Lemmatisation of Medieval Greek: Against the Limits of Transformers’ Capabilities?

Colin Swaelens[†], Pranaydeep Singh[†], Ilse De Vos[*], Els Lefever[†]
[†]Language Technology & Translation Team, [*]VAIA
[†] Groot-Brittanniëlaan 45, Gent, Belgium, [*]Kasteelpark Arenberg 10, Heverlee, Belgium
{author1, author 2, author4}@ugent.be, author3@kuleuven.be

Abstract

This paper presents preliminary experiments for the lemmatisation of unedited, Byzantine Greek epigrams. This type of Greek is quite different from its classical ancestor, mostly because of its orthographic inconsistencies. Existing lemmatisation algorithms display an accuracy drop of around 30pp when tested on these Byzantine book epigrams. We conducted seven different lemmatisation experiments, which were either transformer-based or based on neural edit-trees. The best performing lemmatiser was a hybrid method combining transformer-based embeddings with a dictionary look-up. We compare our results with existing lemmatisers, and provide a detailed error analysis revealing why unedited, Byzantine Greek is so challenging for lemmatisation.

Keywords: Natural Language Processing, Lemmatisation, Byzantine Greek

1. Introduction

Recent developments in the field of natural language processing resulted in great advances in ancient language processing, which is becoming a thriving research field (Riemenschneider and Frank, 2023; Sommerschield et al., 2023). Multiple ancient languages, among which Latin (Mercelis and Keersmaekers, 2022) and languages written in Cuneiform (Sahala et al., 2020), are a topic of interest within the NLP community. Classical Greek is no exception to this. Nevertheless, research has shown that current NLP approaches often suffer from a large decrease in performance when applied to Greek texts that deviate from the classical, literary texts (de Graaf et al., 2022; Swaelens et al., 2023c). In this work, various linguistic pre-processing tasks have been assessed and all display this drop in performance.

In this paper, we want to investigate the capabilities of state-of-the-art transformer-based approaches to lemmatise unedited, Byzantine Greek texts. Lemmatisation or the assignment of a headword to a given token is far from a trivial task when carried out on a non-standardised low-resource language. Ancient Greek might be considered a low-resource language due to its little amount of (available) data and its closed nature, i.e. the corpus is not growing anymore since the language is dead. Byzantine Greek should then be deemed even lower-resourced as it only makes up a small part of the pre-modern Greek corpus, even though it spans a period of nearly ten centuries.[1] Previous NLP experiments (Swaelens et al., 2023a) have shown that the algorithms developed for Classical Greek, are not well suited to perform the same tasks on Byzantine Greek. We investigate why these recent language processing techniques have such a hard time lemmatising Byzantine Greek, a morphological complex language characterised by a horde of phonetic changes.

The remainder of this paper is organised as follows. After a literature review (Section 2), we elaborate on the data (Section 3) that is used in the experiments (Section 4). We provide an extensive error analysis (Section 5), followed by a conclusion and possible directions for future research (Section 6).

2. Related Research

The interest in (assigning) Greek lemmas has known a steep increase over the recent years (de Graaf et al., 2022; Keersmaekers and Van Hal, 2022). The very first lemmatiser for Greek was developed by Packard (1973), as part of the first morphological analysis tool. This lemmatiser was a dictionary-based system: a binary search algorithm matched the stem, outputted by Packard’s morphological analysis tool, to the corresponding stem-lemma pair in the dictionary that was included in the system. In case of ambiguity, as illustrated in Table 1, a domain expert was needed to assign the correct headword.

Stem	**Lemma**
ιδ- *id-*	ὁράω *horaō* ’to see’
ιδ- *id-*	οἶδα *oida* ’to know’

Table 1: Example of ambiguous stem-lemma pair

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

LREC-COLING 2024, pages 10293–10302
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

This system was followed by the still widely-used Morpheus (Crane, 1991), a rule-based approach that further elaborated the work of Packard to perform morphological analysis of classical Greek. The lemmatisation component of Morpheus used a dictionary-based approach too.

The first machine learning approach to tackle lemmatisation for Greek is TreeTagger (Schmid, 1991). TreeTagger, a Markov Model tagger that makes use of a decision tree to better estimate contextual parameters, was initially trained to provide part-of-speech tags and lemmas for English. In a next stage, Schmid extended his TreeTagger to tag German as well as English (Schmid, 1999). At the time of writing, TreeTagger analyses almost 30 languages and is adaptable to other languages.

Some years later, Schmid (2019) developed RNN Tagger, a combination of a morphological tagger and a lemmatiser, this time developed specifically for historical languages. Instead of using decision trees, RNN Tagger combines a character-based bi-LSTM network and a recurrent neural network (RNN), making use of the dl4mt machine translation system (He et al., 2016). Schmid experimented with classical Greek data, provided by the Ancient Greek Dependency Treebank (AGDT) (Celano, 2019), and reports an accuracy of 91.29%.

The Classical Language Toolkit (CLTK) (Johnson et al., 2021) holds, alongside a tokeniser, part-of-speech tagger, and morphological analysis tool, also two lemmatisers. CLTK's default lemmatiser for classical Greek makes use of the Stanza lemmatisation algorithm (Qi et al., 2020), which has been pre-trained on the PROIEL treebanks (Haug and Jøhndal, 2008). The algorithm combines a dictionary-based approach with a neural sequence-to-sequence approach. A classifier is added to the encoder's output to cope with orthography issues, like lowercasing. Accuracy scores of the Stanza lemmatiser on classical Greek are not reported by the authors. The second lemmatiser in the CLTK library is the back-off lemmatiser, developed by Burns (2020). This lemmatiser is a sequence of five algorithms: (1) a dictionary-based algorithm to tag frequently occurring indeclinable words; (2) a unigram-model lemmatiser trained on the aforementioned AGDT; (3) a rule-based lemmatiser based on regular expressions; (4) a variation of the previous, regular expression based lemmatiser that factors in principal-part information; (5) a dictionary-based lemmatiser that makes use of the lemma dictionary included in Morpheus. If none of the five algorithms outputs a lemma, the token itself is returned as lemma. In their assessment of state-of-the-art lemmatisers, Vatri and McGillivray (2020) report an accuracy of 91% on classical Greek poetry and 93% on classical Greek prose.

The GLEM lemmatiser (Bary et al., 2017), in its turn, combines a dictionary-based approach and a memory-based machine learning algorithm, FROG (Bosch et al., 2007). This approach should make GLEM capable of assigning lemmas to out-of-vocabulary words. GLEM initially tries to match the token to a lexicon, made up of PROIEL and the AGDT; if successful, the according lemma is returned, if not, FROG is applied. That means that FROG predicts the part-of-speech of the to-be-lemmatised token, after which GLEM evaluates whether this token and part-of-speech combination has exactly one match in the lexicon. If so, the according lemma is assigned; if not, frequency information is used to assign a lemma from the lexicon. The authors report an accuracy score of 93% on classical Greek prose.

The systems described so far are all developed for and trained on classical Greek literature. The last approach discussed in this literature review however was not. de Graaf et al. (2022) developed a lemmatiser to label Greek inscriptions. These inscriptions are characterised by the use of different alphabets, large dialectal variation and inconsistent orthography. This kind of texts are, just like book epigrams, autographs: they have been carved in stone once, without any subsequent interventions. The Stanza lemmatiser was trained on the inscriptions' data and complemented with an optional lexicon lookup in a lexicon that combines the *Liddel-Scott-Jones Greek-English Lexicon* (LSJ) (George and Liddell, 1968) and the gold lemmas from the training set. The authors report an accuracy score of 85.1% on the Collection of Greek Ritual Norms (Carbon et al., 2017) and 62.2% on the Cretan Institutional Inscriptions (Vagionakis, 2021).

More recently, a neural edit-tree lemmatiser (De Kok, 2021) was developed within the spaCy framework. Where common lemmatisation techniques consist of a rule- and/or dictionary-based approach, the neural edit-tree lemmatiser learns to predict lemmatisation rules from a training corpus. This makes the manual writing of rules unnecessary. The fundamental principle is straightforward: (1) identify the longest common sub-string (LCS) of a token and its lemma, (2) split the token in prefix, LCS and suffix, (3) find the edits to be made to the prefix and suffix to go from token (σβέννυσι *sbennusi*) to lemma (σβέννυμι *sbennumi* (to put out)).

1. σβέννυσι - σβέννυμι

2. σβέννυ (LCS) + -σι (suffix)

3. replace -σι by -μι

This is a simplified representation because multiple, shared substrings may occur, which is accounted for by a recursive algorithm. The reason this approach is the odd one out, is because this algorithm has not yet been trained on Greek and thus no accuracy scores can be reported.

For completeness, we conclude this literature review with the Thesaurus Linguae Graecae (TLG) (Pantelia, 2022), the largest digital corpus of Greek texts written between Homer (800 B.C.) and the fall of Byzantium (1453 A.D.). The TLG provides a lemmatised search engine since 2006 and although their lemmatiser should be capable of lemmatising 98% of all Greek word forms it has seen, no information on its development is provided. The lemmatiser itself is not freely available either.

3. Resources

3.1. Byzantine Book Epigrams

Book epigrams are metrical paratexts, i.e. poems standing next to (παρά para) the main text of a manuscript, written by the person who was copying or simply reading that main text. They are consequently all autographs. They are the thoughts of a scribe wrapped in an elegant poem in the margin of a manuscript, speaking to us directly from the past. Classical texts on the contrary have come to the 21st century indirectly. On top of the copying process, they have been edited and revised by philologists attempting to reconstruct, as good as possible, the so-called *Uhrtext*. Since book epigrams have not been edited, they display quite some orthographic inconsistencies (cf. de Graaf et al.'s inscriptions), of which the itacism is the most notable. The itacism is the shift of the classical Athenian pronunciation of four vowels (ι *i*, η *è*, ε *e*, υ *u*) and two diphthongs (ει *ei*, οι *oi*) to one and the same [i] sound. This made it quite hard for the scribes – some of which were not too acquainted with Greek – to know which [i] should be reflected in the typeface. It is also noteworthy that not one *classical Greek* existed, but that every region had its own dialect until Alexander the Great (ca. 300 B.C.). It is remarkable that, while developing language technology for Greek, some researchers only test on Attic prose and conclude that the algorithm works perfectly fine for "classical Greek", although their test set covers only one town at one given moment in time.

At the time of writing, the Database of Byzantine Book Epigrams (DBBE) (Ricceri et al., 2023) stores 12,192 book epigrams. They are referred to as *Occurrences*, since they display the epigrams exactly as they *occur* in the manuscripts. This means that no editing whatsoever took place during the digitisation of the poems. In addition, the DBBE aims to provide an edited, more readable version of each Occurrence. These records are called *Types*. The Types serve as a kind of representative or umbrella for the Occurrences, since one Type can represent several Occurrences and one Occurrence might be linked to multiple Types. The DBBE currently stores 4,924 Type records. Example 1 shows an Occurrence (1a) with its corresponding Type (1b) and English translation (1c).

(1) a. ὦς περ᾽ ξἔνη χἔρον|τες ἡδῆν π(ατ)ρίδα
hōs per xenè cherontes èdèn patrida
DBBE Occurrence 17870 (v.1)

b. Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα
hōsper xenoi xairousin idein patrida
DBBE Type 2820

c. Just like travellers rejoice upon seeing their homeland[2]

In Example 1a the itacism affected the three middle words: ξἔνη *xenè* should be ξένοι *xenoi* (travellers), χἔροντες *cherontes* should be χαίροντες *chairontes* (rejoicing), and ἡδῆν *èdèn* should be written as ἰδεῖν *idein* (to see). The attentive reader may have noted that the third word in Example 1b is not the participle χαίροντες *chairontes* but instead the indicative χαίρουσιν *chairousin*. This is an example of an editorial intervention that goes further than the correction of orthographic mistakes.

In addition to the Occurrences and Types, the database (Demoen et al., 2023) contains metatextual information, e.g. on the manuscript in which the book epigram is found, but also where it was written and by whom.

3.2. Data sets

As we experimented with different partitions of our data, we composed two training sets: the *classical* training set, consisting of 1.06M classical Greek tokens from the AGDT and PROIEL. Secondly, the *mixed* training set consists of the *classical* training set extended with 5K tokens from the DBBE occurrences. We have put together one validation set: the *classical* validation set consists of ca. 80K classical Greek tokens from the AGDT and PROIEL. Finally, the test set is compiled from the DBBE Occurrences: it consists of 10K tokens.

4. Lemmatisation Experiments

To develop a lemmatiser capable of annotating non-edited, Byzantine Greek tokens, we have investigated seven different approaches, which were also compared to existing lemmatisers for Greek.

4.1. Existing Lemmatisers

Three existing lemmatisers, RNN Tagger, the CLTK back off lemmatiser and GLEM, were tested on our gold standard (Swaelens et al., 2023b). The reason

[2] The translations are provided by the authors.

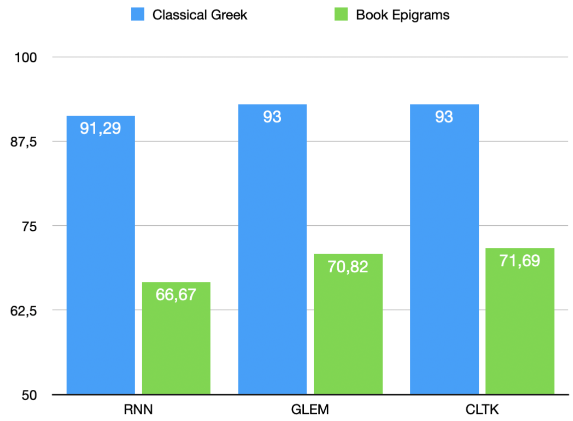

Figure 1: Reported accuracy of existing lemmatisers on classical Greek (blue) and measured accuracy on unedited, Byzantine Greek book epigrams (green).

for doing this is twofold: (1) it provides us with a strong baseline and (2) it clearly proves how different and infinitely more complex the unedited Greek of the book epigrams is, compared to standardised classical Greek. This is illustrated by Figure 1, which shows a drop in accuracy of 20 pp. for the CLTK back off lemmatiser, 22 pp. for the GLEM lemmatiser and 25 pp. for RNN tagger. Although a drop in accuracy was to be expected, the actual decrease in performance was more substantial than anticipated. An in-depth analysis of the errors is provided by Swaelens et al. (2023a), but both the increased use of the perfect tense and the itacism might contribute to the accuracy drop of the existing lemmatisers for Byzantine Greek.

4.2. Experiments with various Lemmatisation Approaches

This section describes the experiments we carried out to develop a new lemmatiser for unedited Byzantine Greek text. The conducted experiments fall into two categories: transformer-based approaches and neural edit-tree approaches. As for the transformer-based approaches, we build on the DBBErt[3] language model (Swaelens et al., 2023b). This model has been trained on a mixture of classical, Byzantine and Modern Greek. Furthermore, it has been fine-tuned to perform part-of-speech tagging, which yielded results competitive to the state-of-the-art models. As for the edit-tree approach, we opt for it because rule-based systems have proven their worth for lemmatisation. This technique, so to speak, learns the rules automatically, which reduces the time of human intervention.

4.2.1. Fine-Tuned Embedding (FT)

The recently-developed transformer-based language model for Byzantine Greek yielded scores competitive to state-of-the-art techniques. Given the competitive results of the fine-tuned DBBErt model, we thought it the perfect start of our experiments to evaluate whether a transformer-based approach might be beneficial to perform lemmatisation. The DBBErt model was first fine-tuned for classification on the *classical training set* and the *validation set*. The model was then evaluated by using the *test set*. It yielded an accuracy score of 56.81%, which is 15 pp. lower than the best baseline lemmatiser, CLTK Tagger.

4.2.2. Change classification head (FT POS/LEMMA)

The DBBErt model has already been fine-tuned to perform part-of-speech tagging, which yields competitive results on the Byzantine book epigrams (Swaelens et al., 2023b). Existing tools already showed that part-of-speech information is beneficial for lemmatisation, which made us fine-tune the DBBErt model that was already fine-tuned on part-of-speech tagging. The *classical training set* was used to fine-tune the embeddings. We added a classification layer on top of the fine-tuned model to perform lemmatisation, which yields an accuracy score of 53.80%.

4.2.3. LSJ & Fine-Tuning (LSJ FT)

Since the fine-tuned embedding did not have a competitive accuracy score, we tried a trivial rule-based approach: if the to-be-lemmatised token matches a headword in the LSJ dictionary, return the headword;[4] if it does not match, let the fine-tuned DBBErt embedding, as described in Section 4.2.1, predict the lemma. With this approach, we hoped to correctly predict uninflected tokens, like adverbs or conjunctions. However, the accuracy score of this approach was even slightly worse than the fine-tuned embedding, yielding an accuracy score of 56.24% on the *test set*.

4.2.4. Training Dictionary & Fine-Tuning (TD FT)

We performed a follow-up experiment building on the dictionary-based approach. This time a Python dictionary was created based on the *classical training set*, with the keys being the tokens and the values a list with the part-of-speech as first element and the lemma as second element. We

[3] The url to the model will be made available for the camera-ready version.

[4] We used the LSJ files made available by Helma Dik and Perseus Tufts.

Token	Lemma
Αἶνος	αἶνος
φλόγα	φλόξ
σβέννυσι	σβέννυμι
τῶν	ὁ
τριῶν	τρεῖς
παίδων	παῖς

Table 2: The data in its original format, as presented by DBBE Occurrence 27019

Token	Lemma
Αἶνος_n	αἶνος
φλόγα_n	φλόξ
σβέννυσι_v	σβέννυμι
τῶν_l	ὁ
τριῶν_m	τρεῖς
παίδων_n	παῖς

Table 3: The data in the token_pos format, presented by DBBE Occurrence 27019

then extended our rule of Section 4.2.3: if the token matches a key in the dictionary *and* its part-of-speech matches the first element of that key's value list, return the lemma; if it does not, let the fine-tuned embedding from Section 4.2.1 predict a lemma. By matching both the token and its part-of-speech, we aimed to reduce errors attributable to the itacism. The preposition εἰς *eis* (to), for example, can be written, among other forms, as *οἰς *ois* or *ἠς *ès*, both of which are possible forms of the relative pronoun. This approach should exclude errors of this nature. It was however not particularly fruitful, as this approach scored an accuracy of 53.17% on the *test set*.

4.2.5. Fine-Tuning with Part-of-Speech (FT POS)

The literature underscores the importance of part-of-speech information when performing lemmatisation. It is, however, not an easy task to incorporate additional linguistic information to fine-tune a transformer embedding. This led us to consider an approach that, at first blush, is less conventional. We have appended the part-of-speech information to the tokens of the *classical training set*. Table 2 shows the original data, Table 3 the data with additional part-of-speech information. Unfortunately, this approach too proved to be ineffective, resulting in an accuracy score of just 47.23% on the *test set*.

4.2.6. Neural Edit Trees (NET)

We conducted two experiments with this state-of-the-art lemmatisation technique. For the first experiment, we trained on the *classical training set*, the *classical validation set* for validation and tested on the *test set*. This yielded an accuracy score of 53%.

During training, an edit has to occur three times to "learn" it. Keeping in mind that the language of our book epigrams is somewhat different from classical Greek, we conducted the experiment again, this time with the *mixed training set* and the *validation set*. By adding 5,000 tokens of the DBBE Occurrences, we hoped that some edits would indeed occur at least three times so that the algorithm could learn and perform better. Nevertheless, the accuracy dropped to 47.85%, a drop of more than 5 pp.

4.2.7. Hybrid Approach (HA)

The study of existing lemmatisers for Greek shows that rule-based systems work better than machine learning techniques, especially when a dictionary is included in the rule-based system (Swaelens et al., 2023a). While analysing the output of our different experiments (see Section 5), we found that closed class words were too often lemmatised wrongly, which made us set up a preliminary hybrid approach. We made a dictionary of all closed-class words, in all possible ways they could occur in Byzantine Greek (cf. the itacism), together with their lemma, and used this dictionary to lemmatise the closed-class words in our test set. The other tokens were predicted by the fine-tuned DBBErt model in the first preliminary experiment, for the second experiment by the neural edit-trees. The closed-class dictionary combined with the fine-tuned embedding resulted in 65.76% accuracy, combined with the neural edit-trees in 62.11%. This brings us to the conclusion that the edit-trees already did a better job in lemmatising the closed-class words.

4.3. Summary of Experimental Results

Figure 2 presents the results of the approaches that we investigated for the lemmatisation of Byzantine Greek unedited text, which are summarised in Table 4. We provide two reference lines to compare our results to: the best performing state-of-the-art Back-Off lemmatiser (71.69%) and Stanza/CLTK lemmatiser (64.99%). Additionally, it is important to mention that the accuracy score on the (classical Greek) validation set of one of the edit-tree experiments was 95%, while the result on the book epigrams was 53%. The same drop in performance was observed for training the classifier on the DBBErt embeddings, where the validation set yielded accuracy scores higher than 90%, while the model only reached 56.81% on the Byzantine test set.

By means of an extensive error analysis, we hope to provide some insight in the results of the experiments.

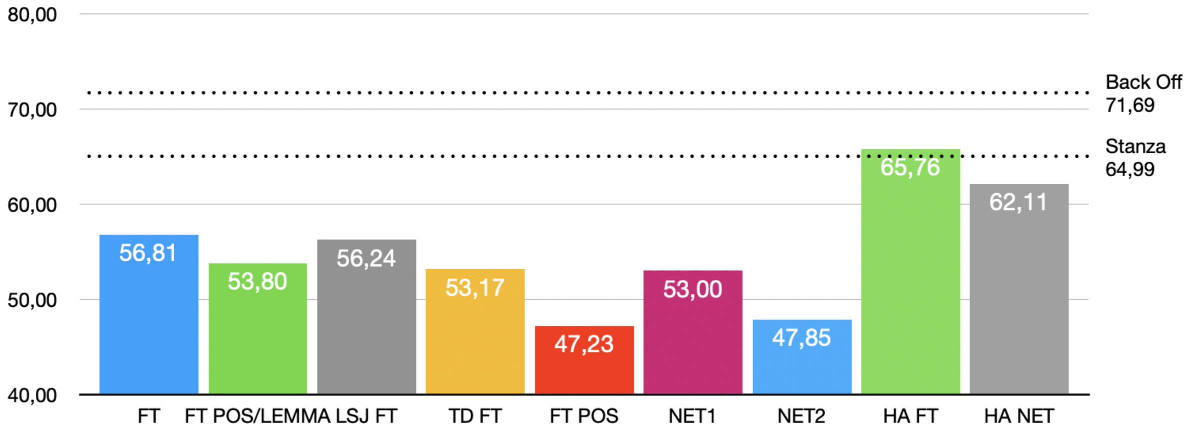

Figure 2: Graph presenting the accuracy scores of the lemmatisation approaches we tested compared to the worst and best baseline.

Lemmatiser	Accuracy
FT	56.81
LSJ FT	56.24
TD FT	53.17
FT POS	47.23
NET1	53.00
NET2	47.85
HA FT	**65.76**
HA NET	62.11

Table 4: Summarising table of the results of all tested lemmatisation approaches.

5. Error Analysis

5.1. Iota Subscriptum

The most prominent problem of all tested lemmatisers, is the absence of the iota subscriptum. In Greek, a iota (ι) following an α, η or ω, is written either underneath that vowel, *subscriptum*, or next to it, *adscriptum*. This is, again, an editorial invention to make the Greek easier to read. These iota's however are very often absent in original texts and the book epigrams are no exception to that.

(2) a. + αἶνο(ς) ϑ(ε)ῶ χάρις τε (καὶ) δόξα, πρέπει· τῶ δόντι τέρμα τῆς γραφῆς φϑᾶσαι σϑένος
ainos theō charis te kai doxa prepei:
tō donti terma tès graphès phthasai sthenos
DBBE Occurrence 17386

b. Praise is due to the Lord, as well as gratitude and glory:
to Him who gives the strength to reach the end of this writing.

Example 2a contains two words in the dative, ϑεῶ *theo* (God) and τω *to* (the). Both of them are usually written with a iota subscriptum (ϑεῷ and τῷ), which is also how they are to be found in the training data. Admittedly, this phenomenon did not occur in the lemmatisation training sets. However, the DBBE was part of the data on which the language model itself was trained. We therefore did not expect the absence of this iota to be this challenging. Preprocessing the data and, if the token has a nominal part-of-speech, adding the iota to it might alleviate this problem.

5.2. Verbal System

The Greek verbal system is characterised by a variety of stems per verb. Every stem represents a combination of a diathesis with an aspect (continuous, punctual, future and perfective). The lemma of a verb is generally the active indicative present, first person singular. To form the aorist (or punctual) stem, the regular verbs, on the one hand, have a quite transparent paradigm. The stem of the present is extended with a sigma. The aspect stems of irregular verbs, on the other hand, can look quite different compared to their present stem, due to, e.g. ablaut or reduplication.

Example 2a contains an aorist participle, δόντι *donti*, from the lemma δίδωμι *didomi* (to give). This aorist participle is built on the short o-grade stem δο- *do*, while the present indicative (lemma form) is built on the reduplicated, long o-grade stem διδω- *dido-*. The fine-tuned embeddings wrongly lemmatised this word as ὀδούς *odous* (tooth), while the edit-trees simply returned the token itself.

Compared to the aorist stem, the perfective stem is more complex. The regular formation of the perfective stem entails adding the reduplication in the e-grade as a prefix to the stem and a kappa as a suffix to the stem. Again, the stem of the irregular verbs can change quite radically, which makes it hard to recognise the lemma. Furthermore, the perfective stem is more common in Byzantine Greek than it is in classical Greek. In our classical training set, for example, 6% of all verbal forms have a perfective stem while in the test set, containing only Byzantine Greek, 11.4% of all verbal forms have a perfective stem. This might contribute to the explanation why so many perfective verbal forms are lemmatised incorrectly.

(3) a. αὔτη βίβλος φέρουσα συντεταγμένους
θαῦμα πρόκειται πᾶσιν ἐξηρημένον
autè biblos ferousa suntetagmenous
thauma prokeitai pasin exèrèmenon
DBBE Occurrence 17013 (vv. 3-4)

b. This book, carrying composed <words>, is about a miracle exceptional to all.

Two perfective forms are to be found in Example 3, both of which are predicted incorrectly by the fine-tuned embeddings as well as the neural edit-trees. The participle συντεταγμένους *suntetagmenous* (composed) consists of the stem ταγ- *tag-*, which was reduplicated by adding τε- *te-*. This perfective stem from the verb τάσσω is preceded by the prepositional prefix συν- *sun*, which brings us to the lemma συντάσσω *suntasso* (compose). A dictionary-based approach would have performed better, provided that the dictionary contains all aspect stems.

The second perfective form in Example 3, ἐξηρημένον *exèrèmenon* (exceptional), is not characterised by reduplication at all. Being a perfective participle from the lemma ἐξαιρέω *exaireō* (take out of), the formation consists of the prepositional prefix ἐξ- *ex-*, the augment ε- *e-* that contracted with the stem αιρη- *airè-* to ηρη- *èirè-* and the regular suffixes to form a participle, -μενον *-menon*. The absence of the iota subscriptum in our example, makes the form even less recognisable as being part of the paradigm of ἐξαιρέω *exaireō*.

The relatively small number of perfectives in the training data might be the reason why the fine-tuned embeddings have such a rough time classifying these forms correctly. As for the edit-trees, this might be an even harder task because of the peculiarities of the DBBE corpus, e.g. the absence of the iota subscriptum, which makes these perfective forms even more dissimilar to their lemma than in classical Greek.

5.3. Nominal Stems

Another problem our lemmatisers faced, are nouns of which the stem is not (completely) visible in their lemma. Nouns that do have the complete stem included in their lemma perform remarkably better. The first word of Example 2a has the stem αἰν- *ain-*, which is part of the lemma αἶνος *ainos* (praise).

The stem of πατρίδα *patrida* (homeland) in Example 1a, however, is πατριδ- *patrid-*. The final consonant of this vowel is not visible anymore in its lemma (πατρίς *patris*) due to assimilation of δς *ds* to ς *s*. The first example is lemmatised correctly, while the second example is not.

(4) a. δί|δου μοι λύσ(ιν) πολλ(ῶν)
ἀμ|πλακημ(ά)τ(ων):
didou moi lusin pollōn amplakèmatōn
DBBE Occurrence 17060 (v. 4)

b. Give me remission for my many faults.

The last word of Example 4 displays another nominal category that is lemmatised wrongly. It consists of the stem ἀμπλακηματ- *amplakèmat-* and the suffix of the genitive plural -ων *-ōn*. The lemma of ἀμπλακημάτων *amplakèmatōn*, however, is ἀμπλάκημα *amplakèma* (fault). The consonant τ *t* at the end of the stem is omitted since Greek words can only end with either a vowel or the consonants ν *n*, ρ *r* and ς *s*. We cannot quite pinpoint why these nominal forms are being lemmatised incorrectly, given that (1) the stem and the lemma by and large consist of the same letters, differing only in the final (cluster of) consonant(s) of the stem, and (2) this kind of nouns appears frequently in both the classical and Byzantine corpus.

5.4. Evaluation Problems

The problems addressed in the previous sections are wrong predictions made by the lemmatisation algorithms. This section describes issues that might not be considered an “error” as such. As pointed out in Section 3.1, Greek was not one, uniform language.

(5) a. τοῖς καθαροῖσι νόον, μυστήρια λαμπρὰ φαείνει·
tois katharoisi noon, mustèria lampra phaeinei
DBBE Occurrence 17014 (v. 2)

b. It reveals the brilliant mysteries to those that are pure of mind.

The lemma of the word νόον *noos* is νόος *noos* (mind), but the Attic dialect uses the contracted form νοῦς as lemma. The edit tree correctly predicted the lemma νόος *noos*. This was then unjustly evaluated as being incorrect, because of our gold standard using the Attic form of the lemma νοῦς *nous*. The same goes for the word βίβλος *biblos* in Example 3. Both the fine-tuned DBBErt model and the edit-trees correctly predicted the Attic form βίβλος *biblos* (book) as lemma, yet our gold standard contains the more general form βύβλος *bublos*. The word συντεταγμένους *suntetagmenous* (composed) in Example 3, already described within the scope of the verbal system, faces the same issue. The lemmatisers predicted the lemma συντάττω *suntatto*, which is correct. In our gold standard, however, the Attic pendant συντάσσω *suntasso* is used. These two forms are, in fact, exactly the same but evaluated as being different and thus incorrect. These are alternating forms, so the predictions of the lemmatisers are unjustly evaluated as being incorrect.

A related issue pertains the irregular adjectives, which share some characteristics addressed previously in Section 5.3.

(6) a. ἐκ τῆς ἀρίστης ἐκλογῆς τε καὶ τέχνης
ek tès aristès eklogès te kai technès
DBBE Occurrence 17013 (v.6)

b. by the best selection and (the) art

The word ἀρίστης *aristès*, for example, is the superlative form of the adjective ἀγαθός *agathos* (good). The gold standard returns the positive grade of every adjective as lemma, while the predictions of both the fine-tuned embeddings and the edit-trees are the superlative itself ἄριστος *aristos* (best). Although not matching our gold standard, it is in fact a headword in the reference dictionary (George and Liddell, 1968), so relaxation of the evaluation might be inevitable.

6. Conclusion & Future Research

This paper presented preliminary experiments to lemmatise unedited Byzantine Greek, exploring the application of either a transformer-based classification system or a more recent lemmatisation technique, viz. neural edit-trees. Neither approach managed to yield competitive results when compared to the best performing existing lemmatiser for Greek. Is it too bold a claim that Byzantine Greek challenges the capabilities of transformers? After all, Greek – particularly its Byzantine variant – is a low-resourced language, with a very complex morphological system that is characterised by, among other things, stem changes and reduplications, which are, in their turn, subject to phonetic laws, and, in case of the book epigrams, orthographic inconsistencies. Linguistic knowledge turned out to be indispensable for tackling this complex task, which led us to explore a hybrid approach. This hybrid approach, combining a more traditional rule-based approach with a machine learning component, yielded our best accuracy score, namely 65.76%.

Despite the initial setbacks, we will continue experimenting with transformer-based approaches. We intend to train a sequence-to-sequence model for Greek and assess its effectiveness in tackling our lemmatisation challenge. Additionally, we will delve deeper into the neural edit-tree approach, fine-tuning its parameters and experimenting with various data partitions. As a last, but maybe most effective way to elaborate on our lemmatisation experiments, we will investigate the impact of adding more linguistic data. The incorporation of the closed-class dictionary notably increased our best lemmatiser with nearly 10 pp., prompting us to consider creating a more extensive dictionary that includes, among other elements, verbs with all their associated stems. Furthermore, we will continue annotating Byzantine Greek texts in order to accumulate sufficient Byzantine Greek data to be included in the training set.

7. Bibliographical References

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.

Gregory 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.

Daniel De Kok. 2021. Neural edit-tree lemmatization for spacy. Last accessed October 7th, 2023.

Henry George and Scott Liddell. 1968. *A Greek-English Lexicon*. Clarendon.

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.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA model for Latin token tagging tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, 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.

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 and Digital Humanities*, On the Way to the Future of Digital Manuscript Studies.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Aleksi Sahala, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020. Automated phonological transcription of Akkadian cuneiform text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3528–3534, Marseille, France. European Language Resources Association.

Helmut Schmid. 1991. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

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.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, pages 1–45.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023a. Evaluating existing lemmatisers on unedited byzantine Greek poetry. In *Proceedings of the Ancient Language Processing Workshop*, pages 111–116, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023b. Linguistic annotation of byzantine book epigrams. *Language Resources and Evaluation*, pages 1–26.

Colin Swaelens, Ilse De Vos, and Els Lefever. 2023c. 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.

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.

8. Language Resource References

Jan-Mathieu Carbon, Saskia Peels-Matthey, and Vinciane Pirene-Delforge. 2017. Collection of greek ritual norms (cgrn). Last accessed October 7th, 2023.

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

Kristoffel Demoen, Gilbert Bentein, Klaas Bentein, Floris Bernard, Julián Bértola, Julie Boeten, Mathijs Clement, Cristina Cocola, Eline Daveloose, Sien De Groot, Pieterjan De Potter, Ilse De Vos, Krystina Kubina, Hanne Lauwers, Paulien Lemay, Renaat Meesters, Marjolein Morbé, Delphine Nachtergaele, Marthe Nemegeer, Joachim Nielandt, Mace Ojala, Lisa-Lou Péchillon, Raf Praet, Rachele Ricceri, Anne-Sophie Rouckhout, Jeroen Schepens, Febe Schollaert, Lev Shadrin, Nina Sietis, Dimitrios Skrekas, Colin Swaelens, Maria Tomadaki, Sarah-Helena Van den Brande, Merel Van Nieuwerburgh, Lotte Van Olmen, Noor Vanhoe, and Nina Vanhoutte. 2023. Database of byzantine book epigrams.

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.

Irene Vagionakis. 2021. Cretan institutional inscriptions dataset. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.