Linguistic Annotation of Byzantine Book Epigrams

Colin Swaelens[1*[0000-0002-3360-8093]], Ilse De Vos[2†[0000-0002-1152-3072]] and Els Lefever[1†[0000-0002-7755-0591]]

[1] LT3, Ghent University, Groot-Brittanniëlaan 45, Ghent, 9000, Belgium
[2] Department of Linguistics, Ghent University, Blandijnberg 2, Ghent, 9000, Belgium

*Corresponding author: Colin.swaelens@ugent.be;
Contributing authors: i.devos@ugent.be, els.lefever@ugent.be;
†These authors contributed equally to this work.

Abstract. In this paper, we explore the feasibility of developing a part-of-speech tagger for not-normalised, Byzantine Greek epigrams. Hence, we compared three different transformer-based models with embedding representations, which are then fine-tuned on a fine-grained part-of-speech tagging task. To train the language models, we compiled two data sets: the first consisting of Ancient and Byzantine Greek texts, the second of Ancient, Byzantine and Modern Greek. This allowed us to ascertain whether Modern Greek contributes to the modelling of Byzantine Greek. For the supervised task of part-of-speech tagging, we collected a training set of existing, annotated (Ancient) Greek texts. For evaluation, a gold standard containing 10,000 tokens of unedited Byzantine Greek poems was manually annotated and validated through an inter-annotator agreement study. The experimental results look very promising, with the BERT model trained on all Greek data achieving the best performance for fine-grained part-of-speech tagging.

Keywords: Byzantine Greek, part-of-speech tagging, morphological analysis, computational linguistics, natural language processing, machine learning, neural networks, language models.

1 Introduction

Byzantine book epigrams are an invaluable resource for the fields of linguistics, sociocultural history, textual transmission, and literary studies. The term book epigram designates poems that are both written in and on books [1] and was introduced into Byzantine scholarship by Lauxtermann [2], who remarked that those poems are intimately related to the production of literary texts and manuscripts. The epigrams tell the reader more about the manuscript they are part of: they present its topic, tell us who authored it, physically wrote it, paid for it, read it, and so on. That makes book epigrams not fundamentally different from epigrams inscribed on other objects (e.g., tombstones, buildings) as they are completely dependent on and only meaningful if considered within their material, physical context. Set apart from the main text of the manuscript,

the book epigrams help the reader in their orientation as to what this main text means, and how the carrier of the text, i.e., the manuscript, came into being.

The unique character of those book epigrams originates in the fact that they are interconnected in many ways.

Firstly, the epigrams are still to be found in their original, unedited form and original context of use, which is, on the contrary, not the case for the available texts from antiquity. For example, Homer's Iliad, first written down around 800 BC, has come down to us through centuries of copying, editing, and recopying by scribes and scholars. Through the centuries, manuscripts have been damaged or destroyed and no original autograph has survived. The Iliad as we can read it today in the edition of e.g., Martin West [3] is West's reconstruction of Homer's text, based on over 800 sources, among which manuscripts, papyri fragments, comments of grammarians, and authors that have cited Homer. Consequently, every readable text from antiquity is a conflation of multiple sources and is edited by a philologist to reconstruct the original text as good as possible. The book epigrams, however, are to be found in their original form on their original carrier without any editorial interference, which makes them unique within the written tradition of Greek.

Secondly, since many book epigrams are composed by scribes, i.e., people who are often non-professional poets, the language of the corpus displays less erudite literary techniques and linguistic developments [4].

Thirdly, book epigrams are often formulaic: verses, half-verses, poems, and chunks of poems recur throughout the corpus, often object of adaption, permutation, and variation.

By now it is clear that those book epigrams give us a glimpse of the contemporary world and language of the scribes, who, in their turn, enriched our modern culture with the inheritance of Classical Antiquity. These invaluable epigrams should therefore be made available for researchers within the fields of Medieval studies and linguistics. The Database of Byzantine Book Epigrams (DBBE) [5], the corpus we are working with, is online available and free to be used for research purposes. Although the texts stored in the DBBE are provided with all sorts of metadata (date, author, place of writing, etc.), the DBBE does not contain any linguistic information on them. Since the existing tools for Greek are not flexible enough for the very challenging nature of the DBBE data (which will be discussed in Section 2.1), there was a need to develop a new linguistic annotation pipeline for both Medieval[1] and Ancient Greek.

This paper reports on the experiments carried out for the development of a morphological analyser for Byzantine Greek. In a first step, various transformer-based language models were trained and compared, while, in a second step, the resulting language models were fine-tuned for the task of fine-grained part-of-speech tagging (or morphological analysis). Since Greek is a highly inflectional language, our part-of-speech tags consist of both the *part-of-speech* and the complete morphological analysis of each word.

---

[1] For this paper, Byzantine and Medieval will be used as synonyms that designate the period from the 5th and 15th century AD.

The remainder of this paper is organised as follows: Section 2 provides an extensive overview of both existing resources and natural language processing techniques for Greek. Next, Section 3 elaborates the process of data collection and annotation, including an inter-annotator agreement study. Once the data is defined, the development of the part-of-speech tagger is described: Section 4.1 dives into the training of the language model, while Section 0 treats the fine-tuning on a part-of-speech tagging task. The experimental results are described and analysed in Section 5, after which a conclusion is drawn in Section 6.

2 Related Research

2.1 Resources

It is very hard to compile a corpus for Medieval Greek since the available resources mainly provide classical or biblical texts. The Thesaurus Linguae Graecae (TLG) [6], being the pioneer of digitising Greek literary texts since 1971, has the biggest digitised collection of Greek texts, spanning from Homer (800 BC) until the fall of Byzantium (1453 AD). Their work, consisting of more than 110 million tokens, coming from more than 10,000 works associated with over 4,000 authors, can be consulted online but is unfortunately not available to be used for research purposes. Nevertheless, we collected as much data as possible, to train both a state-of-the-art language model and a part-of-speech tagger.

A first open-source alternative to the TLG is the Perseus Digital Library (or Perseus Project) [7]. This is a growing digital library of resources for the study of the humanities, focusing on the Greco-Latin world. The project and consequently the corpus started with very few texts and expanded as funding was secured. The Perseus Digital Library currently contains about 13.5 million tokens of Greek literary texts, both prose and poetry, and spans the period from the 8th century BC until the 12th century AD. Almost every text on Perseus is open source and can be downloaded as an XML-file.

The First1KGreek project[2] serves as a complement to the Perseus Digital Library. The project aims to collect every Greek work composed between Homer and 250 AD with an explicit focus on texts that are not to be found at other open-source platforms. It contains over 25.5 million tokens of both prose and poetry. Both the Perseus Digital Library and the First1KGreek project are part of the Open Greek and Latin portal[3], an open-source platform with digital texts, reading tools and software.

In addition to the available plain Greek texts, several initiatives arose that (manually) tagged every word in a text with its part-of-speech, its lemma, its syntactic relation, and (sometimes) even its semantic role. Such annotated corpora are called treebanks. The list of treebanks we describe in this article is not exhaustive but limited to those used in

[2] https://opengreekandlatin.github.io/First1KGreek/
[3] https://opengreekandlatin.org

our research. The Ancient Greek Dependency Treebank (AGDT) [8, 9] was the first treebank to cover an Ancient Greek corpus. The AGDT contains 560,000 tokens from classical prose as well as poetry, which are manually provided with a part-of-speech tag, a morphological analysis, their lemma, and syntactic relations. PROIEL [10] is a treebank that contains the New Testament in five different languages, including the Greek original, but also Herodotus's *Histories* and Sphrantzes's *Chronicles*. The PROIEL project contains some 277,000 tokens for Greek. The Gorman treebank [11] is a collection of exclusively Ancient Greek prose authors and totals some 550,000 annotated tokens. Another open-source resource containing annotated Greek texts is Trismegistos [12], an interdisciplinary portal of papyrological and epigraphical resources. In addition to the texts themselves, Trismegistos provides metadata about those texts facilitating cross-cultural and cross-lingual research. The Trismegistos corpus generally covers papyri (fragments) from the 4th century BC until the 7th century AD, which sums up to 945,776 stored entries or 4,817,824 tokens. Finally, there are the Pedalion Trees [13], containing both prose and poetry from the Perseus project that is missing from the AGDT. Pedalion contains some 320,000 annotated tokens.

For the presented research, we have integrated the data of the Database of Byzantine Book Epigrams, an ongoing project that stores both textual and contextual data of book epigrams from Medieval Greek manuscripts dating from the 5th up to the 15th century.

The intentions of the DBBE are threefold: 1) it aims to provide both reliable transcriptions and readable texts of the epigrams, 2) which are brought together in an interrelated, structured, and searchable corpus, and 3) are provided with references to existing sources and material. The DBBE currently contains over 12,000 unique book epigrams. They are stored both as *occurrences*, the epigrams exactly as they occur in their manuscripts, and as *types*, normalised and thus more readable versions of those same texts. *Occurrences* on the one hand render the text of individual epigrams as faithfully as possible, displaying all idiosyncrasies of the manuscript, e.g., in terms of orthography. *Types* on the other hand serve as umbrellas to which one or more occurrences are linked. Together, the *types* and *occurrences* count 412,529 tokens. When a new book epigram is discovered in a manuscript and added to the DBBE, it is primarily published as it is found in that manuscript and thus as an *occurrence*, after which the DBBE either links it to an existing *type* or, if no similar *type* exists, we create a new one. Since the creation of new *types* needs some time, our aim is to immediately annotate new *occurrences* when they are added to the DBBE. Furthermore, the growing interest in optical character recognition for manuscripts [14] means that more *occurrence*-like texts will become digitally available. Consequently, the development of a linguistic annotation pipeline, capable of processing this kind of very challenging non-normalised texts (as will be further explained in the next paragraph), will be very relevant for other linguistic research as well.

The Greek that is found in the *occurrences* contains a lot of orthographic inconsistencies, that are mainly due to phonetic shifts. These phonetic laws are comprehensively explained by Holton et al. [15]. Given the scope of this paper, we will mainly focus on

the itacism or iotacism[4]. The classical, Athenian pronunciation of the vowels ι [i], η [ɛ], υ [y] and diphthongs ει [eː], οι [oj] shifted to the pronunciation [i][5]. Type 2150, shown in Example 1, renders the normalised Greek text of the three occurrences shown in Examples 2, 3 and 4. Although all three occurrences provide the exact same verse, the itacism confused the scribes, which resulted in an abundance of orthographic variants of that same verse.

1. Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα
   Osper kseni cherousin iðin patriða
   Just like travelers rejoice by seeing their homeland[6]
   DBBE Type 2150

2. + ὥσπερ ξένη χᾶίρου σύν εἰδεῖν πατριδα ·
   + osper xeni cherou sin iðin patriða ·
   DBBE Occurrence 17591

3. Ὥσπερ ξένοι χαίρουσιν ἡδεῖν πατρίδαν ·
   osper xeni cherousin iðin patriðan ·
   DBBE Occurrence 18619

4. + οσπέρ ξένη χεροῦσι ηδὴν πατριδα
   + osper xeni cherousi iðin patriða
   DBBE Occurrence 18746

An example of this orthographic inconsistency is the penultimate word of every verse (*to see*), that is written in its *normalised* Ancient Greek form in Example 1. All three occurrences of that same word are written in a different way in Examples 2-4. Furthermore, these last three examples display the tendency of leaving out the spiritus asper, ἁ [ha], and the spiritus lenis, ἀ [a], which - at the beginning of a word - indicate the presence or absence of the phoneme /h/ respectively. Example 2 renders a spiritus lenis on **εἰ**δεῖν, while Example 3 has a spiritus asper on **ἡ**δεῖν and Example 4 has no spiritus at all on **η**δὴν. A last evolution shown in Examples 2-4, is the disappearance of

[4] The term *itacism*, on the one hand, originates from the new pronunciation of the letter η, that shifted from [ɛta] to [ita]. *Iotacism*, on the other hand, originally refers to the shift in pronunciation of all vowels and diphthongs to [i]. Nevertheless, both terms tend to be used interchangeably.

[5] That shift took place before the 3rd century AD, for which we decided to provide every example of Greek text with its medieval, phonetic transcription (IPA) instead of a classical transliteration.

[6] We used the translations provided at the *type* page of every *occurrence* in the DBBE, except for those translations marked with ‡. These are the author's translations.

the distinction between long and short vowels, e.g. Example 4 has written οσπέρ (*just like*) with an omicron [o], the other two with an omega [o:].

All these phonetic changes had repercussions in the typeface. The scribes definitely knew Greek, but that was clearly not sufficient to distinguish which [i] sound should be written in their own poems. Because of these inconsistencies, a lot of ambiguous forms arise throughout the DBBE corpus, e.g., οσπέρ in Example 4, that denotes the adverb ὥσπερ but might be analysed as an indeclinable form of the pronoun ὅς (his, her). Existing tools to linguistically annotate Greek texts are developed to process standardised, Ancient Greek texts and are thus not capable of dealing with the peculiarities described above. This is why we developed a novel, more flexible machine-learning approach to perform automatic linguistic annotation of Byzantine Greek poems.

2.2 NLP on the Greek Language

Although there has been an increase in attention to NLP research for Ancient Greek over the last few years, interest in the topic dates back at least half a century. Interestingly, the research and development of the first NLP tool for Greek has its origin in an educational context.

David W. Packard wanted to reform the curriculum for teaching ancient Greek at university. He believed that students would be able to read literature much earlier in their curriculum if the initial grammatical instruction and the language phenomena focused on were in alignment with the actual texts they first read. To test that theory, he needed the complete lexical and grammatical analysis of the words that occur in texts suited for first-year students, which resulted in the first tool to perform morphological analysis on 40,000 ancient Greek tokens [16]. Packard's *pipeline* is straightforward. Before the analysis is performed, every input word is first compared to the so-called *indeclinable list*. This is a list of some 800 words that are either not inflected or have a highly irregular inflection. Apparently, half of the tokens in a typical Greek text occur in that list. The other half of the words are then analysed according to Greek morphology rules, described in Smyth's grammar. The algorithm splits the ending from the stem by removing the final character of a word and determining whether that character is an inflectional ending. If that is the case, the remaining part of the word is assumed to be a stem, which is then looked up in a dictionary. Should the stem exist and be consistent with the inflectional ending, one possible analysis is saved, and the algorithm repeats itself until no more possible analyses are found. Packard expressly states three difficulties for Greek: firstly, crasis – the merging of two words into one – is very difficult to detect automatically, which is why the most frequent occurrences of crasis were added to the *indeclinable list*. Secondly, the system could not deal with ambiguous forms, i.e., word forms with multiple possible analyses. The programme printed the most likely analysis together with a warning for the editor to review that form. Thirdly, the programme does not take into account accents nor diacritics, thus creating even more

words that are difficult to tackle. Despite these drawbacks, Packard's system was the base for further development in NLP for Greek.

It laid the foundation, for instance, for Morpheus [17], a morphological analysis tool that is still widely used today. Morpheus looks for possible endings and a stem to which those endings might be attached. These forms are then looked for in a large database of possible Greek forms. What is new, is that Morpheus can take into account diacritics: an improvement that reduces all possible matches with a whopping 23%. The main advantage however of Morpheus compared to Packard's system, is that Morpheus has been gradually developed to be able to deal with more than only Attic, the most-studied dialect of classical Greece. That was a big leap forwards, since standardised Greek was non-existent until the rise of Koinè in the 3rd century BC [18]. According to Miller [19], there were four major dialectical clusters that could look very dissimilar, e.g., the genitive singular of the word 'heaven' was οὐρανοῦ in Attic, but ὠράνω in Aeloic (dialect of Aeolia, Boeotia, and Thessaly). Morpheus is able to recognise both morphological endings and stems from the Greek dialects, which made this a very useful and powerful system to perform morphological analysis. It has, for example, been integrated in the AGDT to speed up manual annotation by displaying all possible analyses of a given word.

Fully automatic morphological analysis, however, is not possible, since Morpheus does not decide which analysis is most likely to be correct in case several analyses are possible. To cope with this disambiguation problem, several part-of-speech taggers have been developed. Celano et al. [20] tested five existing part-of-speech taggers on literary data from the AGDT. This comparative study shows that the Mate tagger [21] outperforms the Hunpos tagger [22], the RFTagger [23], the OpenNLP POS tagger[7], and the NLTK Unigram tagger [24] on classical Greek data. This comparative study was repeated by Keersmaekers [25], who compared Mate tagger, RFTagger, and MarMot tagger [26], as part of the development of a linguistic annotation pipeline for papyrological Greek. Unlike literary Greek, papyrological Greek is to be found in its original, unedited form, a quality it shares with the book epigrams we work with. The aim of the pipeline of Keersmaekers is to enable morphological analysis and syntactic annotation for papyrus texts[8]. That papyrus corpus is analysed best by RFTagger (94.7%), who outperforms the Mate and MarMot tagger for that task. However, every token of the papyrus corpus is provided with a normalised version of that token, on which the morphological analysis was performed.

Johnson et al. [27] founded the Classical Language Toolkit (CLTK), an open-source Python framework dedicated to NLP support for historical languages. The CLTK made available a collection of corpora, which for Greek include the Perseus Digital Library (cf. Section 2.1), the First1Kgreek (cf. Section 2.1) and Lacus Curtius[9]. Furthermore,

---

7 https://opennlp.apache.org

8 Given the scope of this paper, we do not examine NLP approaches for Greek other than tokenisation and part-of-speech tagging.

9 http://penelope.uchicago.edu/Thayer/E/Roman/home.html

several language models and resources were developed for, among other things, probabilistic sentence and word tokenisation, part-of-speech tagging, lemmatisation, and morphological tagging. That all-in-one framework allows philologists with less technical knowledge to make use of existing language technology for Greek.

The Diorisis project [28] in its turn, performed automatic linguistic pre-processing on 820 freely available Greek texts, for the purpose of developing a computational model for semantic change in Ancient Greek. The part-of-speech tagger they used, was the stochastic TreeTagger [29]. Training was done with data from both the AGDT and PROIEL, testing on the AGDT, which resulted in an 91% accuracy score. One particular focus within the Diorisis project was the disambiguation of words coming from different lemmas. That is done by first assigning the part-of-speech tag and only then assigning the most-likely lemma, based on this part-of-speech tag.

Schmid, in his turn, developed a neural-based part-of-speech tagger, RNN Tagger [30], that has been trained for Ancient Greek as well. This resulted in a state-of-the-art accuracy score of 91.29% when tested on the AGDT.

The first transformer-based language model for Ancient Greek was developed by Brennan Nicholson[10]. The model is character-based and is trained to predict missing characters in Greek words.

An exploratory study on automatic linguistic annotation of Byzantine Greek epigrams from the DBBE was carried out by Singh et al. [31]. What makes this study stand out, is that Singh et al. were the first to train a transformer-based language model for Ancient and Medieval Greek and implement it in a part-of-speech tagger for Greek. They retrained a pre-trained Modern Greek language model [32] on Ancient Greek data, be it without diacritics[11]. This language model was then implemented in a part-of-speech tagger, for which the FLAIR [33] architecture was used. The part-of-speech tagger yielded a competitive accuracy score of 86.88%. However, the language model was trained on both Ancient and Modern Greek, while the part-of-speech tagger was trained on Ancient Greek data and evaluated on Medieval Greek data. Furthermore, the evaluation set was compiled of *types* from the DBBE, which, as mentioned above, deviate from the actual Greek we find in the *occurrences*.

3 Data Collection & Annotation

3.1 Data Collection

To develop a part-of-speech tagger that makes use of a transformer-based language model, several data sets were needed. Firstly, we needed as much plain Greek text as possible to serve as input for the language model. Next, both a manually annotated train and test set were necessary for the development of the part-of-speech tagger.

---

[10] https://github.com/brennannicholson/ancient-greek-char-bert

[11] Modern Greek only preserved the acute accent. The circumflex and grave accent lost their discriminative function and disappeared, just like the spiritus asper and the spiritus lenis.

The data collected for the language model counts 127.413.536 tokens consisting of the above-described resources: the complete DBBE (both *types* and *occurrences*), First1Kgreek project, Perseus Digital Library, Trismegistos' papyrus texts, the Modern Greek Wikipedia and Byzantine book epigrams from an edition by Rhoby [34]. All data files have been pre-processed as follows: first, for every XML-file in the corpus, we extracted all text between the <text> and </text> tags, as we did not need any metadata. Next, if the texts were written in bètacode, they were converted into the Greek alphabet. As a last step, all editorial signs (e.g. < > or [ ]) were deleted, except for (...), which signals that text is missing. We kept these indications of lacunae because we believe they contribute to the sentence's structure; if they are deleted, the sentence structure becomes nonsensical. This means that we did not change anything about the texts as found in the online corpora, except for the removal of the editorial signs. All cleaned files were then classified into three time periods, for which we relied on the *Canon of Greek Authors and Works* [35] by TLG.

Period	Number of tokens
Pre-Byzantine	31,467,014
Byzantine	7,952,710
Post-Byzantine	85,575,140
Incerta	52,486
Varia	2,366,183
Total	127,413,536

Table 1. The compilation of the data used for the language models.

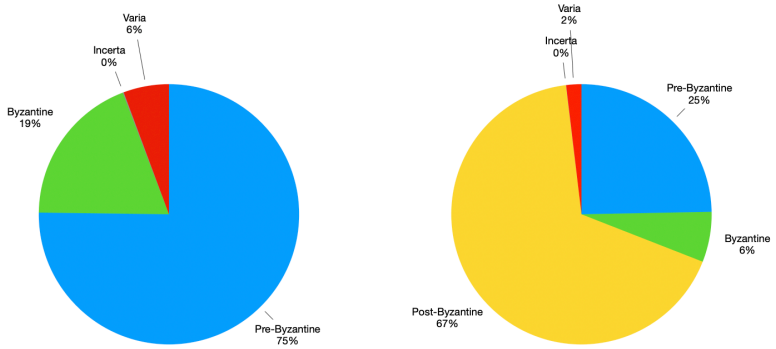

Fig. 1. The left chart shows the distribution of the data set containing pre-Byzantine, Byzantine, incerta and varia, the right one that same data set complemented with post-Byzantine data.

Table 1 shows the partition of our tokens per time period, also displayed in **Error! Reference source not found.**. For our corpus, the pre-Byzantine or Ancient period spans from the 8th century BC up to and including the 5th century AD and consists mainly of First1K, Perseus and Trismegistos. The Byzantine or Medieval period covers the 6th up to and including the 15th century and is mainly composed of DBBE data and Rhoby's epigrams. Finally, the post-Byzantine or Modern period covers the 16th century until today and only contains the Modern Greek Wikipedia data. Works for which no indication about the time of writing exists, e.g., *Vitae Hesiodi Particula*, are marked with *incerta* in the TLG, a strategy we also applied for our data distribution. Composed works, like the *Anthologia Graeca*, are works that contain texts from several centuries, and are therefore marked with *varia*. The texts indicated with *incerta* or *varia* have been added to the pre-Byzantine corpus, because we wanted to guarantee that every text in the Byzantine corpus really is Byzantine.

Next, we have compiled two data sets for the training and evaluation of the part-of-speech tagger. To compile the training set, we extracted all words and their corresponding part-of-speech tags from the AGDT, the Greek texts in PROIEL, the Gorman treebanks and the Pedalion trees. The syntactic and/or semantic tags have been deleted, as they are not within the scope of this paper. Since we aim to annotate unedited Byzantine texts, the test set consists of 10,000 tokens from the DBBE *occurrences* that have been manually annotated. We are aware of the big difference in data between the training set, containing only Ancient Greek, and the test set, containing only Byzantine Greek. To overcome this discrepancy, we are still annotating DBBE *occurrences* to add annotated Byzantine Greek to the training set.

3.2 Inter-Annotator Agreement Study

The aim of the inter-annotator agreement (IAA) experiment is to evaluate whether (1) the label set shown in Table 3 is suitable for this corpus of Byzantine book epigrams, and (2) the manual annotations are reliable and consistent across annotators, which is a prerequisite to use the resulting corpus for training and evaluating our pre-processing pipeline. Since we want the DBBE, when eventually annotated, to be complementary to the AGDT, we based our label set on the one used in the AGDT [36]. However, the label *missing* (indication of lacunae) needed to be added. As the AGDT solely contains edited data, all sentences are corrected to perfection. This is not the case for the *occurrences* in the DBBE, which display all the text's idiosyncrasies regarding orthography and punctuation (as explained in Section 2.1), and they thus also contain lacunae, varying in size from one letter to complete word(group)s. The *missing* label is only used when we cannot be certain of the original word given its context, similar verses, or commentaries.

Example 5 for instance reads τῆ(...), which is the article of the following word δίκης, which it should agree with. We are quite sure that τῆ should thus be annotated as a genitive feminine singular of the article ὁ and therefore it receives the part-of-speech

tag *article*. Example 6 starts with a lacuna followed by a complete word, which complicates the matter as it could be any word or word group, so here the label *missing* is used.

5. ὁ τῆ(…) δίκης πρύτανι(ς)
   o ti(…) dikis pritanis
   The highest magistrate of the law
   DBBE Occurrence 30520

6. (…) χρόνον τε και λό|γους καὶ τὴν φύσιν |
   (…) χronon te kje logus kje tin fisin
   (...) time and also words and the nature
   DBBE Occurrence 30520

The final change has to do with the phenomenon crasis. A crasis arises when two words, of which the first ends in a vowel/diphthong and the second begins with a vowel/diphthong, 'blend' together and the two blended syllables form a single new syllable. Some examples represented in our corpus are κἄν from καὶ (and) and ἄν (modal particle), κἀμοί from καὶ and ἐμοί (to me), and τοὔνομα from τὸ (the) and ὄνομα (name). We have no exact number of how often crasis occurs in the DBBE because there is no exhaustive list of possible forms. However, by querying well-known forms, we detected over 250 instances of crasis. Since the AGDT does not deal with that grammatical phenomenon, we decided to adopt Keersmaekers' approach [25]: analyse the crasis as the part bearing the highest degree of semantic content. In the case of κἀμοί for example, the pronoun ἐμοί bears more semantic content than the conjunction καί so the crasis is analysed as a pronoun in the dative singular.

Occ. Id	17368	18180	18446	19604	20167	21375	22487	22734
Tokens	50	33	9	101	60	43	91	75
Occ. Id	23607	23615	23631	23632	25463	26551	30520	30844
Tokens	10	12	16	19	52	66	354	31

Table 2. The set of epigrams used for the inter-annotator agreement study, summing up to 1,022 tokens.

Three annotators, linguists with profound knowledge of Ancient Greek, were tasked to annotate the data set listed in Table 2 provided with the label set shown in Table 3. All possible labels for each feature We used Sing et al.'s part-of-speech tagger to pre-analyse or bootstrap our data. Due to that bootstrapping, the annotators were able to tag approximately 80 tokens per hour. Upon completion, we performed an inter-annotator study to measure the overlap between the manually verified labels for part-of-speech

and morphological analysis. Because the IAA experiment was carried out with three annotators, we used Fleiss' Kappa [37] for evaluation:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

With p_e, where p_j is the proportion of all assignments to the j^{th} category:

$$p_e = \sum p_j^2$$

and p_o, where N is the number of tokens and n the number of annotators:

$$p_o = \frac{1}{N \cdot n \cdot (n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - N \cdot n \right)$$

To the best of our knowledge, no IAA experiment has yet been conducted for Ancient Greek, let alone for Byzantine Greek. The inter-annotator study resulted in an agreement of 92.72% for the coarse-grained labels (part- of-speech) and 89.83% for the fine-grained labels (complete morphological analysis). The agreement scores are very high, showing *almost perfect* agreement (>90%) for the part-of-speech tagging and morphological analysis in isolation, and very *strong* agreement (80-90%) for the combined label, according to McHugh's IAA evaluation [38]. These scores are very encouraging, especially because we perform part-of-speech tagging on Greek data, for which different tags are often possible and arguments can be made for different analyses of the same word, a problem Packard [**Error! Reference source not found.**] already encountered. This can be illustrated with the word χάριν (*on behalf of)* followed by a genitive. One can argue that its part-of-speech is a noun, χάρις, since its accusative is used in an adverbial way. It is just as valid however to state that χάριν is an adverb *an sich*. However, the high agreement score of the IAA experiment shows that the Greek language does not cause any difficulties for further annotations.

Feature	Possible labels

Part-of-Speech	adjective, adverb, article, conjunction, exclamation, interjection, punctuation, noun, numeral, particle, preposition, pronoun, verb, missing
Person	1, 2, 3, -
Number	singular, plural, dual, -
Tense	aorist, future, future perfect, imperfect, perfect, pluperfect, present, -
Mood	imperative, indicative, infinitive, optative, participle, subjunctive, -
Voice	active, medial, medio-passive, passive
Gender	common, feminine, masculine, neutral, -
Case	nominative, accusative, genitive, dative, vocative, -
Degree	comparative, superlative, -

Table 3. All possible labels for each feature

4 DBBErt Pipeline

This section introduces the system we developed to perform morphological analysis on unedited Byzantine Greek book epigrams. Since existing systems cannot handle this type of data well, we make use of transformer-based language models which do not solely rely on the written form of a word but take into account the context in which it occurs. Because of this feature, we hypothesise that these transformer-based language models can handle the orthographic inconsistency of our corpus quite well. We carry out a comparative study that evaluates three different kinds of transformer architectures: BERT, ELECTRA and RoBERTa. First, each of these three architectures are trained on both an unlabelled data set containing Ancient and Byzantine Greek, and that same data set complemented with Modern Greek. Next, we fine-tune the models on a supervised part-of-speech tagging task, for which manually labeled data are required.

4.1 Language Model

The assumption is that language models should be capable to deal with the major challenge of our corpus: the inconsistent orthography of the DBBE corpus. That is, a language model does not only capture the way a word is written but it also models the word based on the context in which it is used. This contextual representation of a word as a vector of real numbers is called a word embedding, and words with similar meanings are then closer to one another in the vector space. The first language models processed text from left to right and were trained on a corpus of co-occurrence statistics

[39, 40]. At that point, homonyms like χάριν only have one vector representation independent from their use. This has changed since the development of the transformer architecture [41], the main feature of which is self-attention. Self-attention allows a network to directly extract and use information from arbitrarily large contexts. Devlin et al. [42] developed a language model that is based on this transformer architecture: BERT, or "Bidirectional Encoder Representations from Transformers". What makes transformer models like BERT stand out, is that they do not take into account either the left or right context of the word, but takes into account the whole sentence as context. BERT's training is done by (1) a masked language modelling (MLM) task, for which 15% of the words in the corpus were masked and BERT had to generate the correct word, and (2) Next Sentence Prediction (NSP), where the model had to predict a sentence B, given sentence A. These transformer-based language models are state-of-the-art and since the release of BERT, several variants have been developed, each with their own focus. Some examples of these variants are RoBERTa, a more robust BERT model [43], SpanBERT, representing and predicting spans of text [44] and ELECTRA ("Efficiently learning an Encoder that Classifies Token Replacements Accurately"), using a more efficient training than MLM [45]. To develop the optimal language model for Byzantine Greek, we compare three different language model architectures: BERT, ELECTRA and RoBERTa.

RoBERTa and BERT have two main differences: first, RoBERTa makes use of a byte-level BPE tokenizer [46, 47] while BERT makes use of a subword tokenizer; second, RoBERTa does not add next-sentence prediction to the MLM task and trains with larger mini-batches and learning rates than BERT does.

The main difference between BERT and ELECTRA is the way they are pre-trained. The MLM task of BERT's training is a generative task: 15% of the tokens are replaced by [MASK] and a generator predicts the original identity of the corrupted token. ELECTRA's training, however, is done in a discriminative way with *replaced token detection*: the model learns to distinguish real input from plausible but synthetically generated replacements, not from the [MASK] token.

Data. As described in Section 3.1, the Byzantine Greek data set is limited. A language model, however, is very data greedy. Therefore, we created a second data set that supplemented the Ancient and Byzantine Greek set with Modern Greek. These two data sets enable us to evaluate whether Modern Greek contributes to the performance of the architectures to model Byzantine Greek, as Byzantine Greek is situated in time between Ancient and Modern Greek. Henceforth the data set containing Ancient and Byzantine Greek is referred to as the AB set, the one complemented with Modern Greek as ABM set. The composition of both is shown in **Error! Reference source not found.**.

Training of the Language Models. To identify the best possible architecture and data set for Byzantine Greek, the six following combinations of models and data sets have been trained: BERT on AB set, BERT on ABM set, ELECTRA on AB set, ELECTRA

on ABM set, RoBERTa on AB set and RoBERTa on ABM set. These six setups have been trained during 12 epochs with their default parameters' configuration, as pointed out in.

Model	Data set	Vocab size	hidden size	hidden layers	attention heads	max pos embedding
BERT	A B	28,996	768	12	12	512
BERT	A B M	28,996	768	12	12	512
ELECTRA	A B	30,522	256	12	4	512
ELECTRA	A B M	30 522	256	12	4	512
RoBERTa	A B	50,265	768	12	12	514
RoBERTa	A B M	50,265	768	12	12	514

Table 4. The configuration of the six language models and accompanying parameters used for training.

Fig. 2 shows the loss as a function of time of a held-out validation set during training. The loss functions, which represent the distance between the predicted and correct labels, show that all three models trained on the smaller AB data set perform worse than their counterparts trained on the bigger ABM data set. Although the ELECTRA ABM model starts with a loss as high as the three smaller models, its loss drops more than the small models do. It is, however, the worst performing model of the bigger models. Although the BERT ABM model and the RoBERTa ABM model display a very similar graph, the RoBERTa ABM model does outperform the BERT ABM model. Although these loss functions already serve as an indication of how well the different architectures model Byzantine Greek, a proper extrinsic evaluation on an end task is essential.

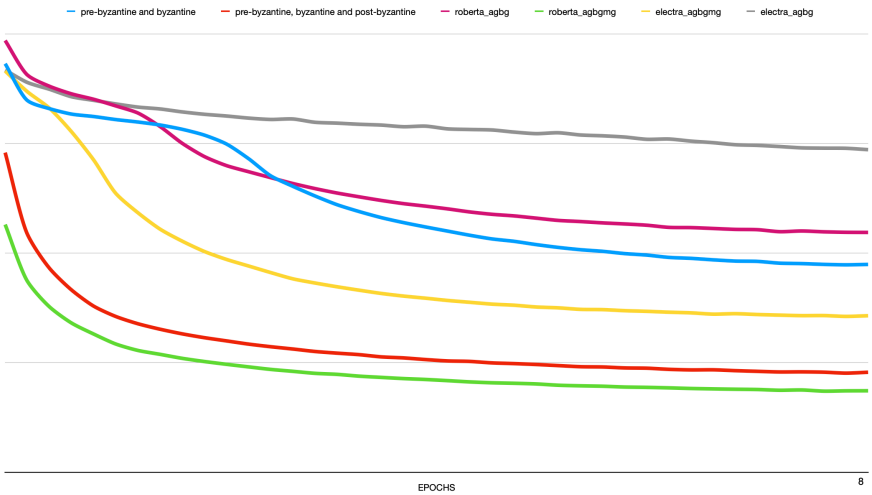

Fig. 2. The loss functions of the trained language models

4.2 Fine-Tuning: Morphological Analysis

This extrinsic evaluation of the language models is carried out by fine-tuning the models for part-of-speech tagging. This task, however, is quite challenging as it implies a complete morphological analysis, consisting of 9 features (cf. Table 3). Such a complex combination of morphological information leads to a very big set of possible labels, viz. 1,057. Because of the high number of labels for a relatively small data set, we dare say this is a quite difficult machine learning task.

Data. The training of a part-of-speech tagger is a supervised machine learning task, which means that labelled data is needed. As described in Section 3.1, we extracted tokens with their morphological analysis from the AGDT, Gorman, PROIEL and Pedalion and we deleted all duplicate texts. The resulting training set summed up to 1,132,120 tokens. For evaluation, we want to focus on the morphological analysis of unedited Greek texts, for which we could rely on our gold standard consisting of 10,000 tokens extracted from the DBBE *occurrences*, which admittedly is more modest in size.

Fine-Tuning the Language Model for Part-of-Speech Tagging. Each language model described in Section 4.1 will be evaluated by fine-tuning it on the task of part-of-speech tagging. We make use of the same architecture, facilitated by the FLAIR framework [33], to train the various part-of-speech taggers based on the six contextual embeddings obtained by our language models. These contextual embeddings are stacked together with randomly initialised character embeddings, which is beneficial for highly inflectional languages [48] like Greek. The stacked embeddings are subsequently processed by a bi-directional long short-term memory (LSTM) encoder and a conditional random field (CRF) decoder, a combination frequently used in sequence tagging tasks. We use a hidden size of 256, set the learning rate to 0.1 and train for 5 epochs.

The validation loss in function of time is shown in **Fig. 3**. The results of these loss functions deviate somewhat from the loss functions displayed in **Fig. 2**. The two ELECTRA models perform worst, irrespective of the data set used. The smaller BERT and RoBERTa models slightly exceed the performance of both the ELECTRA models, but clearly underperform when compared to their bigger counterparts. The functions of the BERT ABM model and RoBERTa ABM model are very similar, but, unlike the language models' loss functions, the BERT ABM model slightly outperforms the RoBERTa ABM model.

As a last element of comparison, we directly fine-tuned the embeddings of the best performing BERT (ABM) model for token classification instead of using them as input in the FLAIR architecture. We did this both for the coarse-grained part-of-speech tags

and the combined label of part-of-speech and morphological analysis. We will refer to these models as DBBErt_pos and DBBErt_morph, respectively.

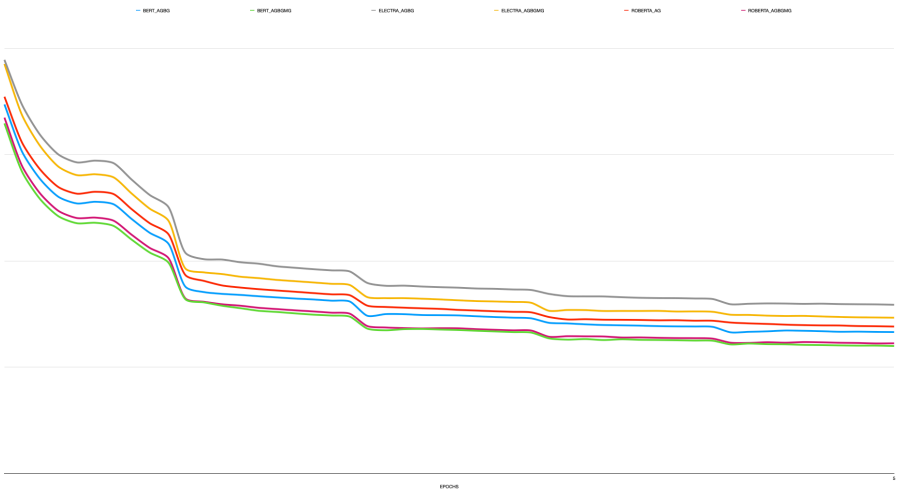

Fig. 3. The loss functions of the fine-grained part-of-speech taggers

5 Results

To the best of our knowledge, this is the first research to evaluate part-of- speech tagging of original, Byzantine Greek texts that were not normalised. The accuracy scores of the coarse-grained part-of-speech tags are shown in Table 5. Keeping in mind that the training is done on normalised, Ancient Greek, the 82.76% accuracy of the BERT ABM model on unedited Byzantine Greek is very satisfying.

The results of the fine-grained part-of-speech tagging are provided in Table 6 and visualised in Figure 4. These results also look very promising, especially given the complexity of the task: every label consists of nine slots and each slot had three to fourteen possible options. Our train set of more than 1 million tokens had a label set of 995 labels, the test set consisting of some 10,000 tokens had a label set of 471 labels. 108 of these 471 labels in the test set, however, do not occur in the train set and can thus not be predicted by our model.

Model	Accuracy	Precision	Recall	F1
BERT AB	77.44%	78.58%	77.44%	77.64%
BERT ABM	**82.76%**	**83.01%**	**82.76%**	**82.67%**
ELECTRA AB	71.62%	73.53%	71.62%	72.09%
ELECTRA ABM	79.65%	79.41%	79.10%	79.07%
RoBERTa AB	71.03%	72.55%	71.03%	71.13%
RoBERTa ABM	76.32%	76.90%	76.52%	76.32%
DBBErt_pos	80.50%	80.50%	82.17%	79.33%

Table 5. Evaluation scores of the coarse-grained part-of-speech tags.

Model	Accuracy	Precision	Recall	F1
BERT AB	63.29%	69.19%	63.29%	62.14%
BERT ABM	**68.57%**	**73.22%**	**68.57%**	**67.32%**
ELECTRA AB	56.99%	64.95%	56.99%	56.01%
ELECTRA ABM	63.27%	69.49%	63.27%	62.42%
RoBERTa AB	56.23%	66.43%	56.23%	55.35%
RoBERTa ABM	61.76%	71.22%	61.76%	60.63%
DBBErt_morph	62.33%	62.33%	64.84%	59.83%

Table 6. Evaluation scores of the full morphological analysis.

These results, however, do not completely confirm our tentative conclusions, which were based on the loss functions of both the language model and the training of the part-of-speech tagger. The RoBERTa AB model displays the worst performance with an F1-score of 55.35% and an accuracy of 56.23%, the ELECTRA AB model performs slightly better with an F1-score of 56.01% and an accuracy of 56.99%. Yet the precision of the RoBERTa AB model is striking: this model seems to predict less false positives than the ELEC- TRA AB model. The same observation holds for the RoBERTa ABM model as well. Although outperformed by both the ELECTRA ABM and BERT ABM models for the other metrics, the precision of RoBERTa ABM is higher.

The BERT AB model and ELECTRA ABM model perform almost identically. ELECTRA ABM shows a slightly better precision and F1-score than the BERT AB model, while accuracy and recall are 0.02% worse. However, **Fig. 4** clearly shows that the BERT ABM model performs best on this task combining part-of-speech tagging and morphological analysis.

The DBBErt fine-tuning performs quite mediocre with an accuracy score of 62.33% and an F1 score of 59.83%. This is not surprising given the basic architecture that was uses (1 classification layer). In future research, we will experiment with more advanced architectures that, for instance, implement a cascaded system that firstly predicts the part-of-speech tag, after which it only predicts the morphological features relevant for that part-of-speech.

As it is difficult to compare the results to previous research, we created two baselines by applying Morpheus, a widely used analyser for Ancient Greek, and RNN Tagger, a neural state-of-the-art tagger, on our novel gold standard described in Section 3.1. Morpheus resulted in an accuracy score of 19.92%. That score, however, needs some interpretation: 44% of the tokens could not be processed by the Morpheus algorithm and 30% of the tokens had multiple possible analyses so only 27% of the test set was provided with only one morphological analysis. RNN Tagger yielded an accuracy score of 65.59% on the morphological analysis. In addition, we also created a frequency-based baseline in order to show how much the transformer model learns from the context rather than from the word alone. This baseline yielded an accuracy score of 34.08%.

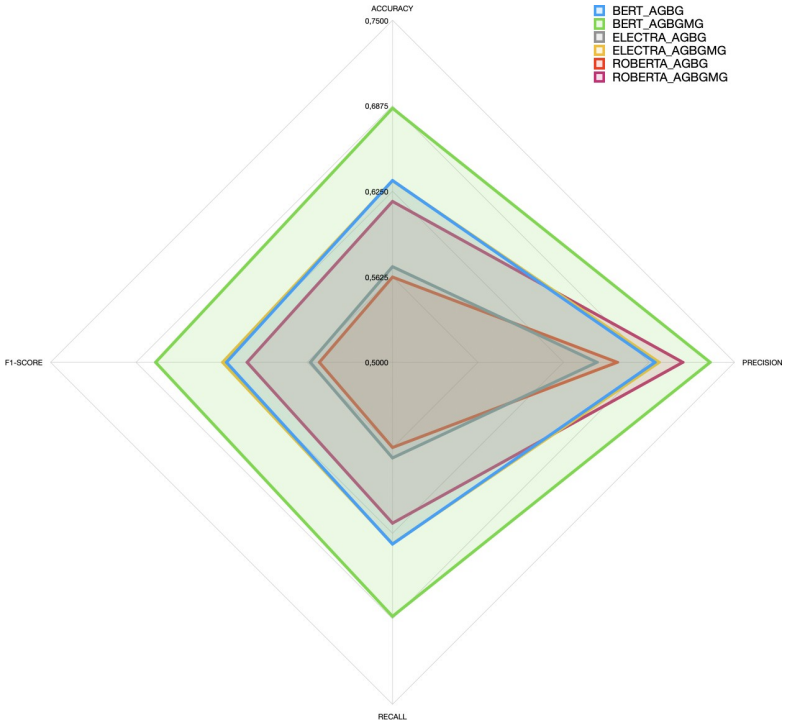

Fig. 4. The accuracy (top), precision (right), recall (bottom) and F1 (left) of our 6 part-of-speech taggers.

5.1 Qualitative Analysis

In addition to the scores that show the performance of the part-of-speech taggers, we are also interested in the mistakes that were made and whether some tendencies could be discovered. For this more in-depth discussion, we compared the labels predicted by the best model, viz. BERT ABM, to our gold standard. The results of this analysis are classified in three categories and are provided with some examples. An overall trend is that the mistakes made by the algorithm strike us as quite "human", i.e., typical learners' mistakes.

Itacism. Words that are written completely different than their Ancient Greek forms due to the itacism, are in general analysed wrongly. A notable example is ἰδῦν /iðin/, the active infinitive aorist of the verb to see usually written as ἰδεῖν. This form, which is extremely affected by the itacism, occurs five times in our test set and is analysed differently each time:

- *adverb*

- *verb, active present participle nominative masculine plural*
- *noun, genitive masculine plural*
- *adjective, accusative neutral plural*
- *noun, genitive neutral plural*

Granted, the word is quite unrecognisable if not pronounced out loud. How- ever, that does not alter the fact that the correct coarse-grained part-of-speech tag, namely verb, is predicted correctly only once.

However, words of which the stem is affected by the itacism but still display a correct suffix, are analysed correctly most of the time. This might be due to the subword tokenizer, that splits the words in meaningful parts. For example, if the last part of the word is a clear suffix that indicates a nominative masculine singular, the word is generally analysed as such, irrespective of whether the stem has been altered or not.

7. + ἄναρχε θεὲ· τῶ τρησώλβιων βήος
   God that cannot be dominated, you, thrice happy life
   DBBE Occurrence 17859

Example 7 ends with the word βήος (*life*), which in Ancient Greek is written as βίος. The stem of the word is spelled differently but the suffix -ος is left unchanged, which explains why the algorithm managed to correctly analyse this word as a nominative masculine singular.

Ambiguous Morphology. Ambiguous forms are not always provided with the correct analysis, in spite of context being captured by the model. The word ἔργα, for example, is either a nominative or an accusative neutral plural. In Example 9 the word φύσις (nature) is nominative and serves as subject, while ἔργα (tasks) is an accusative following the preposition πρός (towards). Even though this sentence already has an unquestionable subject in the nominative, ἔργα was erroneously tagged as a nominative as well.

8. θαῦμα πρόκειται πᾶσιν εξηρημένον
   is for all a miracle to behold
   DBBE Occurrence 17013

9. οἷς ἐστιν φύσις πρὸς ἔργα φιλότιμος
   Those who by nature honour works involving much noble toil
   DBBE Occurrence 17013

Next, our model shows a preference for the masculine gender in its analyses. This manifests itself in two different ways. First, adjectives that display no morphological distinction between masculine and feminine are generally analysed as masculine. The last word of Example 9 for instance, φιλότιμος (*loving honour*), is an adjective that has the same morphology for both masculine and feminine forms. Our model analyses this as a masculine form, notwithstanding the word's agreement with the female noun φύσις (*nature*). Second, adjectives in accusative case look identical for masculine and neutral. The word ἐξηρημένον (*transcendent*) in Example 8 agrees with the accusative neutral θαῦμα (*wonder*) but is tagged as a masculine accusative. This might be ascribed to the gender distribution in our corpus: 49% of the words that have a grammatical gender are masculine, 29% are feminine, 18% is neutral and 5% is *common*, viz. not determinable by morphology nor by context.

A surprising outcome related to gender is the word παῖς (*child*), which can be either masculine or feminine, depending on the person it describes. In our test set, the following forms of the word παῖς occur:

10. Κλαδηφοροῦντες παῖδες Ἑβραίων […] κράζουσιν
    The children of the Hebrews, bearing young branches, are shouting
    DBBE Occurrence 17304

11. ἀγγελίην πολύστενον· αἰ αἰ κλαύσατε παῖδες
    Alas, children, weep for the mournful messages
    DBBE Occurrence 25465

12. Αἶνος φλόγα σβέννυσι τῶν τριῶν παίδων
    The hymn of three children extinguishes the fire
    DBBE Occurrence 27019

Example 10 is annotated as masculine in the gold standard, which might stem from the agreement between παῖδες and the adjective Κλαδηφοροῦντες (bearing branches), which is undoubtedly masculine. However, although neither Example 11 nor 12 has any attributive adjunct that is unarguably masculine or feminine, both are annotated as masculine. This might be due to the fact that the word παῖς is used with a masculine article in 90% of the occurrences in the classical Greek corpus[12]. The algorithm nonetheless tagged both Example 11 and 12 as *common*, while 10 is correctly tagged as masculine.

Another striking *error* is the classification of uninflected nouns, like the Hebrew name Δαυίδ (*David*). This token occurs 8 times in our test set. During manual annotation, we always interpreted the token Δαυίδ within the sentence and labelled it with the

[12] We queried the Philologic database (https://perseus.uchicago.edu) for all collocations of the masculine article with παῖς and all collocations with the female article.

case we would expect that constituent to be. So, if David is the subject of the sentence, we tagged it as a nominative; if it is the sentence's object, we tagged it as an accusative. The model, however, only provides a – correct – coarse-grained part-of-speech tag without further morphological analysis in 5 of the 8 occurrences of Δαυίδ. It also predicted twice a nominative masculine singular and once a dative masculine plural; only one nominative masculine singular was correct, the other two were vocative masculine singular. Although unexpected, the fact that more than half of the occurrences of Δαυίδ has not received a morphological analysis, might be due to the absence of any suffix indicating the case of the word. The uninflected names Ἰεζεκιήλ (*Ezechiel*) and Μανουήλ (*Manuel*), however, are provided with a perfect morphological analysis, while Δανιήλ (*Daniel*) was erroneously tagged as feminine instead of masculine. These examples, unlike the examples of Δαυίδ, clearly display that the model is able to make correct, complete morphological analyses without the presence of a suffix.

Manuscript Writing. Other wrong outputs can be ascribed to typicalities of the texts in our corpus, like the sloppiness of many scribes. The iota subscriptum is every so often omitted in manuscripts and thus in our *occurrences* as well[13]. It is notable that the dative singular of words that have a nominative singular ending with -α or -η, are often analysed wrongly by the model. This is consistent with the training data since the dative of the feminine nouns ending with -α or -η looks the same as the nominative singular because of the omission of the iota. They can be distinguished by taking into account their article – if present –, that is ἡ for a nominative, shown in Example 13 and τῇ/τη (with or without a iota) for a dative, shown in Example 14.

13. ὄψον ἡ ψαλμωδία nominative
as dish singing to the harp ‡
DBBE Occurrence 20204

14. ἡμέρα τῆ τῆς δίκης dative
on Judgement Day ‡
DBBE Occurrence 28563

The in-depth analysis of BERT ABM's predictions shows that the model is quite capable of coping with the peculiarities of unedited, Byzantine texts as described in Section 2.1. This proposition is valid as long as the orthographic inconsistencies do not affect the morphological suffixes, caused either by the itacism or the sloppiness of the scribe.

[13] When an ᾱ, η or ω is followed by a ι (iota), the iota is written either underneath (*subscriptum*) or next to that previous vowel (*adscriptum*).

6 Conclusions & Future Research

This paper reports on the development of a transformer-based part-of-speech tagger for Byzantine Greek. To this end, first two large corpora were compiled: a data set (*AB*) of almost 42 million tokens of Ancient and Byzantine Greek and a data set (*ABM*) of almost 127.5 million tokens containing Ancient, Byzantine, and Modern Greek texts. Both text collections were used to train six different transformer-based language models, that were fine-tuned by means of existing treebanks for the task of part-of-speech tagging. To evaluate the tagger, a gold standard of around 10,000 tokens of DBBE occurrences was manually annotated. An inter-annotator agreement study, which was performed to verify the reliability of the annotations, resulted in very high agreement scores (89.83%).

The development of the part-of-speech tagger involved two main steps. Firstly, we trained a BERT, ELECTRA and RoBERTa language model on both the AB data set and the ABM data set. Each of the embeddings obtained by the language models were then incorporated for training the different part-of-speech taggers. The comparison of these part-of-speech taggers shows that BERT trained on Ancient, Medieval *and* Modern Greek achieves the best results on this task. Although the model leaves room for improvement, an accuracy score of almost 70% on full morphological analysis is very promising, given the difficulty of the task.

In further research, we will keep collecting and annotating data to improve the presented model for part-of-speech tagging. As a next step, we will investigate a linguistic annotation pipeline that combines our part-of-speech tagger with a lemmatiser for Byzantine Greek. We will also experiment with a cascaded approach, where the detailed morphological analysis is performed based on the predicted course-grained part-of-speech category. In addition, we will investigate whether an ensemble of dedicated models, each predicting one of the components of the detailed morphological analysis separately, outperforms the current model that predicts the complete morphological analysis label at once.

Finally, the obtained linguistic annotation will allow to research similarities between the occurrences in the quite entangled corpus of the DBBE. To this end, various methodologies will be investigated to measure orthographic and semantic similarity, e.g. based on word forms, part-of-speech patterns or lemmas, for which we need this linguistic information. The *occurrences*, now linked to one another only if they have a common *type*, could then be "linked" by these similarity detection algorithms. This will result in a more dynamic system to connect related epigrams within the DBBE.

Declarations

- Funding: This study was funded by Ghent University
- Conflict of interest/Competing interests: The authors have no conflicts of interest to declare that are relevant to the content of this article.
- Ethics approval: Not applicable.

- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and materials: The data is available at https://huggingface.co/datasets/colinswaelens/DBBErt , the language model at https://huggingface.co/colinswaelens/DBBErt .
- Code availability: The code for fine-tuning the language model on the part-of-speech tagging task is available at https://github.com/coswaele/DBBErt_POS/blob/main/fine_tuning.py.
- Authors' contributions: Colin Swaelens (the corresponding author) wrote the main manuscript, performed the experiments, and created the figures for the manuscript. Prof. Dr. Els Lefever is the primary supervisor for the PhD project of Colin Swaelens. Dr. Ilse De Vos is co-supervisor for the PhD project of Colin Swaelens. Both supervisors reviewed the manuscript and advised on the experiments.

References

1. Kominis, A.D.: To Byzantinikon Hieron Epigramma Kai Hoi Epigrammatopoioi, Athens (1966)
2. Lauxtermann, M.D., der Wissenschaften. Kommission für Byzantinistik, Ö.A., für Byzantinistik und Neogräzistik, U.W.I.: Byzantine Poetry from Pisides to Geometres: Epigrams in Context. Byzantine Poetry from Pisides to Geometres: Texts and Contexts. Verlag der Österreichischen Akademie der Wissenschaften, Wien (2003). https://books.google.be/-books?id=hJdYzgEACAAJ
3. West, M.L.: Homeri Ilias: Rhapsodias I-XII Continens. Bibliotheca Teubneriana. Saur, Muenchen, Leipzig (1998)
4. Bernard, F., Demoen, K.: Byzantine book epigrams from manuscripts to a digital database. In: Clivaz, C., Meizoz, J., Vallotton, F., Verheyden, J. (eds.) From Ancient Manuscripts to the Digital Era : Readings and Literacies, Proceedings, p. 8. PPUR, Lausanne, Switzerland (2012)
5. Ricceri, R., Bentein, K., Bernard, F., Bronselaer, A., De Paermentier, E., De Potter, P., De Tré, G., De Vos, I., Deforche, M., Demoen, K., Lefever, E., Rouckhout, A.-S., Swaelens, C.: The Database of Byzantine Book Epigrams Project: Principles, Challenges, Opportunities. working paper or preprint (2023). https://hal.science/hal-03833929
6. Pantelia, M.C.: Thesaurus Linguae Graecae, A Bibliographic Guide to the Canon of Greek Authors and Works. University of California Press, Berkeley (2022). https://doi.org/10.1525/9780520388208
7. Crane, G.R.: Perseus Digital Library. Tufts University. Last accessed 30 March 2023 (2023). http://www.perseus.tufts.edu
8. Bamman, D., Crane, G.: The ancient greek and latin dependency treebanks. In: Sporleder, C., van den Bosch, A., Zervanou, K. (eds.) Language Technology for Cultural Heritage, pp. 79–98. Springer, Berlin, Heidelberg (2011)
9. Celano, G.G.: The dependency treebanks for ancient greek and latin. Digital Classical Philology, 279 (2019)
10. Haug, D.T., Jøhndal, M.: 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), pp. 27–34 (2008)

11. Gorman, V.B.: Dependency treebanks of ancient greek prose. Journal of Open Humanities Data 6(1) (2020)
12. Keersmaekers, A., Depauw, M.: Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri, Digital Classics, Heidelberg (2022)
13. Keersmaekers, A., Mercelis, W., Swaelens, C., Van Hal, T.: Creating, enriching and valorizing treebanks of Ancient Greek. In: Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), pp. 109–117. Association for Computational Lin- guistics, Paris, France (2019). https://doi.org/10.18653/v1/W19-7812. https://aclanthology.org/W19-7812
14. Memon, J., Sami, M., Khan, R.A., Uddin, M.: Handwritten optical char- acter recognition (ocr): A comprehensive systematic literature review (slr). IEEE Access 8, 142642–142668 (2020)
15. Holton, D., Horrocks, G., Janssen, M., Lendari, T., Manolessou, I., Toufexis, N.: Phonology, pp. 1–238. Cambridge University Press, Cam- bridge (2019)
16. Packard, D.W.: Computer-assisted morphological analysis of Ancient Greek. In: COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics (1973). https://aclanthology.org/C73-2026
17. Crane, G.: Generating and Parsing Classical Greek. Literary and Linguistic Computing 6(4), 243–245 (1991) https://aca- demic.oup.com/dsh/article-pdf/6/4/243/10889452/243.pdf. https: //doi.org/10.1093/llc/6.4.243
18. Niehoff-Panagiotidis, J.: Koine und Diglossie. Mediterranean language and culture monograph series. Harrassowitz, Wiesbaden (1994). https://-books.google.be/books?id=VurqESJjJSYC
19. Miller, D.G.: Ancient Greek Dialects and Early Authors: Introduction to the Dialect Mixture in Homer, with Notes on Lyric and Herodotus. De Gruyter, Berlin, Boston (2014). https://doi.org/10.1515/9781614512950. https://doi.org/10.1515/9781614512950
20. Celano, G.G.A., Crane, G., Majidi, S.: Part of speech tagging for ancient greek. Open Linguistics **2**(1) (2016). https://doi.org/10.1515/ opli-2016-0020
21. Bohnet, B., Nivre, J.: A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1455–1465. Association for Computational Linguistics, Jeju Island, Korea (2012). https://aclanthology.org/D12-1133
22. Halácsy, P., Kornai, A., Oravecz, C.: Hunpos: an open source trigram tagger. Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, 209–212 (2007)
23. Schmid, H., Laws, F.: Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. COLING '08, pp. 777–784. Association for Computational Linguistics, USA (2008)
24. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pp. 69–72 (2006)
25. Keersmaekers, A.: Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language. Digital Scholarship in the Humanities 35(1), 67–82 (2019) https://academic.oup.com/dsh/article- pdf/35/1/67/32976739/fqz004.pdf. https://doi.org/10.1093/llc/fqz004

26. Müller, T., Schmid, H., Schütze, H.: Efficient higher-order CRFs for morphological tagging. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 322-332. Association for Computational Linguistics, Seattle, Washington, USA (2013). htps://aclanthology.org/D13-1032
27. Johnson, K.P., Burns, P.J., Stewart, J., Cook, T., Besnier, C., Mattingly, W.J.: 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, pp. 20–29 (2021)
28. Vatri, A., McGillivray, B.: The diorisis ancient greek corpus: Linguistics and literature. Research Data Journal for the Humanities and Social Sciences **3**(1), 55–65 (2018). https://doi.org/10.1163/24523666-01000013
29. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK (1994)
30. Schmid, H.: 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, pp. 133–137 (2019)
31. Singh, P., Rutten, G., Lefever, E.: A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. In: The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Co-located with EMNLP 2021, pp. 128–137 (2021). Association for Computational Linguistics
32. Koutsikakis, J., Chalkidis, I., Malakasiotis, P., Androutsopoulos, I.: Greek-bert: The greeks visiting sesame street. In: 11th Hellenic Confer- ence on Artificial Intelligence. SETN 2020, pp. 110–117. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10. 1145/3411408.3411440. https://doi.org/10.1145/3411408.3411440
33. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 54–59. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-4010. https://aclanthol- ogy.org/N19-4010
34. Rhoby, A.: Ausgewählte Byzantinische Epigramme in Illuminierten Handschriften. Verlag der österreichischen Akademie der Wissenschaften, Wien (2018). https://doi.org/10.2307/j.ctv1wxr6c
35. Berkowitz, L., Squitier, K.A.: Thesaurus Linguae Graecae : Canon of Greek Authors and Works, 2nd ed. edn. Oxford university press, New York (N.Y.) (1986)
36. Celano, G.G.A.: GUIDELINES FOR THE ANCIENT GREEK DE- PENDENCY TREEBANK 2.0. Last consulted December 2022 (2018). https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/ guidelines/Greek_guidelines.md
37. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin 76 (5), 378 (1971)
38. McHugh, M. L.: Interrater Reliability: the Kappa Statistic. Biochemia Medica, 22 (3), 276-282 (2012)
39. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)
40. Pennington, J., Socher, R., Manning C. D.: GloVe: Global Vector for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP). (2014)

41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
42. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
43. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
44. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics 8, 64–77 (2020)
45. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
46. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
47. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
48. Vylomova, E., Cohn, T., He, X., Haffari, G.: Word representation models for morphologically rich languages in neural machine translation. arXiv preprint arXiv:1606.04217 (2016)