Evaluating Existing Lemmatisers on Unedited Byzantine Greek Poetry

Colin Swaelens[1], **Ilse De Vos**[2] and **Els Lefever**[1]
[1] LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication, Ghent University
[2] Department of Linguistics, Ghent University
9000 Ghent, Belgium
{colin.swaelens, i.devos, els.lefever}@ugent.be

Abstract

This paper reports on the results of a comparative evaluation of four existing lemmatizers, all pre-trained on Ancient Greek texts, on a novel corpus of unedited, Byzantine Greek texts. The aim of this study is to get insights into the pitfalls of existing lemmatisation approaches as well as the specific challenges of our Byzantine Greek corpus, in order to develop a new lemmatizer that can cope with its peculiarities. The results of the experiment show an accuracy drop of 20% on our corpus, which is further investigated in a qualitative error analysis.

1 Introduction

If Ancient Greek is considered a low-resourced language, Byzantine Greek is even lower-resourced. What Ancient and Byzantine Greek have in common, is that their texts have been continuously copied by hand until the end of the 15th century. So when we read, for instance, Plato's *Apology*, we read a collation of a philologist who aspires to reconstruct Plato's original 4th-century text *based on* the existing Medieval manuscripts; *based on* but not copied from these manuscripts, as linguistic inconsistencies or orthographic mistakes are adapted to fit the dialect in which the text was conceived. Existing NLP tools for historical Greek are developed for this variant of Greek, that was edited to perfection.

However, because of a growing research interest and progress in optical character recognition (OCR) and, even more relevant, handwritten text recognition (HTR) (e.g. Tsochatzidis et al. 2021; Platanou et al. 2022; Ströbel et al. 2022), more and more unedited Greek texts will become available. These unedited texts contain, among other things, lacunae due to a damaged piece of parchment, omissions of words due to sloppiness or fatigue of the scribe or funky orthography due to phonetic changes. Although no substantial HTR-based corpus is available for Greek, two online available corpora do offer the unedited texts from manuscripts: the Trismegistos (Depauw and Gheldof, 2014) project and the Database of Byzantine Book Epigrams (DBBE) (Ricceri et al., 2023). Both Trismegistos and DBBE do store the edited as well as the unedited version of texts found in papyri and manuscripts, respectively. The DBBE provides Byzantine[1] book epigrams, which are metrical paratexts as they are written in the margin, next to (παρά, para) the main text of a manuscript. The literal transcription of these poems are stored as so-called *Occurrences*, which are linked to a normalised version called *Type*.

Our aim is to develop a linguistic annotation pipeline for the latter, unedited Greek texts. The differences between Ancient and Medieval Greek are thoroughly described by Swaelens et al. (Forthcoming 2023), the features relevant for this work are elaborated upon in Section 3. A new approach for part-of-speech tagging and morphological analysis was developed (Swaelens et al., 2023), as the existing techniques are not capable of handling the idiosyncrasies these unedited texts display. Before diving into the development of the last step of the pipeline, i.e. the lemmatizer, we wanted to evaluate existing systems for lemmatisation on our gold standard of unedited, Byzantine Greek texts.

2 Related Research

The first lemmatizer for Greek was developed by Packard (1973), as part of the first lin-

[1] Byzantine and Medieval will be used as synonyms to refer to the period from the 5th until 15th century.

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 111–116,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_013

guistic annotation pipeline. In order to perform morphological analysis, first suffixes are removed to retrieve the stem of every token. Then, a dictionary made by Packard, is searched with a binary search algorithm to find the matching stem. Based on this dictionary search, the algorithm returns the lemma that is linked to the matching stem. If multiple lemmas are possible, a philologist is needed to discern which lemma was the correct one.

In 2003, the biggest online resource of Greek texts, the Thesaurus Linguae Graecae (TLG) (Pantelia, 2022) started their lemmatization project. Few details on the methodology are provided in the paper, except that the TLG *digitised and extracted a large number of headwords from dictionaries*[2]. The authors, however, claim that the lemmatizer is capable of recognising automatically 98.3% of all tokens in the TLG.

RNN Tagger (Schmid, 2019) was developed as the combination of a morphological tagger and lemmatizer for historical texts. Schmid has made use of a character-based bi-LSTM network to cope with – systematic – spelling variations and improve tagging accuracy. The lemmatizer is also based on a recurrent neural network, making use of the dl4mt machine translation system (He et al., 2016). In his experiments, Schmid did also train and test his tagger on the Ancient Greek Dependency Treebank, which resulted in a tagging accuracy of 91.29%.

The Classical Language Toolkit (CLTK), is an NLP framework developed for pre-modern languages (Johnson et al., 2021). This framework stores several lemmatizers, among which a back-off lemmatizer (Burns, 2020) that makes use of several, sequenced lemmatizers. CLTK's default lemmatizer for Ancient Greek makes use of the Stanza (Qi et al., 2020) lemmatization algorithm, that has been pretrained on the PROIEL treebanks (Haug and Jøhndal, 2008). This algorithm consists of a dictionary-based lemmatizer combined with a neural sequence-to-sequence lemmatizer. On the encoder's output of this combination, an additional classifier is added to cope with, among other things, lowercasing. The authors, however, did not provide an accuracy score of how well the algorithm performs on Ancient Greek.

Burns' back-off lemmatizer, which is included in the CLTK, is a sequence of five lemmatizers. The token first passes a dictionary-based lemmatizer to tag frequently occurring, indeclinable words; then it passes through a unigram-model lemmatizer that is based on training data of the Ancient Greek and Latin Dependency Treebanks (Celano, 2019); third in the sequence is a rule-based lemmatizer that makes use of regular expressions; the fourth lemmatizer is a variation of the previous, regular expression-based lemmatizer that factors in principal-part information; finally, the token is passed through another dictionary-based lemmatizer making use of Morpheus' (Crane, 1991) lemma dictionary. Should none of these lemmatizers output a proper lemma, the token itself is returned as lemma. Vatri and McGillivray (2020) report an accuracy of 91% on poetry and 93% on prose.

Where CLTK's default lemmatizer disambiguates ambiguous tokens based on frequency, the GLEM lemmatizer (Bary et al., 2017) makes use of part-of-speech information to disambiguate. Even more interesting, is that GLEM provides a lemma for out-of-vocabulary words. This is achieved by combining a dictionary-based approach with a memory-based machine learning algorithm, called FROG (Bosch et al., 2007). If the to-be-tagged word occurs only once in the lexicon, consisting of the PROIEL and Perseus (Celano, 2019) corpora, GLEM returns the lexicon's lemma; if not, the word is considered ambiguous and FROG is applied. If several lemmas are possible, GLEM evaluates whether there is exactly one match with the part-of-speech tag predicted by the FROG algorithm and the lexicon. If so, the lemma is assigned; if there are several possible or no matching part-of-speech tags, frequency information is used to assign a lemma from the lexicon.

The interest in lemmatizing Greek has increased, proved by Keersmaekers and Van Hal (2022) and de Graaf et al. (2022). That is, both articles discover how corpora which cannot be lumped together with classical, literary Greek prose, could be lemmatized. Keersmaekers and Van Hal, on the one hand, aim to

[2] https://stephanus.tlg.uci.edu/history.php

lemmatize the papyri texts stored in Trismegistos, de Graaf et al., on the other hand, look into lemmatizing Greek inscriptions. Just like the unedited texts we want to tag, these corpora had some peculiarities that deviate from the polished, classical Greek on which the existing lemmatizers are based.

Although several other lemmatizers do exist, they are not part of this assessment because they are either not freely available or do not disambiguate ambiguous word forms. We did not test TreeTagger (Schmid, 1994) since the parameter files[3] do not contain any information on lemmas. Neither Morpheus (Crane, 1991) nor Eleuxis[4] have been part of our comparison as neither of those disambiguate ambiguous tokens.

3 Comparative Experiment

To evaluate the lemmatizers described in Section 2, we annotated about 10,000 tokens from the DBBE *occurrences* (Swaelens et al., 2023). The DBBE *occurrences* are the literal transcription, viz. without any editing, of the text that is found in a manuscript. As already mentioned in Section 1, these *occurrences* are linked to edited, normalised versions called *Types*, as shown in Example 1. Example 1a shows the *occurrence*, the text as it is found in the manuscript Vat.gr.1908[5], Example 1b the *Type* to which the *Occurrence* is linked and its translation (translated by the authors) is given in Example 1c.

(1) a. ὣς περ᾽ ξένη χἔρονἰτες ἡδῆν π(ατ)ρίδα
DBBE Occurrence 17870

b. Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα
DBBE Type 2820

c. Just like travellers rejoice upon seeing their homeland

This example displays one of the main characteristics of the Greek found in manuscripts before they are edited: orthographic inconsistencies. Since the itacism – a phonetic shift that turned η, ι, υ, ει and οι into the phoneme /i/ – has made its introduction in the 3rd century, quite some orthographic inconsistencies are to be found in the manuscripts. In Example 1a both the first syllable, ἡδ- (the stem of the word), and the second, -ῆν (the suffix indicating the Greek infinitive), are affected by the itacism. This makes the word ἰδεῖν almost unrecognisable, which is why we hypothesise that a dictionary-based approach might be put at a disadvantage.

For our comparative study, all lemmatizers discussed in Section 2, CLTK, GLEM, RNN Tagger, and the Stanza tagger, are used to lemmatize our gold standard containing 10,000 tokens of unedited, Byzantine Greek text. Before feeding the data to the lemmatizers, we removed all redundant white spaces and deleted all punctuation.

4 Results

The results of the comparitive experiments are shown in Table 1. First of all, We observe a general accuracy drop of 20% or more compared to the results of the lemmatizers on Ancient, edited Greek. This was expected, because our data is very challenging. Second, the sequential back-off lemmatizer comes out best, performing almost 7% better than the Stanza lemmatizer, which performed worst. To gain more insight in the results of the tested lemmatizers,we performed a qualitative analysis of the system output, which revealed some tendencies of the problems related to our corpus.

Lemmatizer	Accuracy
Stanza	64.99%
RNN Tagger	66.67%
GLEM	70.82%
CLTK	71.69%

Table 1: Performance of existing lemmatizers on Byzantine Greek poetry.

This comparative study uncovered an encoding problem in our test set: the transcriptions of the manuscripts stored in DBBE make use of multiple unicode characters for identi-

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

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

[5]This book epigram is situated on f.118v and online consultable via https://digi.vatlib.it/view/MSS_Vat.gr.1908/0121

cal characters. The acute accent, for example, is present in the DBBE as two different unicode characters. That is, the ί in πατρίδα (Example 1) has two different unicode representations within the DBBE corpus. Consequently, every deviation from the unicode character that is stored in DBBE or its annotations has been evaluated as incorrect. What is more, the Stanza lemmatizer outputs unicode characters that are different from those CLTK and RNN Tagger output.

The diachronic and/or diatopic alterations that are inherent to the Greek language, hinders the evaluation of the taggers as well. Verbs whose stem ends in a velar occlusive, have a lemma that ends either in -ττω (the classic, Athenian variant), or -σσω (other dialects' variant). The token φύλαττε (*keep guard*) has been annotated as coming from the lemma φυλάττω, while all lemmatizers returned φυλάσσω as lemma. Although this is a correct prediction, it was considered as incorrect by the automatic evaluation. In this same category belongs the alteration between ι and υ, observable in the – identical – words βίβλος and βύβλος (*papyrus roll*).[6] The alteration of a word's final consonant, is the last example that fits within this category. The preposition ἐκ (*out*) is written as ἐξ when followed by a vowel. Again, these double forms caused unjust penalties in the lemmatizers' output. In order to cope with these inconsistencies, we harmonised the different outputs, mainly caused by the unicode difference between the *tonos* and *oxia* accent (Tauber, 2019). The new lemmatisation results, however, show a minor impact of the encoding problems and inconsistencies, resulting in improvements of only 0.04 to 0.6 %, which makes no difference for the final ranking of the tested lemmatizers.

The lemmatizers also have a hard time assigning the correct lemma to a verb in the perfect tense. This might be due to the very low presence of this tense in general in Greek. It is, however, surprising that the back-off lemmatizer cannot extract and match the stem of, e.g., πεφευγώς (*having fled*) to its lemma φεύγω (*to flee*). What is even more surprising, is that GLEM did not even return a lemma of this quite frequent word, while it was stated that GLEM could output lemmas it had never seen before.

A GLEM-specific remark is how much this lemmatizer is affected by the absence of the iota subscriptum[7] in, e.g., the dative case. In DBBE this iota is sometimes written, now underneath the vowel, then next to it, and sometimes not written. Not once did it correctly lemmatize τω as a form of the article ὁ, while τῷ has been lemmatized correctly. The iota adscriptum is not yet part of the test set.

5 Conclusion & Future Research

As a last step in the development of our new annotation pipeline that cannot only handle classical Greek texts but also unedited, Byzantine texts, we are exploring the field of lemmatizing Greek. We compared four freely available lemmatizers that are capable of coping with ambiguity: CLTK back-off lemmatizer, GLEM, RNN Tagger and the Stanza lemmatizer. The back-off lemmatizer performed best, which might be attributed to the fact that it combines five different lemmatizers. The error analysis provided us with useful insights, which we will take into account while developing our own lemmatizer for Byzantine Greek.

At the moment of writing, we are looking into a cascaded system that combines a rule-based module with a dictionary look-up as a first step. In addition, a machine-learning component will be developed to handle all tokens that cannot be lemmatized by the first part. We are investigating several possible algorithms, going from a decision tree model to a neural approach. Furthermore, we will need to cope with the abundance of unicode characters and provide a mapping to evaluate our output correctly. We also need to develop a strategy to deal with the alterations that are inherent to the language to make evaluation easier and more correct, namely a mapping of (1) the five ways to write the /i/ sound, (2) the iota subscriptum or adscriptum and (3) forms like -σσω/-ττω. Finally, we will experiment with the presence or absence of diacritics and their possible impact on the machine learning.

[6] This alteration is not to be confused with the itacism; this alteration is already attested before the itacism appeared.

[7] When a long vowel is followed by a iota, ι /j/, the iota is written either underneath (*subscriptum*) or next to that vowel (*adscriptum*).

References

Corien Bary, Peter Berck, and Iris Hendrickx. 2017. A memory-based lemmatizer for ancient greek. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2017, page 91–95, New York, NY, USA. Association for Computing Machinery.

Antal van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for dutch. *LOT Occasional Series*, 7:191–206.

Patrick J Burns. 2020. Ensemble lemmatization with the classical language toolkit. *Studi e Saggi Linguistici*, 58(1):157–176.

Giuseppe G.A. Celano. 2019. *The Dependency Treebanks for Ancient Greek and Latin*, pages 279–298. De Gruyter Saur, Berlin, Boston.

Gregory R. Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5334–5344, Marseille, France. European Language Resources Association.

Mark Depauw and Tom Gheldof. 2014. Trismegistos: An interdisciplinary platform for ancient world texts and related information. In *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, pages 40–52, Cham. Springer International Publishing.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 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. Association for Computational Linguistics.

Alek Keersmaekers and Toon Van Hal. 2022. In search of the flocks: How to perform onomasiological queries in an Ancient Greek corpus? In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 73–83, Marseille, France. European Language Resources Association.

David W. Packard. 1973. Computer-assisted morphological analysis of Ancient Greek. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*.

Maria C. Pantelia. 2022. *Thesaurus Linguae Graecae, A Bibliographic Guide to the Canon of Greek Authors and Works*. University of California Press, Berkeley.

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022. Handwritten paleographic Greek text recognition: A century-based approach. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 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*.

Rachele Ricceri, Klaas Bentein, Floris Bernard, Antoon Bronselaer, Els De Paermentier, Pieterjan De Potter, Guy De Tré, Ilse De Vos, Maxime Deforche, Kristoffel Demoen, Els Lefever, Anne-Sophie Rouckhout, and Colin Swaelens. 2023. The database of byzantine book epigrams project: Principles, challenges, opportunities. *Journal of Data Mining & Digital Humanities*.

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

Helmut Schmid. 2019. Deep learning-based morphological taggers and lemmatizers for annotating historical texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 133–137, New York, NY, USA. Association for Computing Machinery.

Phillip Benjamin Ströbel, Martin Volk, Simon Clematide, Raphael Schwitter, Tobias Hodel, and David Schoch. 2022. Evaluation of HTR models without ground truth material. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4395–4404, Marseille, France. European Language Resources Association.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023. Medieval social media: Manual and automatic annotation of byzantine Greek marginal writing. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 1–9, Toronto, Canada. Association for Computational Linguistics.

Colin Swaelens, Ilse De Vos, and Els Lefever. Forthcoming 2023. Linguistic annotation of byzantine book epigrams. *Language Resources and Evaluation*.

James K. Tauber. 2019. *Character Encoding of Classical Languages*, pages 137–158. De Gruyter Saur, Berlin, Boston.

Lazaros Tsochatzidis, Symeon Symeonidis, Alexandros Papazoglou, and Ioannis Pratikakis. 2021. Htr for greek historical handwritten documents. *Journal of Imaging*, 7(12).

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