Lemmatization for Ancient Greek

An experimental assessment of the state of the art

Alessandro Vatri

University of Oxford, Oxford, UK

alessandro.vatri@classics.ox.ac.uk

Barbara McGillivray

University of Cambridge, Cambridge, UK;

The Alan Turing Institute, London, UK

bmcgillivray@turing.ac.uk

Abstract

This article presents the result of accuracy tests for currently available Ancient Greek lemmatizers and recently published lemmatized corpora. We ran a blinded experiment in which three highly proficient readers of Ancient Greek evaluated the output of the CLTK lemmatizer, of the CLTK backoff lemmatizer, and of GLEM, together with the lemmatizations offered by the Diorisis corpus and the Lemmatized Ancient Greek Texts repository. The texts chosen for this experiment are Homer, *Iliad* 1.1–279 and Lysias 7. The results suggest that lemmatization methods using large lexica as well as part-of-speech tagging—such as those employed by the Diorisis corpus and the CLTK backoff lemmatizer—are more reliable than methods that rely more heavily on machine learning and use smaller lexica.

Keywords

Ancient Greek digital corpora – Ancient Greek morphology – lemmatization – lemmatizer – usability of digital resources – digital lexicography

1 Introduction

The growing availability of texts in digital form is radically changing the landscape in humanities research, and ancient languages are no exception. Ancient Greek and Latin, in particular, benefit from a very active community of researchers enthusiastically engaged in creating digital resources and tools that have the potential to generate new research questions and transform scholarship practices. The Open Greek and Latin initiative,¹ for example, aims to collect texts and software for these languages and to make them openly-available to researchers and students. Developing in parallel with the creation of such resources, some researchers have also outlined new methodological frameworks (McGillivray 2013; Jensen & McGillivray 2017).

The scale of many of the available textual collections is such that traditional close reading needs to be complemented with so-called distant-reading approaches. The most basic and intuitive way to explore the texts is by searching them. Simple searches on raw texts, however, are impractical for morphologically-rich languages. For instance, a very common scenario concerns lemma searches in which the researchers specify a dictionary-entry form (or lemma, such as ἄνθρωπος) and all texts containing a word form of that lemma are returned (such as ἀνθρώπου, ἀνθρώπων, and so on), thus removing the need to specify all inflected forms or spelling variants. This allows researchers to detect and quantify large-scale patterns (e.g., collocations and n-grams defined by lexemes and not simply by strings of characters) and thus dramatically increase the efficiency of their data collection for the study of a wide range of linguistic phenomena.² Building on this, more advanced types of search focussing on syntactic or semantic aspects allow for even more sophisticated enquiries (McGillivray 2021).

We focus here on lemmatization, the process in which each word form in a text is associated to its lemma. In the case of Ancient Greek, this task can be particularly challenging due to this language's high inflectional variability. Moreover, lemmatization is expected to disambiguate ambiguous forms in context, for example by recognizing whether a form like πράξεις in a given passage is the active future indicative of the verb πράσσω or the nominative/accusative plural of the noun πράξις. To do this disambiguation, various features of the surrounding words need to be taken into account, which is not always a straightforward task for a computer program.

¹ <http://www.opengreekandlatin.org>.

² These would include, for instance, the study of word order pattern, of the syntactic behaviour of specific lexical items, of the argument structure of specific verbs, of the semantics of specific verbs in different voices, etc.

Lemmatization is achieved by annotating the text manually, semi-automatically, or automatically (with the help of computer programs called lemmatizers). The latter approach has obvious advantages in terms of time and replicability (McCarthy 2001: 2, 19–21), but can suffer from lower levels of accuracy than are sometimes acceptable. For these reasons, active research in developing and improving lemmatizers for ancient languages is critically important to scale up linguistic analyses and gain new insights.

We present a series of experiments aimed at assessing the accuracy of the state-of-the-art systems for Ancient Greek and briefly discuss the implications for future research directions.

2 State of the art

Several lemmatizers for Ancient Greek are currently available to the scholarly community; see Bary et al. (2017) for an overview. Because Bary et al. show that their system, GLEM (short, presumably, for Greek Lemmatizer), outperforms other state-of-the-art lemmatizers for Ancient Greek, we decided to focus on comparing GLEM with newer lemmatizers which have been developed since GLEM was created, and which are described in this section. Moreover, Gleim et al. (2019) report on lemmatization experiments for Latin and German. In this article, rather than focussing on how lemmatizers work, we assess the quality of lemmatized texts from the point of view of potential end-users (Greek linguists).

We start our overview from the Perseus project,³ which was historically one of the first initiatives to build tools for the automatic linguistic analysis of Ancient Greek (and Latin) texts. Gregory Crane has developed the morphological analyzer Morpheus (Crane 1991), which has been implemented on the Perseus Project digital library hosted at Tufts University. Morpheus returns all possible lemmas (and morphological analyses) for a given word form. The Perseus Under PhiloLogic project,⁴ hosted at the University of Chicago, provides a selection of the texts of the Perseus Project in the reading environment “PhiloLogic”, and implements some manual disambiguation for the Greek morphology and lemmatization. With the aim to (at least partially) disambiguate forms in context, Giuseppe Celano has automatically lemmatized the Canonical Greek literature texts from the Perseus Digital Library⁵ and the texts of the

³ <https://www.perseus.tufts.edu/hopper/>.

⁴ <http://perseus.uchicago.edu>.

⁵ <https://github.com/PerseusDL/canonical-greekLit/releases/tag/o.o.236>.

First Thousand Years of Greek Project,⁶ making use of the lemma information from Morpheus and Perseus Under PhiloLogic as well as part-of-speech information (Lemmatized Ancient Greek Texts [LAGT]).⁷ This approach works well to disambiguate those forms that can be associated to more than one lemma with different parts of speech.

On the other hand, the Classical Language Toolkit (CLTK; Johnson et al. 2014–2019) contains a lemmatizer that disambiguates between multiple lemmas by choosing the most frequent one according to its corpus frequency. With the aim to improve the results of the CLTK lemmatizer, Patrick Burns has developed a backoff lemmatizer (Burns 2019) modelled after the NLTK (Natural Language Toolkit) Backoff part-of-speech Tagger. This system runs different lemmatizers in sequence, so that if one is not able to lemmatize a form, another one is used instead, which gradually reduces the number of unanalyzed forms. The lemmatizers employed in this chain mechanism include: a dictionary-based one, which looks up high-frequency forms in a lexicon; one which learns patterns of co-occurrence via machine-learning systems trained on previously lemmatized data drawn from the Perseus Ancient Greek and Latin Dependency Treebank;⁸ one which relies on matching forms with certain morphological endings via regular expressions; a second dictionary-based lemmatizer using Morpheus lemmas, and one that simply fills gaps by returning the token as the lemma.

The GLEM lemmatizer (Bary et al. 2017) is built from Frog (Hendrickx et al. 2016), an open-source natural language processing tool based on memory-based algorithms (Daelemans & van den Bosch 2010). This machine-learning component is combined with a lexicon look-up in order to use part-of-speech information to disambiguate between multiple lemmas. The lexicon was obtained from the PROIEL (Pragmatic Resources in Old Indo-European Languages) project's annotated text of Herodotus⁹ and the annotated texts from the Perseus project. GLEM chooses the lemma of the most frequent word form/part-of-speech combination according to the PROIEL lexicon whenever a disambiguation purely based on part of speech is not possible. Bary et al. (2017) show that GLEM outperforms both a lemmatizer based on Frog and the CLTK lemmatizer on an unseen text, with an accuracy of 93.0 % vs. 75.6 % and 76.6 %, respectively, making it the state-of-the-art system for Ancient Greek in 2017.

6 <https://github.com/OpenGreekAndLatin/FirstKGreek/releases/tag/1.1.1802>.

7 <https://github.com/gcelano/LemmatizedAncientGreekXML>.

8 http://perseusdl.github.io/treebank_data.

9 <https://proiel.github.io>.

Diorisis (Vatri & McGillivray 2018) is the most recent contribution to the landscape of Ancient Greek lemmatizers. It was developed in the context of the Diorisis Ancient Greek corpus, which contains a large collection of automatically lemmatized and part-of-speech tagged texts from the Perseus Canonical Greek Literature repository, “The Little Sailing” digital library,¹⁰ and the Bibliotheca Augustana digital library.¹¹ Word forms in the texts were automatically matched against a parsed word-form list (counting 911,840 forms) produced by the Perseus Digital Library and included in the free software package Diogenes.¹² Forms were further tagged by part of speech using TreeTagger¹³ (Schmid 1994), trained on the Perseus Ancient Greek Dependency Treebank and the PROIEL project’s treebank (Haug & Jøhndal 2008).¹⁴ The part-of-speech information is exploited to disambiguate forms which may be associated to more than one possible lemma with different parts of speech.

In the rest of this article we describe an extensive evaluation framework that we have developed to assess the accuracy of two publicly available Ancient Greek (AG) lemmatized corpora (Diorisis, LAGT v1.2.5) and two lemmatizers (GLEM, CLTK).

3 Accuracy assessment

In this section we detail the results of two experiments carried out, with the aim to assess the accuracy of both publicly available Ancient Greek (AG) lemmatized corpora (Diorisis, LAGT) and lemmatizers (GLEM, CLTK).

First, we conducted a manual annotation of the lemmatization. We asked three highly proficient readers of Ancient Greek (one postdoctoral researcher, one PhD candidate, and one advanced MPhil candidate) to examine samples of lemmatized texts from each of these sources and assess which word-forms were lemmatized correctly in each corpus or lemmatizer output. After this task was completed (Experiment 1), further tests revealed that GLEM might have underperformed because it appeared to work more effectively if punctuation is removed from the texts to be lemmatized. In the meantime, a new version of the CLTK lemmatizer was released. Since the manual review of lemmatized samples had already resulted in the production of a ‘gold standard’ of

10 <http://www.mikrosapoplous.gr/en/textsien.html>.

11 <http://www.hs-augsburg.de/~harsch/augustana.html#gr>.

12 <https://community.dur.ac.uk/p.j.heslin/Software/Diogenes/>.

13 <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger>.

14 <https://proiel.github.io>.

acceptable lemmatizations for each lemma, the outputs of a re-run of GLEM on an unpunctuated version of the samples and of the new version of the CLTK lemmatizer were simply assessed against such a ‘gold standard’ and further reviewed manually (Experiment 2).

3.1 *Experiment 1*

3.1.1 Data

The accuracy of the output of lemmatizers and lemmatized corpora was compared through the manual examination of lemmatized word-forms in two samples. The texts selected for this purpose are Lysias’ speech *On the Olive Stump* (Lys. 7, which has 1,995 orthographic words) and *Iliad* 1.1–279 (2,074 orthographic words). These texts were chosen for a number of reasons: they are both available from Diorisis and LAGT, and in both corpora they are based on the digital texts released by the Perseus Digital Library under a Creative Commons Attribution-ShareAlike 3.0 United States License. As such, the same source texts could be fed to the GLEM and CLTK lemmatizers.

The first book of the *Iliad* diverges significantly from classical prose both in vocabulary and word order (due, among other things, to the constraints of metre). This makes for an instructive test case for the performance of a lemmatizer such as GLEM, which implements a machine learning tool (Frog) trained on a different variety of AG and uses a relatively limited lexicon for look-ups. We expect that these factors may affect its performance negatively in at least two ways. First, a number of lemmas and dialectal forms of the Homeric text may fail to be captured by the lexicon (even though many Ionic forms would be matched in a lexicon based on Herodotus). Secondly, a context-sensitive lemmatization algorithm (such as one based on probabilistic PoS-tagging or on the recognition of patterns seen in the training data) may underperform when word-order patterns differ significantly from those common in the training set. Conversely, both Diorisis and LAGT use a very comprehensive lexicon based on the Perseus Digital Library data covering much of the archaic and dialectal forms appearing in Homer. Another potential advantage for Diorisis lies in the fact that *Iliad* 1 is included in the Perseus AG Treebank, which was used to train the PoS-tagger employed in the production of this lemmatized corpus.

Lysias 7 is not part of the training set for either GLEM nor Diorisis, for which a PoS-tagger based on the Perseus Ancient Greek Treebank was used (Vatri & McGillivray 2018). Given that this speech is a prose text, one may expect a lemmatizer trained exclusively on a prose text (such as GLEM, which was trained on Herodotus) to perform reasonably well with ‘familiar’ word-order patterns (which would be an advantage for probabilistic or memory-based methods) and a vocabulary that, in principle, may overlap with that of historiography to

a greater extent than that of epic poetry. At the same time, however, the lexicon used by GLEM may not contain matches for a non-negligible number of Attic forms differing from their Ionic counterparts found in Herodotus, which may negatively affect the performance of GLEM.

3.1.2 Pre-processing

We extracted the raw text of both samples (punctuation included) from the corresponding Perseus Digital Library XML documents (<https://www.github.com/PerseusDL/canonical-greekLit>) and fed it to both GLEM and the CLTK lemmatizer. We then produced one spreadsheet per sample containing both the lemmas provided by Diorisis and LAGT and the output of GLEM and the CLTK lemmatizer for each word-form. In each row, column A contains the word-form and columns B to E contain the output lemmas in random order (that is to say, columns B to E do not each correspond to the output of a specific lemmatizer throughout the spreadsheet, but their order is randomized in each row). The purpose of randomization was to prevent manual reviewers from knowing which lemma was the output of which lemmatizer or lemmatized corpus. The key for assigning each output to its source was stored in a different sheet that reviewers were not able to see and was only used once the manual review of the data was completed.

3.1.3 Setup

Reviewers were each provided with identical copies of the spreadsheets (one for Lys. 7 and one for Hom. *Il.* 1.1–279) and were asked to change the font to bold for each cell containing a lemmatization that they judged to be wrong. For this purpose, they were instructed to:

- Disregard any numbers accompanying lemmas as a way to disambiguate between homophonous dictionary headwords (e.g. λέγω₁ should be considered equivalent to λέγω_{#1} or λέγω), since lexica used by different lemmatizers are not expected to be aligned.
- Disregard differences between deponent or active form lemmas for verbs (e.g., δέω vs. δέομαι) for the same reason.
- Disregard differences between adjectives and corresponding adverbs (lexica may or may not list them as separate entries).
- Mark outputs that do not appear to have been lemmatized at all as wrong. These include both those for which lemmatizers returned empty values or error messages and, more subtly, those corresponding to the form as it appears in the text and not to the dictionary entry (e.g., when a form like δὲ, with a grave instead of an acute accent, is returned as a lemma).

3.1.4 Agreement

When the reviewers returned the spreadsheet, Fleiss' κ was calculated as a measure of inter-subject agreement. For Lys. 7 the score was 0.922, while for Hom. *Il.* 1–279 it was 0.865. These scores indicate good overall agreement but prompted us to review cases of disagreement in order to produce a 'gold standard' against which to test the data from different corpora and lemmatizers.

3.1.5 Gold standard

Each lemmatization for which reviewers disagreed was reviewed manually. A good number of cases of disagreement were merely due to errors or slips on part of the reviewers, but others raised interpretive/lexicographical problems and needed to be resolved through the definition of more rigorous criteria. This exercise was also instructive in that it revealed how a significant category of potential end-users of lemmatizers and lemmatized corpora (i.e., general classicists, philologists, and linguists) might approach and attempt to resolve lexicographical dilemmas as they search digital corpora.¹⁵

Problematic cases included the following:

- Forms with diaeresis corresponding to a dictionary headword without diaeresis (e.g. Ἀτρεΐδης for Ἀτρεΐδης): if lemmatizers assign forms with diaeresis to a headword without diaeresis, lemmatization has evidently operated. Conversely, if lemmas of forms with diaeresis merely reproduce the word-form (which would only give a 'correct' lemmatization when the form corresponds to the dictionary headword, e.g., in the nominative singular for nouns or the first person singular of the present indicative for most verbs), one may not conclude that the form was lemmatized at all and the output should be marked as wrong.
- Enclitics lemmatized as orthotonics (e.g., πού): these forms do not inflect, and users of lemmatizers and lemmatized corpora may not know whether the lexica used for the identification of lemmas list the enclitic or the orthotonic form as the headword. In such cases, we thought it reasonable to adopt a usability criterion, namely, to what extent such ambiguities would mar the accuracy of corpus searches. In principle, a researcher could search the lemmatized texts for both (e.g.) πού and πού. Since one may not easily verify whether one or the other lemmatizer/lemmatized corpus consistently contains clitic or orthotonic forms for a given lemma in this category, it is advisable that a researcher should formulate a query searching for both

¹⁵ The evaluation sheets are available at https://figshare.com/articles/Lemmatizer_Evaluation_Sheets_Lys_7_Hom_Il_1-279_/12479675.

forms at the same time, in order to avoid excluding false negatives from the results. Such an inclusive approach might only overshoot (and produce false positives) if the distinction between enclitic and orthotonic forms is lexical, as is the case with τίς, πῶς, ποῦ, πῶ (interrogative) vs. τις, πῶς, που, πῶ (indefinite). That is to say, if a researcher were to look for occurrences of (e.g.) interrogative τίς in annotated texts in which orthotonic forms of the indefinite τις were lemmatized as τίς, the results would contain a potentially unacceptable number of false positives. This would not be the case with e.g. ποτέ and ποτε. As a consequence, the distinction between enclitic and orthotonic lemmatizations has been considered crucial only when lexical, but both lemmatizations have been considered acceptable for all other enclitics.

- We have only accepted clitic lemmas for proclitics appearing as orthotonic in the text. For instance, the only acceptable lemmatization of a form like εἰ (i.e., εἰ ‘if’ with an enclitic accent thrown back onto it) is εἰ; a lemmatization as εἰ simply looks like no lemmatization at all, and we would not expect researchers to search for both the clitic and the orthotonic form of such words.
- A number of forms of the demonstrative ὅς and of the article ὁ are homophonous and their distinction is merely based on their context. In principle, a good lemmatizer should be able to decide based on the function of a word in context. Therefore, this distinction has been adopted in the assessment of the output.
- Lemmatizations of adverbs as separate headwords from the corresponding adjectives (e.g., καλῶς or ὕστερον as their own lemmas) or as forms of the adjective (e.g., καλῶς as a form of καλός or ὕστερον as a form of ὕστερος) have all been considered acceptable, as mentioned above. However, the only lemmatization of the orthotonic form ὥς that we have considered acceptable is as ὥς (as in LSJ) and not as the adverb built on the stem of the relative pronoun ὅς. If a lemmatizer is able to simply assign -ως forms that it does not recognize to corresponding -ος adjectives, this could simply be the result of good guesswork (which, as far as ὥς is concerned, might occasionally yield the linguistically correct interpretation for this form as an adverb meaning ‘thus’). However, one wonders how likely it is that researchers would accept ὅς as a headword for ὥς. If a researcher were to search for orthotonic forms of ὥς as opposed to ὥς, s/he might more effectively search for instances of the (invariable) form instead of the lemma.
- Lemmatizations of suppletive verbs either as one or more headwords (e.g., εἶπον and ἐρῶ as their own headwords or as forms of λέγω) have all been

considered acceptable. The same applies to polythematic comparatives and superlatives with respect to the corresponding positive adjective (e.g., ἀμείνων, ἄριστος, and ἀγαθός).

- The same flexibility has been applied to the lemmatization of dialectal/poetic forms (e.g., γαῖα as its own headword or as a form of γῆ, or -ιη vs. -ια nouns) and forms compounded with the clitic particle περ (e.g., ὅσπερ as distinct or not from ὅς) or γε (e.g., ἔγωγε as distinct or not from ἐγώ). Analogously, no distinction has been made between the following pairs of headwords: ἀρήν and ἀρνός, ἀντιάζω and ἀντιάω, ἀτιμάζω and ἀτιμάω, πάτρη and πάτρα, ἰλάσκομαι and ἰλάσμαι, κοτέω and κοτάω, πόρω and πορεῖν, ἐπέοικα and ἐπέοικε, ὦ and ὦ, κουλεός and κολεόν, ἀποθνήσκω and θνήσκω/θνήσκω. The lemmatization Ἥρη (instead of Ἥρα) for corresponding forms in the text has not been accepted, since it is unlikely that the headword for this theonym is presented in its Ionic form in any dictionary.
- Both οὖ and ἐ have been accepted as headwords for the reflexive pronoun.
- Lemmatizations of ethnonyms such as Δαναοί as *pluralia tantum* have been accepted as correct (cf. LSJ s.v.).
- Lemmatizations of ἄγε as its own headword or as a form of ἄγω have all been accepted as correct. Analogously, lemmatizations of νεώτερος as its own lemma have been accepted, and so have those of οἴκοι as a form of οἶκος.
- We have also accepted lemmatizations of plural forms of personal pronouns as forms of the singular (e.g. forms of ἡμεῖς assigned to the headword ἐγώ).

As a result of the resolution of disagreements based on these criteria, we have compiled a set of acceptable lemmatizations for each lemma, e.g., ἐρέω, λέγω, and ἐρῶ for ἐρέω (*Il.* 1.76), δεῖ, δέω, and δέω₁ for ἔδει (*Lys.* 7.22), etc. In some cases, no lemmatizer or lemmatized corpus provided an acceptable lemmatization (e.g., for ῥῖε, *Il.* 1.47, with the wrong outputs αἰώ, αἰώ, and the entirely non-lemmatized ῥῖε). We found 12 such cases in the sample from Homer (0.58%) and 11 in the sample from Lysias (0.55%).

We also discovered two misspelled words in the sample from Lysias (ὥχετο for ὦχετο, *Lys.* 7.19, and ἔραττες for ἔπραττες, 7.21) and we have therefore reduced the word count of that sample to 1993 for the purpose of calculating accuracies.

3.1.6 Results

As a final step, we calculated the accuracy of each lemmatizer by matching the lemmatizations we judged correct with the corresponding source and computing the percentage of successes. The scores are presented in Table 1, which include binomial proportion 95% confidence intervals.

These scores suggest that Diorisis outperforms other corpora and lemmatizers on both samples. GLEM scores surprisingly low compared to the reported

TABLE 1 Accuracy of lemmatizers and lemmatized corpora

<i>Iliad</i> 1.1–279		Words	2074
<i>Lemmatizer</i>	<i>Successes</i>	<i>Accuracy</i>	<i>95 % C.I.</i>
Diorisis	1861	0.90	0.88–0.91
LAGT	1785	0.86	0.84–0.87
GLEM	1493	0.72	0.70–0.72
CLTK	1347	0.65	0.63–0.67

<i>Lysias</i> 7		Words	1993
<i>Lemmatizer</i>	<i>Successes</i>	<i>Accuracy</i>	<i>95 % C.I.</i>
Diorisis	1942	0.97	0.96–0.98
LAGT	1672	0.84	0.82–0.85
GLEM	1612	0.81	0.79–0.82
CLTK	1293	0.65	0.63–0.67

accuracy of 93 % on an unseen text reported by its developers (Bary et al. 2017). A post-hoc examination of the data revealed that GLEM did not automatically separate punctuation marks from the word-forms to which they are orthographically attached and was thus unable to lemmatize those forms correctly. This called for a re-run of GLEM on the samples and the comparison of its output with the gold standard.

3.2 Experiment 2

As we prepared to re-run GLEM and test its accuracy on both samples, we also became aware that the new CLTK backoff lemmatizer was released while we were completing Experiment 1 and we decided to test its accuracy in the same way. We automatically removed all punctuation marks from the sample texts and fed them to both GLEM and the CLTK backoff lemmatizer.¹⁶ We

¹⁶ The presence or absence of punctuation marks does not affect the performance of either version of the CLTK lemmatizer anyway.

TABLE 2 Accuracy of GLEM (rerun) and CLTK backoff lemmatizer

<i>Iliad</i> 1.1–279		Words	2074
<i>Lemmatizer</i>	<i>Successes</i>	<i>Accuracy</i>	<i>95 % C.I.</i>
CLTK	1879	0.91	0.89–0.92
GLEM	1750	0.84	0.83–0.86

<i>Lysias</i> 7		Words	1993
<i>Lemmatizer</i>	<i>Successes</i>	<i>Accuracy</i>	<i>95 % C.I.</i>
GLEM	1875	0.94	0.93–0.95
CLTK	1855	0.93	0.92–0.94

then automatically matched the respective outputs with the set of acceptable lemmatizations (the gold standard) resulting from Experiment 1 and manually reviewed all negative cases, including those for which no acceptable lemmatization had been produced. GLEM and CLTK output an acceptable lemmatization for a word form that no other method had achieved in four cases each. In seven cases, we added CLTK’s output to the gold standard (e.g., *πλείων* as the headword for *πλείω* alongside *πολύς*); the same was necessary six times for GLEM.

The results of this test are displayed in Table 2 and are presented alongside those of Experiment 1 in Figures 1 and 2 (see next page).

4 Error analysis

Analysing the errors of lemmatizers offers us the opportunity to identify their strengths and weaknesses and lay out an agenda for developers and curators to improve their lemmatization systems. As an initial exploration, we have classified the lemmatization errors of the Diorisis and LAGT corpora as well as those in the output of GLEM and the CLTK backoff lemmatizer. The distribution of errors across categories is sensitive to the token frequency of the types by which each category is defined, and the distribution of such frequencies may vary

FIGURE 1 Aggregate accuracy scores for Homer, *Il.* 1.1–279

FIGURE 2 Aggregate accuracy scores for Lysias 7

TABLE 3 Lemmatization error frequencies for Lys. 7

<i>Error type</i>	Lemmatizer				Total
	Diorisis	GLEM	CLTK	LAGT	
[1]		35	26	216	277
[2]	1	8	11	6	26
[3]			15	24	39
[4]				1	1
[5]	1	10	10	11	32
[6]			1		1
[7]	2		1	1	4
[8]		1		1	2
[9]		3			3
[10]	20	8	6	10	44
[11]	17	47	58	38	160
[12]	7	4	8	12	31
[13]	3	2	2	1	8
Total	51	118	138	321	628

considerably across text types. Based on this observation, we restricted this analysis to the sample from Lysias, on the assumption that the variety of Ancient Greek that it represents has wider dialectal and/or lexical overlaps with that of a great part of the extant Classical Greek texts than the sample of Homer. Incidentally, the average accuracy of the results for Lysias is higher (92%) than that of the results for Homer (88%). As mentioned above, this may be due to restrictions in the size of lexica and to the presence of dialectal forms and diacritics uncommon in prose (diaeresis) in the text of the *Iliad*.

- The following categories of errors emerged from the analysis (see Table 3):
- Errors due either (a) to failures to parse the input and assign it to a lemma or (b) to the lack of a match in the lexica (presumably because of their limited number of entries) [1]. The latter is most likely to be the case with unidentified proper names (not all proper names are included in lexica) [2]. Errors due to the failure to account for variants produced by connected speech phenomena (elision or apocope [3], crasis [4], enclitic accent or clitic forms of εἰμί [5]).

- Errors due to inaccuracies in the lexica and questionable lexicographical decisions. These include the assignment of forms of a simplex to a compound lemma (e.g., *μεμαχημένος* as a form of *συμμαχέω* in the output of CLTK) [6] and the inclusion in the lexica of forms that are not standardly used as headwords or of non-existent forms (e.g., ***ἐπέπειμι* in the Diorisis Corpus) [7].
- Errors due to what look like parsing mistakes (e.g., *ἴσασιν* as a form of *ἰσάζω* in the GLEM output) [8] and, in particular, to parses that seem to neglect the accent (e.g., *κέρδους* from *κερδῶ* instead of *κέρδος* in the output of GLEM; this form was mistaken for the gen. sg. *κερδοῦς*) [9].
- Errors due to homophony (e.g., *προσῆσαν* from *προσάδω* instead of *πρόσειμι* in the output of GLEM) [10]. A large number of these errors could be solved by accurate PoS disambiguation, which apparently failed for a number of forms in systems that included it in their lemmatization algorithm (e.g., *ἦ* was lemmatized as the particle *ἦ* instead of the imperfect 1sg of the verb *εἰμί*, in all outputs except that of CLTK) [11]. Among ambiguous forms prone to inaccurate lemmatization due to homophony, it is possible to identify at least two subsets. One is represented by ambiguous inflected forms of the demonstrative or relative *ὅς*, of the article *ὁ*, and of the reflexive pronoun *οἶ* (e.g., *οἱ* nom. pl. m. of *ὁ* vs. dat. sg. of *οἶ*); their assignment to the correct lemma is a task that has proved particularly difficult for automatic systems [12]. Another subset consists of the assignment of a contracted form of a verb in *-άω*, *-έω*, or *-όω* to the lemma corresponding to a by-form of the same verb showing a different predesinential vowel (e.g., *ἄξιῶ* as a form of the epigraphic by-form *ἄξιάω* instead of *ἄξιόω* in Diorisis and LAGT) [13].

The fact that category-1 errors are particularly frequent across outputs, but virtually absent from the Diorisis Corpus suggests that the automatic lemmatization of Classical Greek is most effective when larger and more comprehensive lexica (in terms of number of both inflected forms and lemmas) are available and used early in the lemmatization algorithm (as opposed to the late position of the Morpheus lexicon in the CLTK backoff chain). This must be combined with the effective handling of graphic variants that alter citation forms (elided, apocopated, contracted forms, clitic accentuation, and grave accents). The large number of category-11 errors suggests that the lemmatization of Ancient Greek is indeed sensitive to PoS-tagging and, as a consequence, would benefit greatly from the development of more accurate taggers. Failures in categories 2, 6, 7, and 13 could be minimized by increasing the quality of lexica.

5 Conclusions

The availability of high-quality automatic lemmatization capabilities is increasingly important to perform large-scale analyses of Ancient Greek texts, as it has the potential to reveal quantitative patterns and new insights into language usage. We have reported on an extensive evaluation of Ancient Greek lemmatizers currently available to the scholarly community.

A number of conclusions can be drawn from these experiments. According to our evaluation, the best results for the sample from Homer are those included in the Diorisis corpus and those produced by the CLTK backoff lemmatizer, whose accuracy scores are not significantly different (at the commonly adopted 0.05 level of significance). LAGT and GLEM (with the punctuation removed) come second and are equally accurate. When it comes to Lysias, Diorisis is a clear winner, with GLEM (without punctuation) and CLTK back-off coming second.

On a more general note, our results indicate that the size and breadth of lexica have a major impact on the efficiency of lemmatizers. Even though GLEM was trained on prose and incorporates a machine learning tool, it does not outperform Diorisis (which was produced with a simpler PoS-tagger and a very large lexicon) in the lemmatization of a classical prose text. Analogously, the CLTK backoff lemmatizer and Diorisis perform very well on the sample from Homer, for which a large lexicon is evidently required. These results call for the maximization and standardization of an AG digital lexicon to be adopted, ideally, by all research groups active in the development of AG lemmatizers.

Our findings suggest that it is already possible to employ the best lemmatizers (complemented with manual checks) to achieve good-quality analyses, and that Diorisis is currently the best resource for this purpose. This contribution has hopefully prepared the ground for even further improvements to the state of the art in automatic lemmatization of Ancient Greek.

Author contributions

Alessandro Vatri designed and ran the experiments, recruited participants, analysed the results, and is responsible for sections 3, 4, and 5.

Barbara McGillivray acquired the funding for the project, oversaw the study, contributed to the high-level design of the study and drafted sections 1 and 2. Both authors gave final approval for publication.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and under the seed funding grant SF042. We thank Martina Astrid Rodda and Roxanne Taylor for their collaboration.

References

- Bary, Corien, Peter Berck & Iris Hendrickx. 2017. A Memory-based lemmatizer for Ancient Greek. *DATECH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, 91–95. ACM Digital Library.
- Burns, Patrick J. 2019. Building a Text Analysis Pipeline for Classical Languages. *Digital Classical Philology*, ed. by Monica Berti, 159–176. Berlin/Boston: De Gruyter Saur.
- Crane, Gregory. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing* 6: 243–245.
- Daelemans, Walter & Antal van den Bosch. 2010. Memory-based learning. *The handbook of computational linguistics and natural language processing*, ed. by Alex Clark, Chris Fox & Shalom Lappin, 154–179. Oxford/Malden, MA: Wiley-Blackwell.
- Gleim, Rüdiger, Steffen Eger, Alexander Mehler, Tolga Uslu, Wahed Hemati, Andy Lücking, Alexander Henlein, Sven Kahlsdorf & Armin Hoenen. 2019. A practitioner's view: A survey and comparison of lemmatization and morphological tagging in German and Latin. *Journal of Language Modelling* 7: 1–52.
- Haug, Dag T.T. & Marius L. Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. *Proceedings of the second workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34 (www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf).
- Hendrickx, Iris, Antal van den Bosch, Maarten van Gompel, Ko van der Sloot, & Walter Daelemans. 2016. Frog, a natural language processing suite for Dutch. Technical report draft 0.13.1. *Radboud University Nijmegen. Language and speech technology technical report series* 16–02. [Retrieved from <https://github.com/LanguageMachine/frog/raw/master/docs/frogmanual.pdf>.]
- Jenset, Gard B. & Barbara McGillivray. 2017. *Quantitative historical linguistics. A corpus framework*. Oxford: Oxford University Press.
- Johnson, Kyle P. et al. 2014–2019. *CLTK: The classical languages toolkit*. <https://github.com/cltk/cltk>.
- McCarthy, Diana. 2001. Lexical acquisition at the syntax-semantics interface: Diathesis alternation, subcategorization frames and selectional preferences. Doctoral dissertation, University of Sussex.

- McGillivray, Barbara. 2014. *Methods in Latin computational linguistics*. Leiden/Boston: Brill.
- McGillivray, Barbara. 2020. Computational methods for semantic analysis of historical texts. *Routledge international handbook of research methods in digital humanities*, ed. by Kristen Schuster & Stuart Dunn. Abingdon/New York: Routledge.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. Manchester (UK).
- Vatri, Alessandro & Barbara McGillivray. 2018. The Diorisis Ancient Greek corpus. *Research Data Journal for the Humanities and Social Sciences* 3: 55–65.