Exploring Large Language Models for Classical Philology

Frederick Riemenschneider
Dept. of Computational Linguistics
Heidelberg University
69120 Heidelberg
riemenschneider@cl.uni-heidelberg.de

Anette Frank
Dept. of Computational Linguistics
Heidelberg University
69120 Heidelberg
frank@cl.uni-heidelberg.de

Abstract

Recent advances in NLP have led to the creation of powerful language models for many languages including Ancient Greek and Latin. While prior work on Classical languages unanimously uses BERT, in this work we create four language models for Ancient Greek that vary along two dimensions to study their versatility for tasks of interest for Classical languages: we explore (i) encoder-only and encoder-decoder architectures using RoBERTa and T5 as strong model types, and create for each of them (ii) a monolingual Ancient Greek and a multilingual instance that includes Latin and English. We evaluate all models on morphological and syntactic tasks, including lemmatization, which demonstrates the added value of T5's decoding abilities. We further define two probing tasks to investigate the knowledge acquired by models pre-trained on Classical texts. Our experiments provide the first benchmarking analysis of existing models of Ancient Greek. Results show that our models provide significant improvements over the SoTA. The systematic analysis of model types can inform future research in designing language models for Classical languages, including the development of novel generative tasks. We make all our models available as community resources, along with a large curated pre-training corpus for Ancient Greek, to support the creation of a larger, comparable model zoo for Classical Philology. Our models and resources are available at https://github.com/Heidelberg-NLP/ancient-language-models.

1 Introduction

Since the beginning of the creation of the *Index Thomisticus* in 1946 (Busa, 1980) and the publication of the Concordance to Livy (Packard, 1968), Classical Philology has been revitalized by the "digital revolution" (Berti, 2019). Today, numerous efforts have been undertaken to make Classical texts digitally available, annotate, and automatically process them. E.g., the Classical Language Toolkit (CLTK, Johnson et al., 2021) offers various tools to process pre-modern languages, in particular Latin and pre-modern Greek.[1]

Recently, we see a surge of the first pre-trained contextualized language models (PLMs) for Classical languages: Latin BERT has been proposed by Bamman and Burns (2020), Ancient Greek (AG) BERT by Singh et al. (2021). Lately, a second AG BERT has been proposed by Yamshchikov et al. (2022). However, both AG BERT models have been pre-trained on a comparatively small pre-training dataset. Moreover, they have been initialized from Modern Greek BERT (Koutsikakis et al., 2020), which limits them to the modern Greek alphabet, ignoring the diacritics of Ancient Greek.

Although numerous richly annotated treebanks are available for Latin and AG, systems have, by now, not been evaluated on a shared benchmark. Given that two popular treebanks for AG have been integrated into Universal Dependencies (de Marneffe et al., 2021), it is surprising that researchers working on AG do not compare to benchmarking results of, e.g., Straka (2018). Hence, a thorough assessment of the performance of the existing models is necessary in order to compare and evaluate their effectiveness for this underexplored language.

While BERT models are known to achieve high performance on a wide range of tasks, encoder-decoder models or multilingual models may often be a better choice, depending on the task. In this work, we explore a variety of language models for Classics in general and Ancient Greek in particular: We introduce GrεTa, GrεBERTa, PhilBERTa, and PhilTa, four PLMs for Classics. GrεBERTa and GrεTa are RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) models trained on Ancient Greek texts, respectively. PhilBERTa and

[1] We use the term "pre-modern" and "Ancient" interchangeably following the convention of calling every pre-modern language stage "Ancient". This is in line with, e.g. Singh et al. (2021), Yamshchikov et al. (2022), and the ISO code standard.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15181–15199

PHILTA are their trilingual counterparts pre-trained on Greek as well as Latin and English data.

We explore the advantages of (i) the two model architectures in (ii) mono- and multilingual pre–training for the mid-resource language Ancient Greek on a variety of morphological, syntactic, and semantic tasks, helping to answer questions, such as: *When to choose one architecture over the other?* or: *How does multilinguality affect a language model?*

Moreover, we publish the first wide-ranging benchmark results to compare our models for AG and Latin to the relevant prior work, establishing new SoTA results for both languages.

In summary, we aim to unify and push forward the current research landscape at the intersection of Classics and NLP with the following contributions:

(i) We introduce four pre-trained language models for Classics: GRε(BERT|T)A and PHIL(BERT|T)A. To our knowledge, we are the first to develop encoder-decoder models for Classics, and multilingual models tailored to both Latin and Greek.

(ii) We evaluate the already existing and our proposed models on several tasks, making many of them comparable for the first time. Furthermore, we outperform the existing Ancient Greek BERT models by a notable margin.

(iii) Our evaluation sheds light on the differences between encoders like ROBERTA and encoders of encoder-decoder models like T5 as well as on the influence of multilinguality on the mid-resource language Ancient Greek. By offering novel model types for AG, we aim to inspire new research and application tasks.

(iv) We develop and publish a large-scale, high-quality pre-training corpus for AG as a contribution to the community.

2 Related Work

Pre-training Data for Ancient Greek. Pre-trained language models require large amounts of unlabeled pre-training data. Ancient Greek and Latin being historical languages, the number of available texts is inherently limited, which makes the creation of a high-quality pre-training corpus even more important. To circumvent this problem, Singh et al. (2021) and Yamshchikov et al. (2022) pre-trained their AG BERT model from a Modern Greek BERT (Koutsikakis et al., 2020). But this approach has two weaknesses: First, there is an important cultural gap between modern and ancient texts that we do not want to introduce into our models. A Modern Greek BERT is familiar with contemporary concepts like cell phones or communism, which are unknown to antiquity, while we intend to use PLMs as a "window" to ancient cultures. Also the style of modern internet documents is fundamentally different from the transmitted ancient texts. Second, and more importantly, continuing pre-training of the Modern Greek BERT prevents us from adapting its tokenizer. AG, however, uses more diacritics, which host important information. By contrast, in our work, we build a tokenizer from scratch that is optimized for Ancient Greek.

In order to boost the data needed to train "pure" models of Ancient Greek, we put special effort into the curation of a large, but high-quality pre-training corpus for AG, leveraging previously unused textual sources. Finally, we evaluate the effect of using additional multilingual pre-training data.

Evaluating Models for Ancient Languages. Morphological and syntactic tasks, such as PoS tagging, dependency parsing, and lemmatization, have always been of interest to researchers of Latin and Ancient Greek. The standard tool for AG morphological analysis is Morpheus (Crane, 1991), a rule-based system, that has also been integrated into many more recent approaches. PoS Tagging has also been performed by various language-agnostic systems trained on AG data (Celano et al., 2016), but their success depends heavily on the chosen dataset: a winning system on one dataset (Celano et al., 2016) achieves the worst results on another (Keersmaekers, 2019). More recently, the CLTK (Johnson et al., 2021) provides a variety of taggers for many tasks. Surprisingly, although numerous richly annotated treebanks are available, systems have, by now, not been evaluated on a common benchmark.[2] E.g., Singh et al. (2021) test their proposed AG BERT on random splits from three popular treebanks, which we cannot compare against. The second AG BERT (Yamshchikov et al., 2022) has only been evaluated on authorship attribution.

As for lemmatization, Vatri and McGillivray (2020) provide an evaluation of three different lemmatizers. However, one of the evaluated candidates was partly trained on test data, which may have influenced its performance. It is noteworthy that,

[2]Cf. also Johnson et al. (2021): "The CLTK lacks formal evaluations of its models' accuracies. [...] Unfortunately, [outside] benchmarks do not yet exist for pre-modern languages."

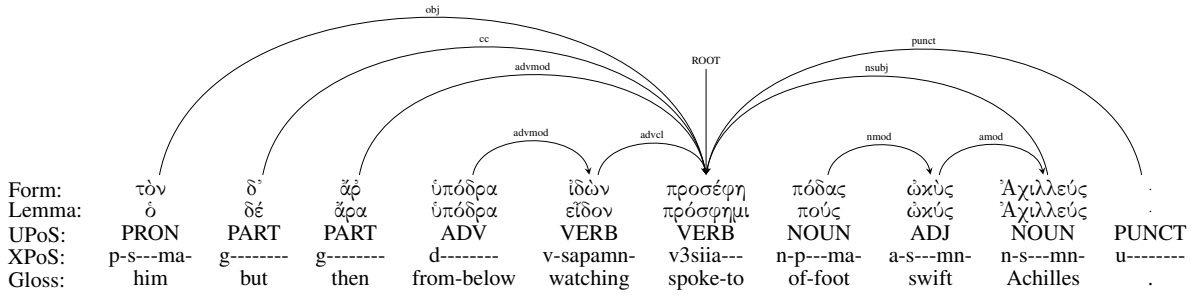

Figure 1: Example sentence (Hom. *Il.* 1.148) with corresponding dependency, lemma, UPoS, and XPoS labels. Translation: "Then watching him grimly from under his brows, swift-footed Achilles spoke to him."

despite the integration of two popular treebanks for AG into Universal Dependencies (UD, de Marneffe et al., 2021), many groups working on AG systems have not compared their models against the results of models benchmarked on UD data, such as Straka (2018). We remedy these issues by evaluating our systems and existing AG BERT models on the two authoritative treebanks covered by UD. The tasks we consider – dependency parsing, lemmatization, coarse, universal (UPoS) PoS tagging and fine-grained, language-specific (XPoS) tagging – are visualized in Figure 1.

For Latin, the issue does not arise thanks to the EvaLatin 2022 campaign (Sprugnoli et al., 2022), which has enabled direct comparison of models and has engendered strong models for Latin. Yet, despite the impressive results achieved in EvaLatin, our trilingual models outperform the existing systems on PoS tagging and lemmatization.

Language Model Architectures. Language models can be categorized into three classes: encoder-only, decoder-only, and encoder-decoder models. Encoder-only models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are best suited for tasks that aim to analyze complete sequences by sequence or token classification. Encoder-decoder models, on the other hand, are typically employed for conditional generation tasks, such as machine translation. Currently, all three models for ancient languages are BERT and thus encoder-only architectures.

We argue that an encoder-decoder model, such as T5 (Raffel et al., 2020), is a useful addition to this encoder-only landscape. First, it enlarges the space of possible NLP tasks for AG, enabling us, e.g., to cast lemmatization as a sequence-to-sequence task and to explore machine translation for ancient languages. Second, it allows us to compare the encoder of an encoder-only model with that of an encoder-decoder architecture, as they are both trained on the same data with a similar pre-training objective. Finally, commonly used multilingual encoder-decoder models like mT5 (Xue et al., 2021) and ByT5 (Xue et al., 2022) are not pre-trained on Ancient Greek texts.

As we aim for optimally trained encoder-only models, we chose RoBERTa over BERT: its dynamic masking strategy exploits the pre-training data better, and it has been shown that BERT's NSP objective can be detrimental (Liu et al., 2019).

3 Pre-trained Language Models for Ancient Greek and Latin

3.1 GrεBERT(a) and PhilBERT(a)

GrεBERTa and PhilBERTa are RoBERTa$_{base}$-, GrεTa and PhilTa are T5$_{base}$-sized models. Both models are pre-trained using a masked language modeling (MLM) objective. Specifically, in the case of RoBERTa, wordpieces are masked during the pre-training process, while for T5, spans are masked. Although it has been shown that multilingual pre-training can lead to gains for low-resource languages through cross-lingual transfer, it remains an open question when exactly it is preferable to use a multilingual instead of a monolingual model (Doddapaneni et al., 2021). To explore the implications of multilinguality for AG language models, we test different capabilities and possible interferences by comparing the different model types.

3.2 PLM Fine-tuning for Downstream Tasks[3]

PoS Tagging. PoS tagging for Ancient Greek typically aims for a complete morphological analysis:

[3]The following descriptions remain neutral to different PLM types by referring to basic transformer components. Where necessary, we will distinguish specific PLM types.

Next to the word class, the model has to predict eight fine-grained morphological attributes.[4] We frame this sequence labeling task as a multi-task classification problem applied to each token, with nine different classification heads per token on top of one shared encoder: We denote a sequence of tokens S of length n as $S = w_1, w_2, \ldots, w_n$ and refer to the contextualized embedding of each token as $\mathbf{e}_i = \text{Encoder}(w_{1:n}, i)$. As Byte Pair Encoding (Sennrich et al., 2016) splits words into subword units, we represent each token using its first subword embedding in the encoder. Each of the nine attributes is predicted using a feed-forward network applied to the last encoding layer, followed by a softmax function. The total loss is calculated as:

$$\mathcal{L}_{\text{total}} = \sum_{m=0}^{8} \frac{1}{9} \mathcal{L}_m$$

We use this approach for the Perseus XPoS dataset. For the other, less-detailed tagsets, we employ a single classification head.

Dependency Parsing. We follow Zhang et al. (2017) who cast dependency parsing as head selection. The model predicts a unique head for each token considered as a dependent. Since the model makes independent predictions, the obtained dependency graph can (in a few cases) be unconnected and is then completed by the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) for building non-projective trees – given that AG allows free word order. While Zhang et al.'s (2017) DENSE parser was based on a bi-directional LSTM, we define the model on top of the final hidden states of the transformer encoders.

Following Zhang et al. (2017), we add an artificial ROOT token w_0 and calculate the probability of the word $w_j \in \{w_0, w_1, \ldots, w_N\}$ being the head of the word $w_i \in \{w_1, w_2, \ldots, w_n\}$ in S as:

$$p_{\text{head}}(w_j | w_i, S) = \frac{\exp(f(\mathbf{e}_j, \mathbf{e}_i))}{\sum_{k=0}^{N} \exp(f(\mathbf{e}_k, \mathbf{e}_i))}$$

where f predicts the score of an edge (w_j, w_i) as follows:

$$f(\mathbf{e}_j, \mathbf{e}_i) = \mathbf{v}^\top \cdot \tanh(\mathbf{U} \cdot \mathbf{e}_j + \mathbf{W} \cdot \mathbf{e}_i)$$

Here, $\mathbf{v}$ is a weight vector and $\mathbf{U}$, $\mathbf{W}$ weight matrices. Dependency labels are predicted in a similar fashion: Let g be a single hidden-layer rectifier network that takes as input the concatenation $[\mathbf{e}_i; \mathbf{e}_j]$. The probability for the label l is then computed as:

$$p_{\text{label}}(l | w_j, w_i, S) = \frac{\exp(g(\mathbf{e}_j, l, \mathbf{e}_i))}{\sum_{l' \in L} \exp(g(\mathbf{e}_j, l', \mathbf{e}_i))}$$

While Zhang et al. (2017) use the representations of their trained DENSE model as input for the label classifier, we resort to the pre-trained embeddings.

Lemmatization. Current systems for lemmatization of AG, such as UDPIPE (Straka, 2018) or GLEM (Bary et al., 2017), are rule-based or use a classifier to predict editing rules that modify a token's pre- and suffixes. However, these complex scripts are not well-suited for a language like AG, which has many irregular forms that involve modifications of the word stem. An alternative approach is to utilize an encoder-decoder model that receives the inflected form, the PoS tag, and (optionally) additional information such as morphological features, as demonstrated for different languages by Schmid (2019) or Wróbel and Nowak (2022).

Yet, these earlier encoder-decoder-based lemmatization models are purely word-based and rely on pre-computed PoS tags or morphological features in a pipeline setting. By contrast, we propose a novel T5-based lemmatization model that is (i) *contextualized*, so that relevant morphological indicators can be inferred by the model on the fly from the token's surrounding context. (ii) The model works *end-to-end*: it receives the surface form of the word to be lemmatized in its full sentence context and predicts its lemma without receiving or predicting PoS tags or morphological features.[5] We mark the t(arget) token to be lemmatized in its context using delimiter tokens <t_tok_beg> and <t_tok_end>. For instance, for the input sentence ξύνοιδα <t_tok_beg> ἐμαυτῷ <t_tok_end> οὐδὲν ἐπισταμένῳ with the marked inflected t(arget) token ἐμαυτῷ, we expect as output the lemma ἐμαυτοῦ. We also experiment with providing, in addition, the target word as a sequence of individual characters, delimited by an additional separator token <t_tok_sep>: ξύνοιδα <t_tok_beg> ἐμαυτῷ <t_tok_sep> ἐ μ α υ τ ῷ <t_tok_end> οὐδὲν ἐπισταμένῳ.

Semantic and World Knowledge Probing Tasks. So far, we considered only morphological and syn-

[4] Person, number, tense, mood, voice, gender, case, and degree. Usually, many of these attributes are left empty. E.g., only adjectives have the attribute "degree".

[5] However, multi-task learning for joint morphological analysis *and* lemmatization is an interesting option that we did not pursue here.

tactic tasks. However, to evaluate the models more comprehensively, it is necessary to also test their semantic and world knowledge. Since such benchmarks do not exist for AG or Latin, we create two small datasets to evaluate these aspects. Inspired by Talmor et al. (2020), we test whether the language models can **distinguish synonyms from antonyms**. For this task, we input a sentence, e.g., τὸ χρήσιμον καὶ τὸ ἀγαθόν: <mask> ὁμοῖά ἐστιν ("the useful and the good: they are <mask> similar"), and the model has to predict either οὐχ ("not") or πάντως ("very"). Talmor et al. (2020) cast a similar task for English as a zero-shot MLM prediction problem using BERT and RoBERTa. However, with our prompt, the models always predict οὐχ ("not"), regardless of the provided word pairs. Experiments with variations of the prompt have led to similar difficulties. Hence, we evaluate this task in a few-shot setting, fine-tuning the MLM-head on 10 to 50 shots of synonyms and antonyms each, to prepare them for the task.

Similarly, we construct a dataset of **family relationships** between (mythical) heroes and gods. Here, the model is given a phrase, such as Τηλέμαχος ὁ τοῦ <mask> παῖς ("Telemachus, son of <mask>"), and has to predict the correct entity (in this case, Odysseus). For this task, we test the models in a zero-shot setting. However, this task cannot be solved by most encoder-only models, as the masked names typically consist of more than a single wordpiece. Thus, for this task, we evaluate only GRεTA and PHILTA, which can predict full entity names. By comparing the mono- and multilingual variants, we assess the models' acquired world knowledge as well as potential effects that may be induced by multilingual training: Given that Greek and Roman mythology share many of these gods, yet by different names, the multilingual model may be able to acquire additional knowledge from the Latin pre-training data, to solve the task formulated in Ancient Greek. We describe both datasets in Appendix B.2.

3.3 Acquisition of Pre-training Data

Ancient Greek. To cover a wide range of dialects, topics, and time periods of Ancient Greek, we make use of four different data sources: (the Greek part of) the Open Greek & Latin Project,[6] the CLARIN corpus Greek Medieval Texts,[7] the Patrologia Graeca,[8] and the Internet Archive (IA).[9] While the first three sources contain born-digital textual data, the IA online library provides books in the public domain along with their OCR transcriptions.

However, we found the partition of texts labeled as Ancient Greek in the IA to be incomplete and noisy: only a small fraction of the books containing AG text was labeled as such, and only few of them were transcribed with OCR settings supporting Greek characters. We hence extracted a novel data partition from the IA that was then fully re-OCRed by the Internet Archive to ensure correct transcription. To select a large number of high-quality texts, we applied a complex retrieve and filter procedure, focusing not only on (i) text quality, but also on (ii) collecting purely Ancient Greek texts, avoiding inclusion of texts in different languages, such as Latin, English, or German that can co-occur in the same book, and on (iii) filtering duplicates.

Latin and English. Acquiring pre-training data for Latin was facilitated by the Corpus Corporum project,[10] a meta-repository of Latin corpora that offers a comprehensive representation of the Latin language. All this data was kindly offered to us.

For English, we collect pre-training data from various sources, aiming for texts that are related to antiquity, by being focused on the same topics that we find in ancient texts – as opposed to modern themes. To this end, we utilize English translations of Latin and Ancient Greek texts as pre-training data. Furthermore, we ensure that the amount of English data is of similar size as the ancient texts, to prevent the models from being overwhelmed by a large number of English texts.

Statistics of pre-training data in Table 1. More details on corpus creation and statistics in Appendix C.

3.4 Pre-training Process

Even though our filtering of the IA corpus resulted in high-quality texts, the corpus is necessarily noisier than the born-digital texts. We therefore start pre-training on the IA data, and continue with the born-digital texts. Our tokenizers and the multilingual variants are trained on the born-digital texts only. For further pre-training details, see Appendix A.

[6] https://opengreekandlatin.org/.
[7] https://inventory.clarin.gr/corpus/890.
[8] https://patristica.net/graeca/.
[9] https://archive.org/.
[10] https://www.mlat.uzh.ch/.

Language	Dataset	Number of Tokens
Ancient Greek	Open Greek & Latin	30.0 million
	Greek Medieval Texts	3.3 million
	Patrologia Graeca	28.5 million
	Internet Archive	123.3 million
	Overall	**185.1 million**
Latin	Corpus Corporum	**167.5 million**
English	Overall	**212.8 million**

Table 1: Statistics of the pre-training datasets. Only Open Greek & Latin is used by Singh et al. (2021) and Yamshchikov et al. (2022) for their AG BERT models. Token counts determined by UNIX command `wc -w`.

4 Experiments

We run the experiments outlined in Section 3.2 to provide insight into the performances achieved by different model types and in relation to prior SoTA.

4.1 Datasets

Ancient Greek. For the PoS tagging, dependency parsing, and lemmatization tasks, we evaluate the PLMs for AG on the data provided by the Perseus and the PROIEL datasets, which are both integrated into Universal Dependencies 2.10 (de Marneffe et al., 2021).

To probe our models for semantic and world knowledge (see Section 3.2), we use our newly constructed datasets, described in Appendix B.2.

Latin. For Latin, we resort to the treebank used in EvaLatin 2022 (Sprugnoli et al., 2022), which covers three tasks: PoS tagging, lemmatization, and feature identification. Since no data for dependency parsing is provided, we restrict our evaluation to PoS tagging and lemmatization. In EvaLatin, instead of constructing test data by drawing samples from the initial data set, the test data exhibits different degrees of distribution differences in relation to the training data. For each task, three test sets are provided: The *Classical* set belongs to the same genre and time period as the training data, but comes from an author not included in the training data. The *Cross-genre* data includes two works that belong to different genres, yet being written during roughly the same time period. The *Cross-time* test set is based on text written in the 15th century, which is significantly later than the texts of the training data.

In Table 2, we summarize the diverse tasks under consideration with their corresponding metrics, the used evaluation datasets, the model architectures, and the pre-trained language models that are applicable to the respective task. Further details, including dataset statistics, are provided in Appendix B.1.

4.2 Models and Baselines

Ancient Greek. To showcase the capabilities of a recent system tailored to AG, we report the results of the taggers provided by the Classical Language Toolkit (Johnson et al., 2021).[11] As a baseline, we use the currently best-performing system, UDPipe (Straka et al., 2019), a transformer-based multi-task architecture that utilizes multilingual BERT, trainable word embeddings, and character embeddings.[12] In addition, to directly assess the benefits of using our monolingual model, we replace this multilingual BERT with our GrεBERTa model.

For PoS tagging and dependency parsing, we further compare to both prior encoder models trained on AG texts. We use the PoS tagger and DENSE (Section 3.2) to evaluate both AG BERT models as well as our GrεBERTa and PhilBERTa models. We apply the same approach to GRETA's encoder (GRETA-ENC) to investigate its behavior.

For lemmatization, we compare the performance of CLTK and UDPipe with that of our full-fledged T5 models. To predict a lemma during inference, we use beam search with a width of 20.

Latin. For Latin, we report the results of both teams that participated in the EvaLatin 2022 competition: Team Kraków (Wróbel and Nowak, 2022) utilizes the XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2020) model for PoS tagging, team KU-Leuven (Mercelis and Keersmaekers, 2022) employs an ELECTRA model. For lemmatization, Wróbel and Nowak (2022) use ByT5$_{\text{small}}$ (Xue et al., 2022), a multilingual encoder-decoder model similar to mT5 (Xue et al., 2021) that operates on UTF-8 bytes instead of subwords. Mercelis and Keersmaekers (2022) implement a cascaded approach that resembles the Greek lemmatizer GLEM (Bary et al., 2017): If a rule-based lookup

[11] From the multiple taggers offered by the CLTK we choose the one that achieved best performance on the validation set. For the Perseus dataset, this is a TnT tagger (Brants, 2000), while for PROIEL, it is Stanza (Qi et al., 2020). Note, however, that it is not possible to directly compare the results to those of the other models, as they were trained on different data splits and using an older version of the dataset (cf. https://github.com/cltk/greek_treebank_perseus).

[12] We report scores of the most recent, unpublished version of UDPipe (https://ufal.mff.cuni.cz/udpipe/2/models#universal_dependencies_210_models) and the scores obtained when training UDPipe ourselves.

	PoS Tagging		Dependency Parsing		Lemmatization
	UPoS	XPoS	Unlabeled	Labeled	
Task Description	PoS tagging with universally applicable, coarse PoS tags	PoS tagging with language-specific, fine-grained tags; complete morphological analysis in the case of Perseus	predicting the head of each token in text	predicting the head and relation type of each token in text	predicting the lemma of each token in text
Metric	Accuracy	Accuracy	UAS	LAS	Accuracy
Datasets	Perseus ✓ PROIEL ✓ EvaLatin ✓	Perseus ✓ PROIEL ✓ EvaLatin ✗	Perseus ✓ PROIEL ✓ EvaLatin ✗	Perseus ✓ PROIEL ✓ EvaLatin ✗	Perseus ✓ PROIEL ✓ EvaLatin ✓
Model Architecture	Encoder + Classification Head	Encoder + Classification Head(s)	Encoder + DENSE	Encoder + DENSE	Encoder-decoder
PLM Instances	(GRɛ|PHIL)BERTA, (GRɛ|PHIL)TA-ENC	(GRɛ|PHIL)BERTA, (GRɛ|PHIL)TA-ENC	(GRɛ|PHIL)BERTA, (GRɛ|PHIL)TA-ENC	(GRɛ|PHIL)BERTA, (GRɛ|PHIL)TA-ENC	(GRɛ|PHIL)TA

Table 2: Summary of the tasks under consideration.

returns multiple lemmata, the system tries to disambiguate between these possibilities by means of the predicted PoS tag. To further clarify any remaining ambiguities, a classifier is trained to select the correct lemma from the available options.

5 Results

Ancient Greek. We present the results for **PoS tagging** and **dependency parsing** for Ancient Greek on the Perseus dataset in Table 3. The PROIEL dataset seems to be easier to solve, as all models achieve performances that are much closer to each other (see Appendix D). Since the overall trends are consistent across both datasets, we focus our discussion on the results on the Perseus dataset.

As seen in Table 3, the CLTK performs clearly below all other models on both tasks. While the CLTK is not directly comparable to the other models (see fn. 11), the evaluation still provides a perspective on the capabilities of the *de facto* only available framework for processing AG text.

UDPIPE provides a strong baseline, which AG BERT (Yamshchikov et al., 2022) is unable to consistently outperform. By contrast, all other PLMs show clear gains over UDPIPE. The monolingual, encoder-only GRɛBERTA model consistently performs best on all tasks. Interestingly, the performance of GRETA-ENC on PoS tagging is slightly worse than that of PHILBERTA, while it achieves better results for dependency parsing. This trend has also been observed in initial experiments. We elaborate on the behavior of GRɛTA-ENC and PHILBERTA in Section 6.

Results for **Lemmatization** are shown in Table 4. Here, augmenting UDPIPE with GRɛBERTA's pretrained embeddings does not lead to better scores. We attribute this to the tokenization process and refer to our discussion in Section 6. GRɛTA, on the other hand, demonstrates strong encoder-decoder capabilities and significantly outperforms UDPIPE. Providing GRɛTA with the individidual characters of the target word leads to a small gain.

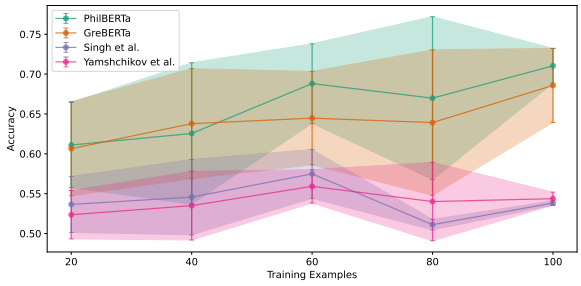

Figure 2: Synonym/antonym disambiguation accuracy for growing few-shot sizes: GRɛBERTA and PHILBERTA vs. AG BERT models. We use equal amounts of synonyms and antonyms (a run with 20 samples includes 10 synonyms and 10 antonyms). We use k-fold cross-validation. Error bars show standard deviation.

The results of the **Synonym/antonym disambiguation task** are visualized in Figure 2. Again, GRɛBERTA and PHILBERTA demonstrate higher scores compared to the AG BERT models. We observe the same for GRɛTA and PHILTA (cf. Figure 4 in Appendix D). Our monolingual models and their multilingual counterparts perform almost equally, especially taking into account the overlapping standard deviation bands. We see a minimal trend for PHILTA to gain over GRɛTA in Figure 4, but our small datasets do not allow drawing firm conclusions on their relative performance.

Finally, we report zero-shot results for the **Family relationship task** in Table 5. As the T5-based models have been pre-trained to predict multiple masked spans at once, they tend to predict, for each sample, more than a single entity. We interpret the output as a ranked list and report recall@k, evalu-

Model	PoS Tagging		Dependency Parsing	
	UPoS	XPoS	UAS	LAS
CLTK	68.83	47.21	59.21	43.24
UDPIPE (official)	92.88	85.60	80.32	74.53
UDPIPE (ours)	92.36 (0.09)	84.72 (0.06)	78.74 (0.04)	73.14 (0.06)
UDPIPE + GRεBERTA	95.74 (0.06)	90.95 (0.07)	86.30 (0.14)	82.15 (0.14)
AG BERT (Singh et al., 2021)	94.92 (0.18)	88.27 (0.27)	84.03 (0.12)	78.80 (0.37)
AG BERT (Yamshchikov et al., 2022)	92.50 (0.03)	84.56 (0.13)	80.34 (0.11)	74.22 (0.21)
GRεTA-ENC	94.44 (0.14)	89.03 (0.13)	87.32 (0.04)	83.06 (0.07)
PHILBERTA	95.60 (0.21)	90.41 (0.18)	86.99 (0.06)	82.69 (0.06)
GRεBERTA	**95.83** (**0.10**)	**91.09** (**0.02**)	**88.20** (**0.11**)	**83.98** (**0.21**)

Table 3: PoS tagging and dependency parsing results on the Ancient Greek Perseus dataset. The results are averaged over three runs with different random seeds, and the standard deviation is indicated in parentheses, except for the CLTK and UDPIPE (reported results). Note also that the CLTK is not trained on exactly the same data as the other models and therefore not strictly comparable.

Model	Accuracy
CLTK	76.10
UDPIPE (official)	86.70
UDPIPE (ours)	84.50 (0.09)
UDPIPE + GRεBERTA	85.56 (0.06)
PHILTA	90.02 (0.02)
PHILTA + Chars	90.66 (0.01)
GRεTA	90.80 (0.10)
GRεTA + Chars	**91.14** (**0.10**)

Table 4: Lemmatization results for Ancient Greek on the Perseus dataset. Results are averaged over three runs, with standard deviation in parentheses, except for the CLTK and UDPIPE (reported results).

Model	k = 1	k = 5	k = 10	k > 10
GRεTA	4.39	9.65	10.53	10.96
PHILTA	3.07	8.33	11.40	11.84

Table 5: Zero-shot family relationships task (recall@k).

ating whether the correct entity is contained in the first 1, 5, 10, and >10 predictions, restricting the maximum sequence length to 50 wordpieces.

Latin. The PoS tagging and lemmatization scores on EvaLatin 2022 are reported in Table 6. While the performance scores of all models are rather close to each other, our trilingual models consistently outperform the EvaLatin 2022 participant systems across all three subtasks. PHILBERTA reaches even higher scores than KRAKÓW-OPEN on PoS tagging, which leverages additional annotated data. On lemmatization, PHILTA similarly outperforms KRAKÓW-CLOSED on the Classical, Cross-genre, and Cross-time subtasks by 2.25, 1.78, and 0.23 percentage points, respectively, but does not outperform KRAKÓW-OPEN on the Cross-genre and the Cross-time subtask.

	Model	Classical	Cross-genre	Cross-time
UPoS	KRAKÓW-OPEN	97.99	96.06	92.97
	KRAKÓW-CLOSED	97.61	94.62	92.70
	KU-LEUVEN	96.33	92.31	92.11
	PHILBERTA	**98.23** (**0.06**)	**96.59** (**0.15**)	**93.25** (**0.12**)
Lemmatiz.	KRAKÓW-OPEN	97.26	96.45	92.15
	KRAKÓW-CLOSED	95.08	91.62	91.68
	KU-LEUVEN	85.44	86.48	84.60
	PHILTA + Chars	**97.33** (**0.04**)	**93.40** (**0.13**)	**91.91** (**0.04**)

Table 6: PoS tagging and lemmatization results (EvaLatin 2022 dataset). KRAKÓW-OPEN uses additional data.

6 Analyses and Discussion

Training Behavior of GRεTA-ENC. While GRETA-ENC and GRεBERTA are of similar size (Table 7) and pre-trained with comparable objectives, GRεTA-ENC performs slightly worse than GRεBERTA. One reason may be that in a T5 model, some important information is distributed across encoder and decoder. This raises the question of whether encoders in encoder-decoder models are trained suboptimally, and whether improvements may be obtained by combining separately pre-trained encoders and decoders, or by pre-training the encoder before adding the decoder. Another reason may be that the encoder is not accustomed to using its classification head. Here again, it may be advantageous to pre-train the encoder before extending it to encoder-decoder pre-training.

In Figure 3 we compare the PoS tagging validation accuracy of GRεTA-ENC to that of a randomly initialized T5 encoder (same size). GRεTA-ENC performs much worse than the randomly initialized model after one epoch, reaching only approximately 6%. However, while the randomly initialized model stagnates, GRεTA-ENC outperforms the randomly initialized model after two epochs, significantly improving its performance thereafter.

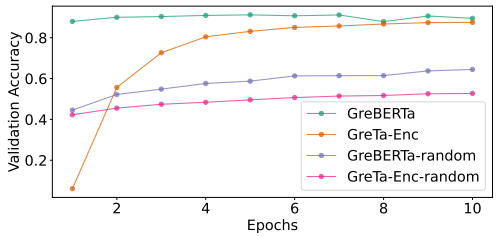

Figure 3: Validation accuracy (AG XPoS Tagging on Perseus) for GRETA-ENC and GREBERTA and randomly initialized counterparts over various training epochs.

By contrast, GREBERTA reaches a high validation accuracy already after one epoch. We see the same trend with different random seeds and for dependency parsing, but it is most apparent in Perseus XPoS tagging.

Lemmatization as a Character-based Task. As seen in Table 4, augmenting UDPIPE with GREBERTA does not lead to significant improvement for lemmatization. This we attribute to the tokenization process. GREBERTA uses wordpieces, which contain little information about individual characters. We hypothesize that UDPIPE ignores the GREBERTA embeddings for lemmatization and instead relies on its own additional character embeddings. Accordingly, explicitly providing GRETA with the individual characters of the inflected word form leads to a slight increase in performance.

This explanation can also shed light on the success of the *UTF-8 bytes-based* BYT5 model for lemmatization in Latin. This model was chosen by Wróbel and Nowak (2022), after initial experiments with the *wordpiece-based* MT5 (Xue et al., 2021) had underperformed. Future work on (AG) lemmatization could therefore investigate whether Byte Pair Encoding-based models can be augmented with character embeddings as additional input.

Effect of Multilinguality. Table 3 shows that PHILBERTA consistently performs slightly worse compared to monolingual GREBERTA on morphological and syntactic tasks. We attribute this to the *curse of multilinguality* (Conneau et al., 2020): the capacity of the trilingual models is split between three languages. Still, both models achieve strong results on AG and Latin tasks and can be especially useful in tasks that require multilingual knowledge, as in MT or glossing tasks. Our small-sized knowledge probing tasks show very similar performance for both model types. While the size of our data does not allow for firm conclusions, this is in line with Kassner et al. (2021), who find no improved knowledge representation in multilingual PLMs.

7 Conclusion

We introduce four strong language models for Classical Philology, including the first encoder-decoder PLMs for Ancient Greek and Latin. We rigorously benchmark our models and prior work on various tasks, demonstrating strong improvement over the SoTA. We showcase the versatility of encoder-decoder models, (i) by offering a novel end-to-end contextualized lemmatization model for AG and Latin, with a greatly simplified architecture that clearly outperforms prior work; (ii) while MLM in encoder-only models is restricted to single-token predictions, our T5-based models exhibit great flexibility for formulating probing tasks, which help exploring what models learn from pre-training data. Considering the two investigated model dimensions, our work (iii) sheds light on differences between the encoders of T5 vs. ROBERTA, where the former tends to exhibit slower learning curves; (iv) our monolingual models outperform the multilingual ones in monolingual morphological and syntactic tasks, without clear trends on small-scale semantic and knowledge probing tasks.

Limitations

While we aim for a comprehensive analysis of existing methods (such as lemmatization) and model types for Ancient Greek and other Classical languages, there are limits to exhaustively exploring the full space of variations and rigorously evaluating their impact on model performance. For example, we could not comprehensively evaluate the effects of (i) the pre-training corpora, as we did not re-train a BERT model for Ancient Greek, to pin down the exact difference between prior BERT models (which were trained on smaller data before) and our own models, which are based on inherently stronger model types; similarly, we did not induce Latin ROBERTA and T5 models, to confirm the differences between mono- and multilingual models for language-specific Latin tasks. (ii) In a similar vein, we did not compare different model sizes. However, we studied prior work and scaling laws and believe that the base model is appropriate for the size of our training data. Further factors of this type concern (iii) hyperparameter settings and (iv) other factors in isolation.

Not only do we miss sufficient computational

resources to perform such manifold ablations and comparative assessments, we also considered the carbon footprint that such experiments cause and which does not stand up to the insights that could possibly be gained from more experiments.

For these reasons, we focused on two selected dimensions of variants that we believe to be valuable for a community interested in Classical languages:

(i) We tried to answer questions as to when multilingual models can be profitably used, and (ii) aimed to showcase various potential advantages of encoder-decoder models, which by now have not been considered in studies on Classical languages.

Another clear limitation lies in the size of the demonstrated semantic and knowledge probing tasks. (i) They are of small size, and we cannot, therefore, draw firm conclusions as to, e.g., the effect of multilinguality. Also, the synonym/antonym disambiguation task is presumably the most difficult one. As a counter-balance, we used a more tangible task for knowledge probing, by choosing family relationships, which we expect to be frequently found in the pre-training corpora.

(ii) A further limitation we find for the knowledge probing tasks resides in the size of our trained models and the underlying pretraining data. This limitation could be one that is not easy to overcome. But we still encourage the community to create similar probing task datasets. Future work may find appropriate ways of data augmentation, or transfer learning methods that are applicable to historical languages so that further progress and insight will be possible.

Ethics Statement

It is a computationally demanding task to pre-train large language models. However, transfer learning opens the possibility to fine-tune our pre-trained models, which showed strong performances, in a reasonable amount of time.

The texts utilized for pre-training the models may well exhibit biases related to ancient perspectives of the world. We do not view this as an issue, as the proposed language models for historical languages are intended for academic use and do not have practical, everyday applications.

Acknowledgments

We are deeply indebted to the Internet Archive team for their continuous support by creating new OCR transcriptions of the misclassified Greek books, and to our anonymous reviewers for their comments, which have helped to significantly improve the paper. We thank Nina Stahl and Thomas Kuhn-Treichel for their help in creating our semantic and knowledge probing tasks, as well as Jan Ctibor and Philipp Roelli for providing us with the invaluable *Corpus Corporum* data. Finally, we acknowledge and thank for crucial support from the Google TPU Research Cloud program, for granting us access to their TPUs.

References

David Bamman and Patrick J Burns. 2020. Latin BERT: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

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

M. Berti. 2019. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*. Age of access? Grundfragen der Informationsgesellschaft. De Gruyter.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

R. Busa. 1980. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, 14(2):83–90.

Giuseppe G. A. Celano, Gregory Crane, and Saeed Majidi. 2016. Part of Speech Tagging for Ancient Greek. *Open Linguistics*, 2(1).

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *arXiv preprint arXiv:2107.00676*.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Alek Keersmaekers. 2019. 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.

Manfred Kern, Alfred Ebenbauer, and Silvia Krämer-Seifert. 2003. *Lexikon der antiken Gestalten in den deutschen Texten des Mittelalters*. Walter de Gruyter.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

D. W. Packard. 1968. *A Concordance to Livy*. A Concordance to Livy. Harvard University Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Helmut Schmid. 2019. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 133–137, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Jan Hajič. 2019. Evaluating contextualized embeddings on 54 languages in

pos tagging, lemmatization and dependency parsing. *arXiv preprint arXiv:1908.07448*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

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

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based part-of-speech tagging and lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in Plutarch's Shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 665–676, Valencia, Spain. Association for Computational Linguistics.

Hyperparameter	GRεBERTA	PHILBERTA	GRεTA	PHILTA
Adam ϵ	$1 \cdot 10^{-8}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-8}$	$1 \cdot 10^{-8}$
Adam β_1	0.9	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999	0.999
Attention Dropout	0.1	0.1	0.1	0.1
Attention Heads	12	12	12	12
Batch Size	128	256	512	512
d_{ff}	—	—	2048	2048
d_{kv}	—	—	64	64
d_{model}	—	—	768	768
Hidden Dropout	0.1	0.1	0.1	0.1
Hidden Size	768	768	—	—
Learning Rate (LR)	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
LR Scheduler	linear	linear	linear	linear
Nb. of Layers	12	12	$2 \cdot 12$	$2 \cdot 12$
Nb. of Parameters	126 mill.	135 mill.	223 mill.	247 mill.
Train Epochs	50, 100	0, 100	50, 100	0, 100
Warmup Steps	0	0	10000	10000
Weight Decay	0	0	0.01	0.01

Table 7: Pre-training hyperparameters.

A Training Details

A.1 Pre-training Details

We pre-train the monolingual models for 50 epochs on the Internet Archive corpus and continue pre-training for 100 epochs on the born-digital texts, the trilingual models were trained for 100 epochs on the born-digital texts. The tokenizers were trained on the born-digital data only. GRεBERTA and PHILBERTA were trained on an NVIDIA A100-PCIE-40GB, GRεTA and PHILTA on a Google TPU v2-8. Training took between 3 and 7 days. Further details in Table 7.

A.2 Fine-tuning Details

We train every Greek model for 50 epochs on an NVIDIA GeForce GTX 1080 Ti, evaluating the model after every epoch on the validation set and using early stopping with a stopping patience of 5. As the EvaLatin dataset does not provide a validation set, we use 2% of the training data as the validation set. Furthermore, since the EvaLatin dataset is larger than the Greek datasets, we set the maximum number of training epochs to 20 for the Latin models. Depending on the treebank and the task, training the models took approximately 1 hour (PoS tagging), 5–7 hours (dependency parsing), and 1–3 days (lemmatization). Further details in Table 8. We did not experiment with different hyperparameter settings, as our main goal was to provide comparable and wide-ranging benchmarking results.

Hyperparameter	
Adam ϵ	$1 \cdot 10^{-8}$
Adam β_1	0.9
Adam β_2	0.999
Batch Size	32
Early Stopping Patience	5
Learning Rate	$1 \cdot 10^{-4}$
Learning Rate Scheduler	linear
Random Seeds	42, 1, 2
Train Epochs	50
Weight Decay	$1 \cdot 10^{-5}$

Table 8: Fine-tuning hyperparameters.

	Perseus	PROIEL	EvaLatin
Sentences (train)	11 476	15 014	15 785
Sentences (dev)	1137	1019	—
Sentences (test)	1306	1047	1960
Sentences (total)	13 919	17 080	17 745
Tokens (train)	159 895	187 033	316 573
Tokens (dev)	22 135	13 652	—
Tokens (test)	20 959	13 314	45 544
Tokens (total)	202 989	213 999	362 117
Lemmata	13 413	9348	10 357
Forms	41 304	32 591	54 133
UPoS Tags	14	14	16
XPoS Tags	847	27	—
Dependency Relations	25	33	—

Table 9: Statistics of the Perseus, PROIEL, and EvaLatin datasets.

B Downstream Task Details

B.1 Universal Dependencies and EvaLatin 2022

For PoS tagging, UD provides universal PoS tags (UPoS) and language-specific PoS tags (XPoS). UPoS consists of 17 tags used for all languages covered by UD.[13] XPoS tags, on the other hand, can follow a dataset-specific annotation scheme. While the XPoS tags of the PROIEL dataset are similar to the UPoS tags, the Perseus dataset aims for a complete morphological analysis (cf. Section 3.2).

See Table 9 for further details and Table 2 for an overview. In line with common convention, we report the accuracy for both PoS tag sets. For dependency parsing, we report the unlabeled attachment score (UAS) and the labeled attachment score (LAS). The UAS indicates the percentage of tokens that have been assigned the correct head, whereas for the LAS, both the predicted head and the dependency label have to be correct. All results are obtained from the official evaluation script.[14]

[13] In the case of AG, 3 of these 17 tags are not used.

[14] https://universaldependencies.org/conll18/conll18_ud_eval.py and https://github.com/CIRCSE/LT4HALA/blob/master/2022/data_and_doc/conll18_ud_eval_EvaLatin_2022_rev2.py.

B.2 Semantic and World Knowledge

Semantic Knowledge. We asked a graduate student and a doctoral candidate in the field of Classics to gather synonym and antonym pairs. Such word pairs can be nouns and substantivized adjectives or substantivized infinitives. We then utilized a predefined template to generate sentences that incorporate the collected pairs. As this template does not ensure grammaticality, the annotators manually edited the sentences. Subsequently, the sentences were independently reviewed by both annotators, deduplicated, and then verified by a professor of Ancient Greek. All three annotators participated on a voluntary basis and were not compensated for their contributions. One of the annotators is also a co-author of this paper.

141 synonym and 146 antonym pairs were collected. While we publish all 287 examples, we drop 5 randomly selected antonym pairs in our experiments to ensure that the number of synonym and antonym pairs is equal. We train all language models for 10 epochs using a batch size of 4 and report the averaged, cross-validated results.

World Knowledge. This dataset was compiled by one of the previous annotators who is not a co-author of this paper. The annotator gathered 228 examples with 11 different relations by reading through Hesiod's *Theogony* and by drawing inspiration from Kern et al. (2003), a lexicon that contains comprehensive information about mythical figures.

C Acquisition of Pre-training Data

C.1 Ancient Greek Pre-training Data

Open Greek & Latin Project.[15] The Open Greek & Latin Project is an umbrella project covering various subprojects that aim toward the development of open-access corpus linguistic resources for Latin and Classical Greek. Two of them, the Perseus Digital Library and the First Thousand Years of Greek project, contain Ancient Greek texts, mostly covering texts that are typically associated with classical antiquity, such as Homer, Plato, Herodotus, Euripides, and Plutarch. Already in this corpus, we find a wide variety of dialects and language stages. The Open Greek & Latin Project contains approximately 30 million tokens.

Greek Medieval Texts.[16] The Greek Medieval Texts corpus offered by CLARIN covers writings from the fourth to the sixteenth century AD. It contains religious, poetical-literary and political-historical texts as well as hymns and epigrams. Strictly speaking (and as the name suggests) the corpus contains texts of late antiquity, and in particular, Medieval Greek. We argue, however, that Ancient Greek and Medieval Greek, although different language stages, are strongly connected to each other and that our language models benefit from seeing more diverse data during pre-training. This corpus contains about 3.3 million tokens and is licensed under the CC BY-NC 4.0 license.

Patrologia Graeca.[17] The Patrologia Graeca is a large collection of important Christian texts written in Greek, dating from the first until the fifteenth century AD. Since not all texts are machine-readable and available, we are restricted to those out of copyright texts that are made accessible (around 28.5 million tokens).

Internet Archive.[18] The Internet Archive is an online library that provides texts obtained from public domain books via OCR. We found out that only a small fraction of the books containing Ancient Greek text was labeled as such. Moreover, we discovered that even less books were transcribed with OCR settings that allowed Greek characters. As a result, many high-quality scans of Ancient Greek texts were transcribed into incomprehensible sequences of non-Greek characters. For example, the verse ὦ γύναι ἦ μάλα τοῦτο ἔπος νημερτὲς ἔειπες[19] is transcribed as `& yvvai, ff pdXa tovto Stto^ vrjpepTe^ e€/.7r€9*`.

Even though transcriptions of this nature may seem useless at first glance, they are nevertheless helpful in identifying documents that have been incorrectly treated as non-Greek texts, for many common words are relatively consistently transcribed. τοῦτο ("this"), for example, is often transcribed into `tovto`. By searching for all books that contain the word `tovto`, we can identify potential Greek texts. This approach allows us to avoid the computationally intensive task of applying Greek OCR to every book in the Internet Archive, and instead focus our efforts on a more targeted search. All candidates are then filtered more aggressively: If

[15] https://opengreekandlatin.org/.

[16] https://inventory.clarin.gr/corpus/890.

[17] http://patristica.net/graeca/.

[18] https://archive.org/.

[19] Hom. *Il.* 3.204.

a candidate contains the five (presumably) Greek stopwords tovto (τοῦτο), kal (καί), tov (τόν), to (τό), and yap (γάρ) more than ten times, the candidate is considered to contain Greek text.

We argue that this method effectively minimizes false positives while retaining a high recall: Since Greek stopwords like τοῦτο ("this") and καί ("and") should be present often enough in every book with a significant amount of text, our approach should correctly classify them as Greek. Non-Greek texts, on the other hand, should hardly contain all five stopwords more than ten times.

This procedure yields 25378 books, on which the Internet Archive applies OCR with Ancient Greek as a target language. While our method reliably detects Greek texts, it does not ensure a high scan (and therefore also text) quality. In order to use solely high-quality data, we keep only lines in which more than 90% of tokens are also present in the born-digital vocabulary. A similar approach is used by Bamman and Burns (2020), who use Latin texts from the Internet Archive as pre-training data for Latin BERT. They "retain only those books where at least 40% of tokens are present in a vocabulary derived from born-digital texts". We argue that it is more sensible to include or disregard individual lines instead of whole books: Almost every Greek text contains a Latin or English introduction, and many books are equipped with a translation. Thus, our method not only ensures high-quality data but also removes non-Greek text parts.

Finally, to ensure that our dataset does not contain any significant duplications, we remove all instances of repeated text exceeding 300 characters. After this aggressive filtering, we have approximately 123.3 million tokens left. To demonstrate its quality, we show 40 samples randomly drawn from the dataset in Table 10.

C.2 English Pre-training Data

By collecting English translations of ancient texts, we focus on texts that are strongly connected to antiquity. We gather these texts from various sources: The Perseus Digital Library[20] and the Internet Classics Archive[21] provide born-digital open-access translations of Classical Greek and Latin texts. Similarly, the Documenta Catholica Omnia database[22] contains a large amount of primarily catholic texts in many languages, of which we select the English partition for our use. Finally, we utilize Lexundria,[23] Loebulus,[24] and the Project Gutenberg to add (often proofread) scans of books in the public domain. While Lexundria and Loebulus are

τῆς Μασίστεω γυναικός, ἐούσης καὶ ταύτης ἐν-
θαῦτα. ὡς
πίστις ὑμῶν· φοβηθέντες δὲ ἐθαύμαζον, λέγοντες πρὸς ἀλλήλους,
ὑποληπτέον' ἡ γὰρ πέψις τοῖς μὲν ἐν τῷ ἄνθει μᾶλ-
λον
ἀνέπαυσαν γὰρ τ. ἐμὸν πνεῦμα κ. τὸ ὑμῶν
εἰ Σατανᾶς ἀνέστη ἐφ' ἑαυτόν
πρόσωπον ναοῦ Κυρίου, μετὰ τὸ ἀποικίσαι Ναβουχοδονόσορ
ἐκείνοις δὲ ὄντος ἀεὶ τοῦ ἐπιχειρεῖν καὶ ἐφ ἡμῖν εἶναι δεῖ τὸ προαμύνασθαι.
ἑξακοσίους ὁπλίτας εἰς τὴν πόλιν ἄ ἄγει. ἐν τῷ στρατεύματι
ἔχοντι τοῦ Γερμανικοῦ συναγορεύειν μέλλοντος,'
νοοῦν εἴη, ὅτι ἄλλου τὴν νόησιν λαμβάνον οὐ τὸ
ἐὰν δὲ μὴ τούτοις δύνῃ χρῆσθαι,
μου- ἐφ' ὑμᾶς' ὑμεῖς.δὲ καθίσατε ἐν τῇ πόλει Υ ʿΙερ- ουσαλὴμ
καὶ νοητῆς τελειότητος.
μένον οὐκ ἐπίστευσαν.
τίον ἀράτω Ἰησοῦς
διδόντα ὑετὸν ἐπὶ τὴν γῆν, ἀποστέλλοντα ὕδωρ
ταρασσέσθω ὑμῶν ἡ καρδία, μηδὲ δευλιάτω. ἠκούσατε ὅτι ἐγὼ
τὴν ζημίην ἐπέθηκαν. Ζυώδεκα δέ μοι δοκέουσι πόλιας ποιή-
ἐστι' τὰ δὲ γενητὰ, ἔξωθεν ὄντα, πρόσκειται, ὡς τῇ
ὁ δὲ Κλεομένης τὸν ἱρέα ἐκέλευε τοὺς εἵλωτας ἀπὸ τοῦ
ἅπαξ ἀλλὰ πολλάκις.
ἐλθόντος. καὶ αὖθις ἔδοξε τούτου χάριν καὶ
κερματίζῃς αὐτό, ἐκεῖνοι πολλαπλασιοῦσιν, εὐλαβούμενοι
καὶ προλάμψαν τὸ ἐραστὸν αὐτοῦ καὶ τὸν κρυπτό- μενον
πεντακοσίας, οὺς πάντας ἡ τοῦ δεσπότου χάρις καὶ φιλανθρωπία διατρέφει.
ταύτης ἰδίᾳ προετρέπετο τὸν Σικόπαν κοινωνῆσαι
οὐδὲ παναρμονίου ἡμῖν δεήσει ἐν ταῖς ὠδαῖς τε καὶ
σημεῖα τοῦ τοῦτον συχοφαντεῖν ἐγχαλοῦντ' ἀφορ- μήν.
συμπεριλαμβανομένων καὶ ψυχῆς καὶ τῶν ἐν
πλὴν ἐξ ὠκυβόλων εἴ ποτε τόξων
σφι ἄρτισις περὶ τὸ σῶμα ἐστί.
μὴ πέσῃς κατέναντι ἐνεδρεύοντος
ο Εἰς τοῦτο ἐσυνέργησαν οἱ πρῶτοι τῆς γενεᾶς τῆς,
χωρίων ἢ οἰκιῶν ὑπῆρχον, πωλοῦντες ἔφερον τὰς τιμὰς
ᾧ δὲ περὶ ἑκάστην μεθοδον᾽) φιλοσοφοῦντι καὶ μὴ'
τῶν τῆς. παιδὸς γάμων, Ζεὺς διαλύσας ἐπέτρεψεν
ὑμῶν. πόλεις αἱ πρὸς νότον συνεκλείσθησαν, καὶ
οὐκ ἦν ὁ ἀνοίγων: ἀπωκίσθη Ιουδας,
πειρασμούς. Περὶ ταύτης ἡ Γραφὴ (ά. Κορ,
ἔπεσεν ἐπὶ πρόσωπον αὐτοῦ προσευχόμενος
ζητεῖ' οἶδεν. γὰρ ὁ-πατὴριὑμῶν ὁ οὐράνιος

Table 10: 40 randomly drawn lines of the Internet Archive pre-training dataset.

[20] http://www.perseus.tufts.edu/hopper/.

[21] http://classics.mit.edu/.

[22] http://www.documentacatholicaomnia.eu/.

[23] https://lexundria.com/.

[24] https://ryanfb.xyz/loebolus/.

restricted to providing translations of Latin and Greek texts, the Project Gutenberg offers a more diverse range of literature. Therefore, we use only books from Project Gutenberg that are tagged with the keyword "Latin". We report detailed statistics in Table 11.

Language	Dataset	Number of Tokens
Ancient Greek	Open Greek & Latin	30.0 million
	Greek Medieval Texts	3.3 million
	Patrologia Graeca	28.5 million
	Internet Archive	123.3 million
	Overall	**185.1 million**
Latin	Corpus Corporum	**167.5 million**
English	Perseus	10.8 million
	Classics Archive	4.9 million
	Lexundria	2.8 million
	Loebulus	14.0 million
	Project Gutenberg	28.7 million
	Documenta Catholica Omnia	151.7 million
	Overall	**212.8 million**

Table 11: Statistics of the pre-training datasets. Only Open Greek & Latin is used by Singh et al. (2021) and Yamshchikov et al. (2022) for their AG BERT models. Token counts determined by UNIX command `wc -w`.

D Further Results

Model	Accuracy
CLTK	96.51
UDPIPE (official)	94.71
UDPIPE (ours)	93.87 (0.05)
UDPIPE + GRεBERTA	94.17 (0.05)
GRεTA	97.40 (0.02)
GRεTA + Chars	**97.48 (0.02)**

Table 12: Lemmatization results on the Ancient Greek PROIEL dataset. The results are averaged over three runs with different random seeds, and the standard deviation is indicated in parentheses, except for the CLTK and UDPIPE (reported results).

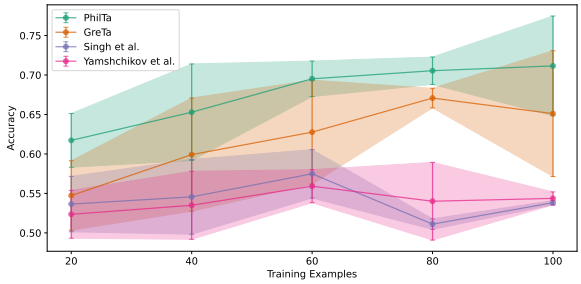

Figure 4: Synonym/antonym disambiguation accuracy scores for different few-shot training set sizes, for GRεTA and PHILTA against AG BERT models. The models are always given equal amounts of synonyms and antonyms, e.g., when using 20 training instances, the models are given 10 synonyms and 10 antonyms. We evaluate all models using k-fold cross-validation and report standard deviation as error bars.

Model	PoS Tagging		Dependency Parsing	
	UPoS	XPoS	UAS	LAS
CLTK	97.10	97.47	76.81	73.39
UDPIPE (official)	97.77	98.05	86.05	82.14
UDPIPE (ours)	97.99 (0.05)	97.68 (0.06)	85.64 (0.17)	81.70 (0.25)
UDPIPE + GRɛBERTA	98.56 (0.02)	**98.70** (**0.03**)	89.75 (0.16)	86.59 (0.15)
AG BERT (Singh et al., 2021)	97.98 (0.02)	98.14 (0.05)	88.50 (0.09)	84.72 (0.18)
AG BERT (Yamshchikov et al., 2022)	97.19 (0.06)	97.42 (0.08)	86.61 (0.21)	82.12 (0.15)
GRɛTA-ENC	98.16 (0.02)	98.31 (0.03)	89.93 (0.08)	86.48 (0.08)
PHILBERTA	98.15 (0.16)	98.45 (0.05)	**90.32** (**0.13**)	86.43 (0.61)
GRɛBERTA	**98.60** (**0.03**)	**98.70** (**0.04**)	90.28 (0.03)	**86.84** (**0.12**)

Table 13: PoS tagging and dependency parsing results on the Ancient Greek PROIEL dataset. The results are averaged over three runs with different random seeds, and the standard deviation is indicated in parentheses, except for the CLTK and UDPIPE (reported results). Note also that the CLTK is not trained on exactly the same data as the other models and therefore not directly comparable.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss general limitations in a section a dedicated section titled "Limitations". Furthermore, we acknowledge that the experiments on our small probing datasets do not allow to draw firm conclusions in Section 5 and Section 6.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The paper's claims are summarized in the Abstract and explicitly listed at the end of the Introduction (Section 1).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We use pre-trained Language Models for Ancient Greek, introducing them in Section 1, elaborating on them in Section 2 as Related Work, and using them in our experiments in Section 5 and Section 6. Furthermore, we pre-train Language Models, elaborating on them in Section 3 and using them in our experiments (Section 5 and 6) as well. Finally, we create a pre-training corpus for Ancient Greek, described in Section 3 and in Section C. The downstream task datasets that we use are introduced in Section 4 and in Section B.

☑ B1. Did you cite the creators of artifacts you used?
We cite the creators of the datasets we used in Sections 1 and 2. We cite relevant prior work and language models that we compare to in Sections 1 and 2. We elaborate on the datasets we use in Section 4.1 and specify the version.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The licenses for the data are discussed in Section C.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We specify the intended use for our pre-training dataset in Section 1 and the intended use for our language models in Section 1 and 2.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Given that we create our dataset utilizing open-domain texts from antiquity, we do not consider anonymization to be a significant concern.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Our pre-training corpus for Ancient Greek is described in Section 3 and in Section C. Our language models are documented in Section 3 and in Section A.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
For the Universal Dependencies and EvaLatin datasets that we used, we report the statistics in Appendix B. We report the creation of the pre-training and probing corpora and their statistics in Appendix B and C.

C ☑ Did you run computational experiments?

We elaborate on our experiments in Section 4. Pre-training and fine-tuning details can be found in Appendix A.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report pre-training and fine-tuning details in Appendix A.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We use the official evaluation scripts for our dataset, mentioned in Section B.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

The collection of our probing task data is described in Section B.

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
They were informed orally in a brief introductory session.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We report details about how the annotators were paid and how they were recruited in Appendix B.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Details about consent are reported in Appendix B.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We report the characteristics of the annotator population in Appendix B.