarly, the wordform σιωπῆι, ambiguous between the two lemmas σιωπάω, a verb meaning 'to be still', and σιωπή, a noun meaning 'silence', was lemmatized as the first, while the correct one was the second.

5.2 Influence of the Lexicon Lookup

The optional lexicon lookup of AGILe has improved the performance, as 141 more corrected lemmatizations were gained. For example, the lemma for the word form πεντακοσιᾶν was predicted as παντακόσιοι, then corrected to πεντακόσιοι, identical to the gold lemma. When the correction does not lead to a correct lemma, the corrected lemma is often further from the truth than the predicted lemma: for example, for the form βόληι AGILe predicts βόλη, a reasonable prediction according to the morphology of ancient Greek, but actually incorrect as this lemma does not exist and the gold lemma is the verb βούλομαι. The predicted

[27] The wordform ἡμέδιμνα was wrongly predicted as ἡμέδιμνα, without any change to the wordform, and corrected to μέριμνα, instead of the gold ἡμέδιμνον.

βόλη, absent from the lexicon, was corrected to πύλη, the least costly solution but less resembling the gold βούλομαι than the predicted βόλη. In 478 cases the call to the lexicon did not lead to the correct lemma and in 21 cases of these the predicted lemma was correct, but, since it was not in the lexicon, AGILe's correction led to an incorrect outcome; all of these 21 cases are unique names, rare words or *hapaxes*. For example, the lemma for the wordform ἄφαμμα, occurring only twice in the corpus,[28] is ἄφαμμα, and the lemma for the wordform Ῥοδίων should be Ῥόδιος; both of these were correctly predicted by AGILe, but then erroneously corrected to respectively ἔφαμμα and Ῥόδος, since the gold lemmas are not in the lexicon.

5.3 Comparison with Other Lemmatizers

Compared to the other four evaluated lemmatizers (see Table 3), the performance of AGILe on the test set is the highest, 85.1%. From a preliminary comparison between the wordforms incorrectly lemmatized by AGILe and by the already existing lemmatizers, it becomes clear that AGILe improves lemmatization on all points identified in the conclusion of Section 3.2, with the only exception being the phenomenon of crasis.[29] Since rules were built in to skip over the digamma and the *h*, these cases are now mostly lemmatized correctly. Moreover, AGILe is able to correctly deal with dialectal variation either immediately or after correction. Immediately in cases such as the form ἁμέραι (gold: ἡμέρα), where the predictions were as follows: ἁμέρα (UDPipe), ἁμύρα (CLTK), ἁμέραι (CLTKb), ἥμερος (GLEM), ἡμέρα (AGILe predicted); after correction for example for the wordform βομόν (gold: βωμός), wrongly predicted by all the other tested lemmatizers –as βομός (UDPipe), βοῦμ (CLTK), βομόν (CLTKb) and βομόν (GLEM)–, but corrected by AGILe from the wrongly predicted βομός to the right βωμός. The specific subject matter of the CGRN, with its focus on rituals of sacrifice, also enhanced vocabulary recognition of animal names (ὅϊς, αἶξ), place names and locations (Ἀθήνη, ἀκρόπολις, Μουνυχιών), names of gods (Λητώ, Βάκχιος), and words beginning with ἱερ-/ ἱαρ- (e.g. ἱαροποιοὶ and ἱερείων). The correct lemmatization of most wordforms of ὅϊς, even if 8 wordforms of this lemma are still wrongly lemmatized, brings an improvement in 84 cases if compared to the other tested lemmatizers and the correct identification of words beginning with ἱερ-/ ἱαρ- adds another 40 correct lemmas, always compared to the other tested lemmatizers. Another improvement is found for the 'more difficult' verb forms such as ἀναλώσει and ἐγδανεισάτωσαν, for which we report the lemmatizations produced by AGILe and the other tested lemmatizers. ἀναλώσει was wrongly lemmatized by the other tools as ἀναλώσω (UDPipe), ἀναλύω (CLTK), ἀναλώσει (CLTKb) and ἀνάλωσις (GLEM), while AGILe predicted the correct ἀναλίσκω; ἐγδανεισάτωσαν was also wrongly lemmatized by the other tools, as ἐγδανεισάτωσαν (UDPipe), γδανεισυτόω (CLTK), ἐγδανεισάτωσαν (CLTKb) and ἐγδανεισάτωσαν (GLEM), and only AGILe predicted the gold ἐκδανείζω.

6 Generalizability of AGILe

To better determine the specificity and the generalizability of AGILe, we also tested it on some literary data, as well as on epigraphic data other than the CGRN. To assess its performance on literary data, we tested AGILe on the PROIEL corpus (13,314 tokens) and obtained an accuracy of 73.6%. This is obviously lower than the accuracy obtained on CGRN and it is also lower than the accuracy reported on PROIEL for the UDPipe models (∼94%, see Table 1). This suggests that AGILe indeed specializes on inscriptions and that other models are still better suited for lemmatizing literary texts.

To further evaluate AGILe's performance on epigraphic data, we tested its performance on another available small corpus of inscriptions, the Cretan Institutional Inscriptions (Vagionakis, 2021).[30] This collection includes 600 inscriptions written between the 7th and the 1st century BCE. The provenance of most inscriptions is Crete itself, but the corpus also includes texts from other parts of the Greek and Roman world, if their content concerns Crete; the corpus covers roughly the same timespan as the CGRN and includes various types of texts, such as decrees, political treaties, dedications and epitaphs. The texts focus on Cretan institutions and political entities. There is thus a reasonable overlap with the CGRN in its geographical and chronological range, in the types of text, and some of its subject matter (mostly, when the inscriptions concern religious institutions, such as priesthoods), but there also are many differences. On these texts, AGILe achieved an accuracy of 62.2%. There were 2471 errors, of which 1336 unique (taking out repetitions of the same error); of these 1336 unique errors, 955 cases went wrong only once, 381 more than once. On these results we performed a detailed error analysis. For this purpose, we selected a portion of all the errors, by including all the most frequent ones (the errors occurring 9 or more times, for a total of 34 unique cases) and by randomly adding other 234 errors, for a total of 268 unique errors (including three cases in which AGILe did not output any prediction). In this way we analysed 838 of AGILe's errors on this dataset. It turned out that in 513 of these errors, AGILe is actually correct (61% of the initially reported 838 "errors"). The reported errors were in most cases only due to a difference in lemmatization conventions. For example, the authors of the Cretan Institutional Inscription corpus chose the Cretan dialectal norm as the lemma, e.g. τύχα for the form τύχαι instead of the LSJ form τύχη. The result was that AGILe's out-

[28] ἄφαμμα is the only one of the 21 cases of detrimental correction to occur more than once in the CGRN.

[29] This is particularly challenging to handle, since a wordform with crasis should be lemmatized to two lemmas; these cases were lemmatized in the CGRN as only one of the two, as in τἐνδοσθίδια (= τὰ ἐνδοσθίδια), which was lemmatized as ἐνδοσθίδια.)

[30] Repository at https://github.com/IreneVagionakis/CretanInscriptions/. The Corpus is also searchable through an interactive interface, at https://ilc4clarin.ilc.cnr.it/cretaninscriptions/en/texts/about.html.

put τύχη was evaluated by the system as 'wrong', though it is actually correct. In our random sample of 835 'errors', we counted 504 of such cases. There were 9 other cases in which we consider AGILe's output to be actually correct, though inconsistent with lemma in the CII corpus: when AGILe's output was capitalized, but CII's lemma was not or vice versa, when AGILe's form was in a different voice (e.g. active instead of the medio-passive in the CII), when the gold lemma in the CII was actually wrong.

If we generalize our impressions from the sample to the whole datase, the fact that in 61% of the initially reported errors in our sample the answer is actually correct, suggests that AGILe's accuracy is actually much better than 62.2% and that perhaps in up to six out of ten cases of the initially reported "errors" AGILE was in fact correct, achieving a hypothetical accuracy of 85% (= 62% + 0,61*38%), similar to our reported accuracy on the CGRN. We also tested GLEM on the Cretan Institutional Inscriptions corpus, as it was the best performing lemmatizer on the CGRN out of those presented in Section 3. On this new data, GLEM achieves an accuracy of 51.2%, which is ten points lower than that observed on the CGRN (61.5%), though also in GLEM's case, probably, many cases are false negatives, due to differences in choices during lemmatization. This suggests that (i) this dataset might be more difficult or in any case even further away than the CGRN in terms of language variety from the literary data GLEM was trained on; and that (ii) AGILe appears indeed more specialised than the other at dealing with inscription data. It should still be noted that we do not expect AGILe to perform equally well on inscriptions irrespective of time, place, and genre; at least, though, on the basis of the data it has been trained on, we hypothesize that AGILe will have a good performance on inscriptions from various parts of the Greek world until approximately the 1st century CE. However, such limitation in time is intrinsic to tools trained on language data, on both ancient and modern languages.

7 Conclusions

We have developed a lemmatizer specifically designed for ancient Greek inscriptions, AGILe, with improved performance on this kind of texts, in comparison to the available lemmatizers for ancient Greek. The lemmatizer has been released and it is open-access. We believe this is an important step towards facilitating the research of many Classics scholars all over the world. In this respect, we have already contacted the PHI, which hosts one of the most largely consulted corpora of ancient Greek inscriptions, to explore possibilities of integrating the tool in their platform.

Through a detailed error analysis, we showed that there is room for improving the accuracy of lemmatization, especially in the process of lexicon lookup; one option could be a weighted lookup, giving priority to the left part of the word, where the root generally is, to help the lemmatizer to select the correct lemma from the lexicon in the cases in which it contains more lemmas with the same edit distance from the predicted one, and a wrong, completely unrelated lemma is selected only because it comes before the correct one in the lexicon. However, this could introduce new errors elsewhere. Another possibility, once the lemmatizer is integrated in an online corpus of inscriptions, is to allow users to choose the correct lemmatization out of a set of options, following the example of the Perseus Digital Library, where users can select the best morphological analysis for each wordform when more are proposed.

Furthermore, an avenue that could be in theory explored for future improvement is to consider integrating other processing levels, such as POS-tagging. To integrate POS-tagging into the lemmatization process, a POS-tagger trained on ancient Greek inscriptions would be necessary and its training would require an adequate amount of (manually) labelled epigraphic data, which is at present unavailable. The currently available POS-taggers for ancient Greek are indeed all trained on literary texts, such as the one included in the CLTK pipeline for ancient Greek, based on Stanza and trained on data from the PROIEL treebank (Johnson et al., 2021, 23). For this tagger, no accuracy is reported. The GLEM lemmatizer does also leverage POS-information provided by the Frog POS-tagger, retrained on the text of Herodotus from the PROIEL treebank (Bary et al., 2017, 93). The accuracy of the Frog lemmatizer trained and tested on Herodotus is reported as 83.0%, against the 90.6% achieved by the GLEM POS-tagger, which improves Frog's performance by using a lexicon. When the two POS-taggers are trained on Herodotus and tested on Thucydides, the performance drops to 67.5% for Frog and 78.47% for GLEM, suggesting performance on inscriptions might be even lower. Moreover, we do not have information about how beneficial for GLEM's lemmatization POS-tagging is. At this stage it is unclear whether training a POS-tagger specifically on (and for) ancient Greek inscriptions would be successful, considering the characteristics of these texts, and therefore there concretely is room for improving lemmatization thanks to POS-tagging.

Lastly, we would like to mention that this work has highlighted a lack of standardization in gold standard creation for lemmatization across different corpora of ancient Greek, which surely hinders progress and must be urgently addressed by the community. To optimize interoperability of corpora and projects, and the development of yet more Machine Learning-based tools for the automatic processing of Ancient Greek, we would suggest to follow a single standard for lemmatization future corpora of inscriptions, by always using the base forms present in the LSJ.

8 Acknowledgements

This work is partially supported by the Young Academy Groningen (YAG) through a PhD scholarship for Silvia Stopponi. We are grateful to the anonymous LREC reviewers whose comments have contributed to improving a previous version of this paper.

This article was made possible through the financial support of Anchoring Innovation. Anchoring Innovation is the Gravitation Grant research agenda of the Dutch National Research School in Classical Studies, OIKOS. It is financially supported by the Dutch ministry of Education, Cul-

ture and Science (NWO project number 024.003.012). For more information about the research program and its results, see the website www.anchoringinnovation.nl.

9 References

Bary, C., Berck, P., and Hendrickx, I. (2017). A memory-based lemmatizer for Ancient Greek. In *ACM International Conference Proceeding Series*, volume Part F129473, pages 91–95, New York, NY, USA, jun. Association for Computing Machinery.

Bodel, J. P. (2001). *Epigraphic evidence: ancient history from inscriptions*. Psychology Press.

Burns, P. J. (2019). Building a Text Analysis Pipeline for Classical Languages. In Monica Berti, editor, *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, pages 159–176. De Gruyter Saur, Berlin, Boston.

Burns, P. J. (2020). Ensemble lemmatization with the Classical Language Toolkit. *Studi e Saggi Linguistici*, 58(1).

Celano, G. G. (2019). The dependency treebanks for ancient greek and latin. In Monica Berti, editor, *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, pages 279–298. De Gruyter Saur.

Colvin, S. (2007). *A historical Greek reader: Mycenaean to the koiné*. Oxford University Press.

Haug, D. T. and Jøhndal, M. (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)*, pages 27–34.

Hendrickx, I., van den Bosch, A., van Gompel, M., and van der Sloot, K. (2016). Frog, a natural language processing suite for dutch. *Language and Speech Technology Technical Report Series, Radboud University Nijmegen*, 16(2).

Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., and Mattingly, W. J. B. (2021). The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online, August. Association for Computational Linguistics.

Keersmaekers, A. (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*, 35(1):67–82, feb.

Keersmaekers, A. (2020). *A Computational Approach to the Greek Papyri: Developing a Corpus to Study Variation and Change in the Post-Classical Greek Complementation System*. KU Leuven.

Liddell, H. G. and Scott, R. (1940). *A Greek-English Lexicon. Revised and Augmented throughout by Sir Henry Stuart Jones with the Assistance of Roderick McKenzie*. Oxford: Clarendon Press.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bellato, S., Bengoetxea, K., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Sadde, S., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichi-

nava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Vincze, V., Wallin, L., Washington, J. N., Williams, S., Wirén, M., Woldemariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2018). Universal dependencies 2.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pantelia, M. C. (2001). Thesaurus linguae graecae digital library. *University of California, Irvine http://www. tlg. uci. edu¿.*, 2:2019.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Reynolds, L. D. and Wilson, N. G. (1991). *Scribes and scholars: A guide to the transmission of Greek and Latin literature*. Oxford: Clarendon Press. 3rd ed.

Smith, D. A., Rydberg-Cox, J. A., and Crane, G. R. (2000). The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

Straka, M., Straková, J., and Hajič, J. (2019a). Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. *ArXiv.org Computing Research Repository*.

Straka, M., Straková, J., and Hajič, J. (2019b). UDPipe at SIGMORPHON 2019: Contextualized Embeddings, Regularization with Morphological Categories, Corpora Merging. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Stroudsburg. Association for Computational Linguistics.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Vagionakis, I. (2021). Cretan institutional inscriptions: A new epidoc database. *Journal of the Text Encoding Initiative*.

Vatri, A. and McGillivray, B. (2018). The diorisis ancient greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65.

Vatri, A. and McGillivray, B. (2020). Lemmatization for ancient greek: An experimental assessment of the state of the art. *Journal of Greek Linguistics*, 20(2):179–196.