Lemmatisation & Morphological Analysis of Unedited Greek: Do Simple Tasks Need Complex Solutions?

Colin Swaelens[1], **Ilse De Vos**[2] and **Els Lefever**[1]
[1] LT3, Language and Translation Technology Team, Ghent, Belgium
[2] VAIA - Flanders AI Academy, Leuven, Belgium
{colin.swaelens, els.lefever}@ugent.be, ilse.devos@kuleuven.be

Abstract

Fine-tuning transformer-based models for part-of-speech tagging of unedited Greek text has outperformed traditional systems. However, when applied to lemmatisation or morphological analysis, fine-tuning has not yet achieved competitive results. This paper explores various approaches to combine morphological features to both reduce label complexity and enhance multi-task training. Specifically, we group three nominal features into a single label, and combine the three most distinctive features of verbs into another unified label. These combined labels are used to fine-tune DBBERT, a BERT model pre-trained on both ancient and modern Greek. Additionally, we experiment with joint training – both among these labels and in combination with POS tagging – within a multi-task framework to improve performance by transferring parameters. To evaluate our models, we use a manually annotated gold standard from the Database of Byzantine Book Epigrams. Our results show a nearly 9 pp. improvement, demonstrating that multi-task learning is a promising approach for linguistic annotation in less standardised corpora.

1 Introduction

The development of automatic linguistic annotation for Ancient Greek began in the 1970s, initially driven by pedagogical needs (Packard, 1973), and soon expanded to accommodate the language's diverse dialects (Crane, 1991). Since then, researchers in natural language processing (NLP) have continuously improved automatic annotation systems, addressing the limitations of earlier approaches. Why, then, develop yet another tagger more than thirty years after the first? The answer lies in the evolving nature and availability of the data.

Most existing systems have been optimised for *standardised* Ancient Greek – texts that reflect critical editions rather than original, unedited sources. However, the increasing availability of 'imperfect' or fragmentary texts introduces new challenges that traditional taggers struggle to handle. Example 1 illustrates some of these challenges: Example 1a presents the diplomatic transcription of a book epigram from manuscript *Vat. gr.* 169, whereas Example 1b shows the edited, readable version of that same epigram provided by the Database of Byzantine Book Epigrams (DBBE) (Ricceri et al., 2023). Even without knowledge of ancient Greek, it is clear that the orthography of Example 1a deviates from the standardised version in Example 1b.[1] A second challenge in Example 1 is the irregular use of diacritics, which are essential for disambiguating tokens with identical orthography. However, in Example 1a, these diacritics appear inconsistently, making disambiguation more difficult.

(1) a. + η μεν φηλος πέφηκ(ας), ἥσελθ(ε) χερον·
è men phèlos pefèk(as), èselthe cheron
ἡ δ ἐχθρὸς και βάσκανος· καὶ γεμ(ὸν) ὦλος δόλον·
è d echthros kai baskanos· kai gem(on) ōlos dolon·
πῶρο πῶρο | πεφευγε τὴν πυλ(ὴν) ταυτ(ὴν).
pōro pōro | pepheuge tèn pyl(èn) taut(èn)
DBBE Occurrence 27108

b. Εἰ μὲν φίλος πέφυκας εἴσελθε χαίρων.
Ei men philos pephykas eiselthe chairōn.
Εἰ ἐχθρὸς καὶ βάσκανος καὶ δόλου γέμων
Ei echthros kai baskanos kai dolou gemōn
πόρρω ἄπελθε τῆς δε τῆς βίβλου.
porrhō apelthe tès de tès biblou.
DBBE Type 4235

[1] These inconsistencies are due to the phonetic change called *itacism*.

Findings of the Association for Computational Linguistics: ACL 2025, pages 7681–7689

c. If you are a friend, come in, be welcome.
If you are an enemy, a slanderer and full of wiles
Away, go far away from this book.[2]

A final challenge concerns corpus uniformity. The DBBE, for instance, expands manuscript abbreviations and, when necessary, applies the correct diacritics to the expanded parts. This raises broader questions about how other datasets handle abbreviations and how automatic transcription systems approach this issue.

The rise of automatic systems for optical character recognition (OCR) and hand-written text recognition (HTR) has led to a growing influx of unedited Greek texts, where existing annotation systems generally fail to perform reliably.

To address the challenges presented above, we move beyond minor tweaks and reconsider how linguistic annotation can be approached. First, we examine previous methods and their limitations, then turn to multi-task learning to assess its potential for improving annotation quality. Building on recent research in part-of-speech tagging for unedited Greek texts (Swaelens et al., 2023), we extend the annotation process by enriching part-of-speech tags with lemma information and relevant morphosyntactic features. These include case, gender, and number for nominal tokens, as well as voice, mood, and tense for verbal tokens. Through targeted experiments, we evaluate whether combining features improves accuracy and where the approach still struggles. Finally, we reflect on the broader implications for computational philology and suggest directions for future work, including the potential of multi-task learning to fine-tune a transformer model pretrained on Greek.

2 Literature Review

Unlike raw-text corpora, linguistically annotated resources for Ancient Greek remain sparse. The largest annotated corpus, *Thesaurus Linguae Graecae* (Pantelia, 2022), is subscription-based and not available for research. However, two comprehensive annotated corpora have recently been released: the GLAUx corpus (Keersmaekers, 2021) and the *Opera Graeca Adnotata* (OGA) (Celano, 2024a). The GLAUx corpus consists of 20M tokens, spanning from the 8th c. BC. to the 4th c. AD. It is primarily annotated automatically with part-of-speech tags, morphological features, lemmas, and syntactic information, while a smaller subset has been manually annotated, using data from several smaller treebanks: PROIEL (Haug and Jøhndal, 2008), the Ancient Greek Dependency Treebanks (Bamman and Crane, 2011), Pedalion Trees (Keersmaekers et al., 2019), Gorman Trees (Gorman, 2020), Harrington Trees[3] and the treebank of Aphthonius' Progymnasmata[4]. The OGA corpus, in contrast, contains 40M tokens, all annotated automatically.

Early systems for ancient Greek linguistic annotation were rule- and dictionary-based taggers (Packard, 1973; Crane, 1991). While still widely used, these approaches face persistent issues: they do not resolve ambiguous tokens, fail to process out-of-vocabulary tokens, and lack mechanisms to handle crasis (the contraction of two words into one).

The emergence of transformer-based language models (Vaswani et al., 2027) has addressed two of these challenges by incorporating contextual information when tagging tokens. However, crasis remains problematic for automatic solutions. Singh et al. (2021) introduced the first transformer-based language model for ancient Greek, Ancient Greek BERT, which was initialised using a modern Greek BERT model (Koutsikakis et al., 2020) and subsequently trained on monolingual data from the First1KGreek Project[5], the Perseus Digital Library[6], the PROIEL Treebank and Gorman's Treebank. Since the base model was pretrained on Modern Greek, the fine-tuned model for Ancient Greek can only handle tokens without diacritics. Nonetheless, it achieves state-of-the-art results for part-of-speech tagging and lemmatisation. Swaelens et al. (2023) pretrained a BERT model on Greek texts ranging from Homeric poetry (800 BC.) to Byzantine literature and Modern Greek Wikipedia, then fine-tuned it for part-of-speech tagging and lemmatisation, with a focus on unedited Greek texts. While its part-of-speech tagging performance was competitive to existing taggers, its lemmatisation results were unsatisfactory (Swaelens et al., 2024b). Similarly, Swaelens et al. (2024a) demonstrated that transformer-based

[2]Translation by *anonymised*.

[3]https://perseids-publications.github.io/harrington-trees/

[4]https://github.com/polinayordanova/Treebank-of-Aphtonius-Progymnasmata

[5]https://opengreekandlatin.github.io/First1KGreek/

[6]https://www.perseus.tufts.edu

morphological annotation requires further refinement, as its performance has yet to reach competitive levels. A comparative study evaluating state-of-the-art models for morphosyntactic annotation and lemmatisation (Celano, 2024b) found that the Dithrax model yielded best results for morphosyntactic annotation, while GreTa, a multilingual LLM (Riemenschneider and Frank, 2023), performed best on lemmatisation. This multilingual transformer model trained on Greek, Latin, and English was fine-tuned for multiple downstream tasks, including part-of-speech tagging and lemmatisation (Riemenschneider and Frank, 2023).

In this paper, we investigate a multi-task learning approach to predict part-of-speech, lemma and fine-grained morphological analysis. Multi-task learning in machine learning is inspired by human learning, where knowledge transfers between related tasks (Zhang and Yang, 2017). Assigning part-of-speech tags, morphological features, and lemmas are highly interdependent tasks, making them well-suited for experimentation with multi-task learning.

3 Data

We fine-tuned DBBErt, a BERT model trained for classification tasks of ancient Greek. This language model is pretrained on a dataset of 127.413.536 tokens which spans nearly 3,000 years, ranging from Homeric poetry to medieval literature and modern Greek Wikipedia data (cf. supra); this language model is described in greater detail by Swaelens et al. (2023). The data used for fine-tuning is extracted from the Pedalion Trees. From the xml-files, we extracted the attributes part-of-speech tag (e.g. v-pppanm-) and lemma for each token. This part-of-speech tag is the combination of nine slots: part-of-speech, person, number, tense, mood, voice, gender, case, and degree. An example sentence is shown in Table 1. The Pedalion treebanks are particularly valuable, as they include the Trismegistos papyrus corpus (Depauw et al., 2014), the only freely available resource containing unedited Greek. This set for fine-tuning sums up to 5,808,465 tokens. Finally, model performance was evaluated on a test set of 10,000 tokens from the DBBE Occurrences. As illustrated in Example 1a, the Occurrences provide diplomatic transcriptions of manuscript texts, preserving their irregular orthography and other peculiarities, making them a suitable benchmark for unedited Greek.

Token	Lemma	Postag
ταῦτα	οὗτος	p-p---na-
δέ	δέ	d--------
εἶπε	λέγω	v3saia---
πρός	πρός	r--------
ἀπάτην	ἀπάτη	n-s---fa-
.	.	u--------

Table 1: An example sentence from the Pedalion Treebanks as we extracted them with the part-of-speech tags (including morphological features) and lemma.

4 Methodology

Building on previous attempts to achieve competitive results for both automatic morphological analysis and lemmatisation, we conducted two categories of experiments: single-task and multi-task learning. The single-task experiments differ from previous approaches in that they neither predict all features as a single label, nor treat each feature separately. Initially, all nine morphological features were combined into a single label as illustrated in Table 1 (Swaelens et al., 2024a), resulting in a set of 1,057 labels. This number proved too large to achieve competitive results. Fine-tuning each feature slot separately yielded strong results for part-of-speech tagging, but failed for other features. A cascaded model, where features were predicted based on their part-of-speech, also did not perform adequately (Swaelens et al., 2024a). Therefore, we present follow-up single-task experiments, in which we aimed to reduce the number of labels, while in multi-task experiments, we sought to leverage interactions between features to improve predictions.

4.1 Single-task training: Combination of Morphological Features

To reduce the number of possible labels, we grouped features into two categories: **nominal features** (case, gender, and number), as these features define nouns, adjectives, articles, and pronouns. The **verbal features** (voice, mood, and tense) are chosen as these are the only morphological features that are shared by all verbal forms.

Other verbal features – such as person and number for finite verb forms – were excluded to prevent an explosion (x12) in label combinations. Participles are rather peculiar, as they share both nominal and verbal features, which are thus both assigned. If a feature was not applicable to a given token, it

Token	Lemma	Nominal	Verbal
ταῦτα	οὗτος	pna	---
δέ	δέ	---	---
εἶπε	λέγω	s--	aia
πρός	πρός	---	---
ἀπάτην	ἀπάτη	sfa	---
.	.	---	---

Table 2: The example sentence of Table 1 with the new labels from the reduced label set.

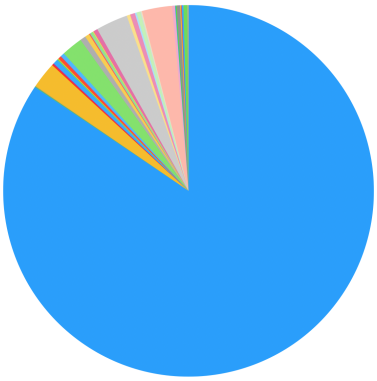

Figure 1: Partition of the verbal labels, with the most occurring label (84%) being the null label --- (in blue).

was assigned the null label (---). Table 2 shows the same example sentence as Table 1 with these new labels.

For instance, the verb form εἶδεν (3rd person singular active indicative aorist of ὁράω 'to see') received a null label --- for nominal features and *aia* for verbal features. Conversely, the participle ἴδων (active participle aorist nominative masculine singular of the same verb ὁράω 'to see'), received the nominal label *smn* and the verbal label *apa*.

We fine-tuned the pre-trained DBBErt model separately for nominal and verbal features using the following conditions: a learning rate of 2e-5, 15 training epochs, and a maximum sequence length of 512 subwords. For evaluation, the performance of both the nominal and the verbal classifier was measured using accuracy and weighted F-score to account for label imbalance, which is evident from Figure 1.

4.2 Multi-task fine-tuning

Unlike single-task models, which employ a single transformation layer atop the transformer model, multi-task models benefit from parameter sharing and transfer learning by training on multiple related tasks simultaneously. This concept is illustrated by Figure 2. A multitude of feature combinations could be jointly trained. However, considering computational costs and sustainability concerns, we decided to conduct the following experiments as we expected them to yield the best results.

Part-of-Speech & Lemma Driven by domain expertise, we jointly fine-tuned DBBErt for part-of-speech tagging and lemmatisation, as knowing the part-of-speech might help disambiguate several possible lemmas and, vice versa, the lemma can help decide what the correct part-of-speech is. Training conditions remained the same as in the single-task experiments: a learning rate of 2e-5, 15 epochs, and a maximum sequence length of 512 subwords. The multi-task setup involved simultaneous loss computation and updates for both classification heads. The model is henceforth referred to as MT_pos_lemma.

Nominal & Verbal Features Qualitative analysis of the single-task experiments (see Section 4.1) showed that finite verbs were tagged with the label for singular *s*-- or plural *p*--, while participles received both a verbal and a nominal feature. The two tasks are thus not completely independent, resulting in a second experiment with multi-task training of DBBErt where the nominal features are jointly fine-tuned with the verbal features. This model is henceforth referred to as MT_verb_nom.

Part-of-Speech & Nominal The performance of taggers assigning the nominal features lags behind, compared to the ones assigning the verbal features. Our analysis of both the single- and multi-task nominal taggers revealed that the nominal tagger does not consistently assign nominal features to nouns, while the verbal tagger does assign verbal features to verbs. However, we observed that the nominal tagger occasionally assigns nominal labels to verbs, and the verbal tagger sometimes assigns verbal labels to nominals. To mitigate these errors, we trained a multi-task model on both verbal and nominal tagging, aiming to improve feature assignment consistency. This model is henceforth referred to as MT_pos_nom.

5 Results

This section presents the quantitative and qualitative evaluation results for the lemmatisation (Section 5.1) and morphological analysis (Section 5.2) tasks.

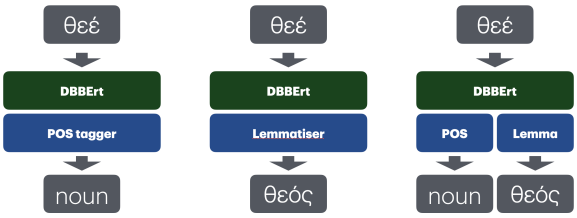

Figure 2: Conceptual visualisation of single-task fine-tuning our transformer model for POS-tagging and lemmatisation, and multi-task fine-tuning for both tasks simultaneously.

5.1 Lemmatisation

The multi-task training on part-of-speech tagging and lemmatisation boosted the performance of the lemmatizer compared to the state-of-the-art. Previous experiments in which a fine-tuned transformer model was combined with a dictionary-based approach yielded an accuracy score of 65.76%, whereas the best performing traditional lemmatizer has an accuracy score of 71.69%, both tested on our evaluation set (Swaelens et al., 2023). The current experiment displays an increase in tagging accuracy of 8 pp. regarding the state-of-the-art, as shown in Figure 3, and a slightly increase of 1 pp. for part-of-speech tagging. The following sections provide a qualitative analysis of the predicted output, which displayed some tendencies in the errors.

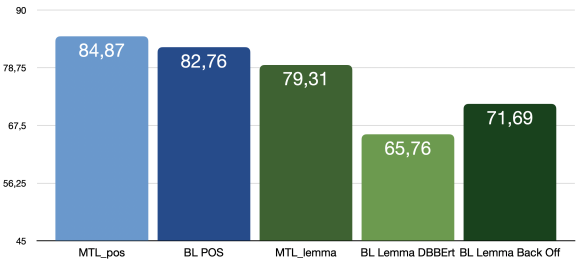

Figure 3: The results of MTL_pos_lemma for POS-tagging (MTL_pos) and lemmatisation (MTL_lemma) compared to state-of-the-art results for POS-tagging (BL POS) and lemmatisation with transformer models (BL Lemma DBBErt) and with traditional models (BL Lemma Back off).

5.1.1 Orthography

A recurring problem is the dialectal variation in Greek, resulting in different spellings of the same word. In Example 2, two words have a lemma that can be written in multiple ways: βίβλους (papyrus, book) and θαρρήσας (having confidence). The former could be lemmatised correctly as either βίβλος or βύβλος, whereas the latter has the lemmas θαρρέω and θαρσέω. The label of both tokens was erroneously considered as a mistake, because the spelling of the lemma differed from the gold standard. Automatic post-processing to make the spelling uniform would only be possible with a comprehensive list of all potentially affected words, which is not yet available.

5.1.2 Cognates

The system encounters difficulties in selecting the correct lemma when processing cognates – words that share a common etymological origin. The second verse of Example 2 has two words, each having one or multiple cognates, that were both tagged wrongly. The word ἄθλους is the accusative plural of ἆθλος (contest), but the system assigned the lemma ἄθλον (prize of contest), which is impossible because the accusative plural of ἄθλον is ἄθλα. The word ὁσίων, in this example a nominal use of the adjective ὅσιος (holy) but was tagged as the cognate ὁσία (divine law).

(2) πόθω γ(ὰρ) | προσφέρω σοι τ(ὰς) δέκα | βίβλ(ου)ς
pothō g(ar) | prospherō soi t(as) deka | bibl(ou)s
βί(ου)ς ἄθλ(ου)ς τε μαρ|τύρων (καὶ) ὁσίων:
bi(ou)s athl(ou)s te mar|tyrōn (kai) hosiōn
πρέ|σβις τούτ(ου)ς προβαλλόμ(εν)ο(ς) θαρρήσ(ας):
pre|sbis tout(ou)s proballom(en)o(s) tharrhès(as):
ὁ σῇ προνοίᾳ πρόεδρο(ς) | ἰωάννης
ho sè pronoia proedro(s) | iōannès

With a fond desire I offer you these ten books the lives and fights of martyrs and saints presenting these as messenger, entrusting your providence, I, bishop John.
DBBE Occurrence 17060

5.1.3 Subword tokenisation

For training a transformer-based language model, texts are first passed through a tokenizer. In the case of BERT, the tokenizer is WordPiece, which does not split the texts in words but in subword units (Wu et al., 2016). Our analysis showed that these subwords cause errors, where the system predicts words with a spelling "similar" to the correct lemma. Similar should be read in quite a broad

sense, like sharing the same prefix, a stem with the same consonants, sharing the same letters, etc., as displayed in the following list with predictions (left) vs. gold standard (right).

- εὐαγγελίζομαι (euangelizomai) vs. εὐαγγελιστής (euangelistès): same stem
- ἀκριβής (akribès) vs. ἠκριβωμένως (èkribōmenos): same stem
- οὕτος (houtos) vs. αὕτος (hautos): almost identical spelling
- ἐξαίρω (exairō) vs. ἐξαιρέω (exairō): different stem
- εὐθύς (euthys) vs. εὐχή (euchè): same initial letters

5.2 Morphological Features

The results for the morphological analysis presented in Table 3 and Table 4 are highly promising. Previous experiments did not assign labels with only nominal features or only verbal features, but used either a nine-slot label or a cascaded approach. This yielded accuracy scores of 62.33% and 58.29%, respectively. As a baseline, we will use the performance of the model fine-tuned on the nine-slot label, which differs only in the fact that it marks redundant features with a -, whereas our labels do not have these redundant slots. Compared to this baseline, labelling nominal features increased with almost 20 pp., whereas tagging verbal features increased with 32 pp. A slight increase of performance (less than 1 pp.) is observed in the multi-task models, which was also computationally more efficient. If we combine the output of the best-performing model for part-of-speech tagging, lemmatisation and morphological analysis, we get an accuracy score of 70.10% and a weighted F1-score of 70.48%, an increase of 8 pp. compared to the best-performing morphological tagger so far.

The single-task experiments yielded some unexpected results. Despite a larger label set (114 labels), predicting verbal features outperformed predicting nominal features (90 labels). This discrepancy likely stems from the distinctiveness of voice, mood, and tense, whereas nominal features often exhibit greater ambiguity. The following qualitative analysis provides further insights into the error patterns of the systems, which are roughly

Model	Accuracy	Weigh. F1
Single-task	0.8278	0.8206
MT_verb_nom	0.8339	0.9101
MT_pos_nom	**0.8409**	**0.8413**

Table 3: Performance of the all fine-tuned DBBErt models for classification of nominal features.

Model	Accuracy	Weigh. F1
Single-task	0.9478	0.9469
MT_verb_nom	**0.9540**	**0.9525**

Table 4: Performance of the all fine-tuned DBBErt models for classification of verbal features.

divided into nominal and verbal features. However, some overlap between the two sections was inevitable.

5.2.1 Nominal features

Ambiguity All models tend to default to the masculine gender when faced with ambiguous word forms – unsurprising given that 48% of the gender feature is masculine in our training set. Additionally, the single-task model shows a preference for the nominative case, particularly in neuter nominals, where nominative and accusative forms are morphologically identical. Errors are also frequent in the vocative case. In Greek, nouns ending in -ος, typically take a vocative form in -ε, while other paradigms simply retain the nominative case. Example 3 illustrates this with two vocatives: ῎Αναξ (leader, king), which is identical to its nominative form, and χριστέ, the vocative of χριστός (Christ). The single-task model incorrectly tagged ῎Αναξ as nominative, while χριστέ received the label *s--*, a tag intended for singular verbs – an anomaly that was not anticipated. However, the MTL_verb_nominal model did tag χριστέ correctly as a singular vocative masculine, suggesting that the joint training made the inclination to tag this as a verbal form less evident.

(3) ῎Αναξ ὑπάρχων χ(ριστ)έ | μου τῶν ἀνάκτων· (...)
Anax hyparchōn ch(rist)e | mou tōn anaktōn· (...)
teuxanta biblon | tèn nean theiō pothō
τεύξαντα βίβλον | τὴν νέαν θείῳ πόθῳ·
Christ, being the King of kings (...)
who made this new book with pious love
DBBE Occurrence 17017, vv. 1 & 4.

Iota subscriptum The DBBE Occurrences, and consequently our test set, differ from critical editions in certain scholarly conventions, such as the treatment of the *iota subscriptum*. If the long vowels ᾱ, η, or ω are followed by an iota (ι), the iota is typically written beneath (sub-scriptum) the vowel.[7] In DBBE, this iota is absent in almost all occurrences. The tokens most affected by this phenomenon are singular dative forms ending in -ῳ or -ῃ. The last two words of Example 3, θείῳ πόθῳ, are such datives, written without the iota. No model tagged these words correctly. The single-task model identified them as masculine genitive singular (as it did with most such datives), while the MTL_verb_nominal labelled θείῳ as masculine genitive singular, and πόθῳ as dual masculine nominative. The preference for a label with a dual number is rather unexpected, given that the dualis form makes up 0.2% of the number labels in our training set. The absence of *iota subscriptum* in female paradigms result in identical spelling for both nominative and dative forms. The last verse of Example 2 contains two female words in dative case, σῇ προνοίᾳ (with your providence). The noun was identified as a nominative, but its accompanying adjective did not receive a label at all.

Uninflected words Uninflected nominals are relatively rare in Greek. Most are names or titles of foreign origin, such as Δαυίδ (David) or Φαραώ (Pharaoh). In Example 4a φαραὼ functions as the object of the participle βυθίσας (to sink), which agrees with the subject μωσης (Moses). However, the single-task system tagged φαραὼ as nominative, just like μωσης. The multi-task system did not assign any label.

(4) a. ἄρδην βυθίσας φαραὼ μωσης λέγει
ardèn bythisas pharaō mōsès legei
After he completely sunk down the pharaoh, Moses spoke
DBBE Occurrence 19263

b. δα(υὶ)δ προφήτ(ου) καὶ βασιλέ(ως) μέλος
da(ui)d profètou kai basile(ōs) melos
Song of David, prophet and king.
DBBE Occurrence 19826

None of the taggers assigned the correct morphological features to the Greek token for David. Unlike *pharaoh*, *David* never received a case label but was consistently tagged with *s-m*, masculine singular. Example 4b is indeed masculine singular, but in the genitive case as it modifies μέλος (song). Initially, we assumed this was due to its indeclinability, but *pharaoh* serves as a counterexample, suggesting that this is not the – sole – cause.

[7] Some critical editions instead write the iota next to the vowel, in which case it is termed *iota adscriptum*.

5.2.2 Verbal features

Most errors in the verbal taggers stem from assigning verbal labels to nouns. The minimal performance difference between the single-task verbal tagger and MTL_verb_nominal is due to a lower number of nouns being erroneously labelled with verbal features. This indicates that verbs received the same labels in both models. The characteristics of the error analysis for the verbal taggers are closely related to those discussed in Section 5.2.1.

Plural genitives All Greek plural genitives end in -ων, which leads all our models to confuse nouns and verbs. Many such words were tagged by the verbal taggers as present participles in a plural genitive, despite being nouns or adjectives. Some examples in the test set are δενδρέων (tree), εὐαγγελιστῶν (evangelist), and εὐκλεῶν (famous). However, the opposite also occurs: verbs like χαλινῶν (participle of χαλινόω, to bridle) were assigned the null label by the verbal tagger, while the nominal tagger identified them as neuter genitive plural.

Iota The absence of the *iota subscriptum* leads to tagging errors in in both the verbal and nominal paradigms. Masculine datives lacking an *iota subscriptum* are frequently misidentified as present indicative forms, which is understandable given that the first-person singular always ends in -ω. A particularly striking error is found in the case of πόθῳ in Example 3. This token was tagged as an active subjunctive aorist, which is unexpected, as aorists are typically marked by a sigma following the stem.

Vocatives Vocatives ending in -ε are frequently confused with imperatives, as illustrated in Example 5. The vocative χριστέ in Example 3 is tagged as an active imperative aorist. This is surprising, as the active imperative aorist ends in -ον whereas the imperative present ends in -ε.

(5) βροντῆς γόνε βρόντησου ὑψόθ(εν) μέγα:
brontès gone brontèsou hypsoth(en) mega
καὶ σήμανον πῶς ἡ προάρχιος φύσις:
kai sèmanon pōs hè proarchios physis
θ(εὸ)ς βροτός τε καὶ θ(εὸ)ς πάλιν μένει:-
th(eo)s brotos te kai th(eo)s palin menei:-

Son of the Thunderer, thunder loud from above
and illustrate how the infinite nature
God, stays mortal and again divine.
DBBE Occurrence 26341

Our qualitative analysis highlights the potential of multi-task learning for improving linguistic annotation in non-standardised corpora, paving the way for future developments within the field of ancient language processing.

6 Conclusion

This study examined the potential of multi-task learning for the linguistic annotation of unedited Greek texts, focusing on part-of-speech tagging, lemmatisation, and morphological analysis. By introducing a novel approach that groups nominal and verbal features into combined labels, and fine-tuning a transformer-based language model within both single- and multi-task frameworks, we demonstrated significant improvements in tagging accuracy. Our results suggest that multi-task learning enhances feature prediction, particularly in challenging cases involving dialectal and orthographic variation.

Despite these advances, challenges remain present. While verbal feature tagging achieved a very low error rate, nominal feature tagging still exhibits inconsistencies, particularly in handling ambiguous case forms and uninflected words. Additionally, lemmatisation errors persist due to orthographic variation and closely related cognates. Addressing these challenges may require further label design, incorporation of external lexical resources, or integrating post-processing techniques.

Future research should explore extending this approach by integration of character-level models or phonological representations, which may improve the system's ability to process orthographic inconsistencies more effectively. As digital access to non-standardised Greek texts expands, improving automatic annotation techniques will be crucial for facilitating linguistic analysis and broader applications in computational philology. Furthermore, we will evaluate this approach to Latin book epigrams, to test whether this methodology is transferable across (ancient) languages.

Limitations

The lack of punctuation in the DBBE Occurrences makes it hard to perform syntactic analysis. This, however, would be an interesting factor to investigate or use in multi-task training, as the morphological features are very tightly connected to the syntactic structure. We furthermore did not fine-tune all possible combinations for multi-task training, keeping in mind its computational and environmental costs. The decisions on which combinations would be used in the multi-task experiments, were based on domain expertise. Finally, we have to keep in mind that almost all training data for ancient Greek is extracted from critical editions, and is thus standardised. This means that we have to keep annotating non-standardised Greek texts, so that its partition in the data becomes large enough to influence the system's performance.

References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Giuseppe G. A. Celano. 2024a. Opera graeca adnotata: Building a 34m+ token multilayer corpus for ancient greek. *Preprint*, arXiv:2404.00739.

Giuseppe G. A. Celano. 2024b. A state-of-the-art morphosyntactic parser and lemmatizer for ancient greek. *Preprint*, arXiv:2410.12055.

Gregory Crane. 1991. Generating and parsing classical greek. *Literary and Linguistic Computing*, 6(4):243–245.

Mark Depauw, Tom Gheldof, L Bolikowski, V Casarosa, P Goodale, N Houssos, P Manghi, and J Schirrwagen. 2014. Trismegistos. an interdisciplinary platform for ancient world texts and related information.

Vanessa B. Gorman. 2020. Dependency treebanks of ancient greek prose. *Journal of Open Humanities Data*.

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.

Alek Keersmaekers. 2021. The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 39–50, Online. Association for Computational Linguistics.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings*

of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), pages 109–117, Paris, France. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

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.

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.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

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

Colin Swaelens, Maxime Deforche, Guy De Tré, Ilse De Vos, and Els Lefever. 2024a. How relevant is part-of-speech information to compute similarity between greek verses in a graph database? In *Proceedings of the first workshop on Data-driven Approaches to Ancient Languages (DAAL 2024)*, pages 33–43. Language & Translation Technology Team (LT3).

Colin Swaelens, Pranaydeep Singh, Ilse de Vos, and Els Lefever. 2024b. Lemmatisation of medieval Greek: Against the limits of transformer‘s capabilities? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10293–10302, Torino, Italia. ELRA and ICCL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2027. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Yu Zhang and Qiang Yang. 2017. An overview of multi-task learning. *National Science Review*, 5(1):30–43.