A State-of-the-Art Morphosyntactic Parser and Lemmatizer for Ancient Greek

Giuseppe G. A. Celano
Leipzig University
Faculty of Mathematics and Computer Science
Institute of Computer Science
celano@informatik.uni-leipzig.de

Abstract

This paper presents an experiment comparing six models to identify state-of-the-art models for Ancient Greek: a morphosyntactic parser and a lemmatizer that are capable of annotating in accordance with the Ancient Greek Dependency Treebank annotation scheme. A normalized version of the major collections of annotated texts was used to (i) train the baseline model Dithrax with randomly initialized character embeddings and (ii) fine-tune Trankit and four recent models pretrained on Ancient Greek texts, namely GreBERTa and PhilBERTa for morphosyntactic annotation and GreTA and PhilTa for lemmatization. A Bayesian analysis shows that Dithrax and Trankit are practically equivalent in morphological annotation, while syntax is best annotated by Trankit and lemmata by GreTa. The results of the experiment suggest that token embeddings are not sufficient to achieve high UAS and LAS scores unless they are coupled with a modeling strategy specifically designed to capture syntactic relationships. The dataset and best-performing models are made available online for reuse.

1 Introduction

In recent years, a few open-access annotated Ancient Greek (AG) corpora, such as *Opera Graeca Adnotata* (OGA) (Celano, 2024) and the GLAUx corpus (Keersmaekers, 2021), have been made available online. These corpora enable searches for morphosyntax and lemmata across a wide range of AG texts, thus filling the gap left by resources such as the *Thesaurus Linguae Graecae*, whose subscription-based query engine is limited to word forms and lemmata.

Because of the token count in the order of millions, the morphosyntactic annotation and lemmatization of the above-mentioned open-access corpora are feasible only if performed automatically. This raises a number of questions about which recent technology would be best suited for that purpose.

OGA v0.1.0 annotations (Celano, 2024) relied on the COMBO parser (Rybak and Wróblewska, 2018), which, despite being accurate,[1] was built on TensorFlow 1 and is not actively maintained anymore. The GLAUx corpus employed RFTagger (Schmid and Laws, 2008), Lemming (Müller et al., 2015), and the Stanford Graph-Based Dependency Parser (Dozat et al., 2017) for annotation of, respectively, morphology, lemmata, and syntax: the models perform well (see Keersmaekers, 2021, for details), but have not been released, and therefore cannot be reused.

For these reasons, the current paper presents a comparison of six models to identify and release state-of-the-art models for morphosyntactic annotation and lemmatization that can annotate literary AG sentences according to the annotation scheme of the Ancient Greek Dependency Treebank (AGDT) and can be used in production to process a large number of texts. To promote future machine learning-based studies on AG, the models and the normalized version of the AG texts used for training—and now documented with their alleged composition dates for the first time—are released.[2,3]

In Section 2, related work is reviewed, while Section 3 describes the dataset used for training. In Section 4, the experiment and the architectures of the different models compared are presented: the results of their training are reported with a Bayesian statistical analysis in Section 5 and discussed in Section 6. Finally, concluding remarks are contained in Section 7.

[1] https://git.informatik.uni-leipzig.de/celano/combo_for_ancient_greek.
[2] https://git.informatik.uni-leipzig.de/celano/morphosyntactic_parser_for_oga.
[3] https://git.informatik.uni-leipzig.de/celano/lemmatizer_for_oga.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 48–65, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-005

2 Related work

The explosion of machine learning in NLP has generated an ever-increasing number of resources, the reuse of which, however, is often not possible or straightforward due to the many different variables involved in each system.

The most recent endeavor comparable to the work presented here is Keersmaekers and Van Hal (2023). Building on Keersmaekers (2021), they documented the parsing and lemmatization of a large corpus consisting of literary and papyrological AG texts annotated according to the AGDT annotation scheme. Interestingly, they conducted experiments to increase LAS and UAS scores, in which the original data were transformed before training: for example, elliptical nodes were deleted and the annotation style for coordination modified. The reported results show some UAS and LAS increases in absolute terms. The models, however, have not been released.

Most recent systems for morphosyntactic annotation and lemmatization were trained on the Universal Dependencies data, which consist of two treebanks, the Perseus treebank and the PROIEL treebank,[4] for a total of about 416K tokens—notably, the size of the UD treebanks is less than half of that of the data annotated with the AGDT annotation scheme used in the present study (see Section 3).

The UD treebanks implement the UD annotation scheme differently, and therefore creation of a single model still represents a challenge: Kostkan et al. (2023) provided a joint spaCy model for morphosyntactic annotation and lemmatization that seems to achieve good overall performance.[5]

A number of studies reported on the creation of token embeddings for AG by using the large amount of texts available online (Singh et al., 2021; Yamshchikov et al., 2022). Most recently, Riemenschneider and Frank (2023) benchmarked a number of models for Ancient Greek and Latin. They show that their pretrained language model GreBERTa achieves the highest performance scores for UPOS, XPOS, UAS, and LAS in absolute terms when fine-tuned on the UD Perseus treebank (95.83, 91.09, 88.20, and 83.98, respectively); lemmatization is best performed by a T5 model they call GreTa, which achieves an F1 score of 91.14.

3 The dataset

The dataset used for training, validation, and testing consists of the following treebanks:[6] (i) the Ancient Greek Dependency Treebank[7] (Celano, 2019; Bamman and Crane, 2011), (ii) the Gorman Trees[8] (Gorman, 2020), and (iii) the Pedalion Trees.[9]

All treebanks were natively annotated using the AGDT annotation scheme, and together they represent by far the largest morphosyntactically annotated dataset for literary AG texts—and one of the largest treebanks in absolute terms: the token count of the texts before normalization is $1,277,310$ and, after it, $1,260,863$.

As Table 1 shows, the final dataset comprises a plethora of texts of different genres—including poetry, history, and philosophy—and periods, ranging from about the 9th century BCE to the 4th century CE (more details are provided in Appendix E). Even though the dataset is not balanced across genres and periods, it is still representative of most text types written in Ancient Greece during the above-mentioned time span.

3.1 Normalization

Since the final database consists of texts from different sources, which were annotated by many different scholars (sometimes adopting different conventions), automatic normalization of the original texts was attempted to foster consistency and therefore performance of machine learning algorithms.

Before training, all the relevant fields, i.e., word form, lemma, POS tag, syntactic head and relation, needed some non-trivial format standardization, especially to handle the case of null or clearly erroneous values. Syntactic trees also had to be modified if cycles were detected.

[4] Recently, the PTNK treebank (about 39K tokens) has been added, but, as far as we are aware, it has not yet been used for machine learning experiments.

[5] The scores for the model odyCy_joint on the UD Perseus treebank test set reported at https://centre-for-humanities-computing.github.io/odyCy/performance.html are 95.39 (POS tagging), 92.56 (morphological features), 78.80 (UAS), 73.09 (LAS), and 83.20 (lemmatization). It is, however, not clear whether the evaluation script used is that of the CoNLL 2018 Shared Task (https://universaldependencies.org/conll18/evaluation.html), which is commonly used in similar studies, including the present one. Since this script does not allow for cycles and multiple roots, we suspect that the reported scores would be lower, if it had been used.

[6] Data licences can be found at the links to the data specified below.

[7] https://github.com/PerseusDL/treebank_data/releases/tag/v2.1_IGDS.

[8] https://github.com/vgorman1/Greek-Dependency-Trees.

[9] https://github.com/perseids-publications/pedalion-trees.

Author	Genre	Century	Tokens
Hesiod, Homer	poem	−9/8	255,375
Sappho, Mimnermus, Semonides	lyric	−7	5,510
Homeric Hymns	hymns	−7/6	3,968
Aesop	fable	−6	5,221
Antiphon, Lysias, Isocrates	oratory	−5	30,679
Aeschylus, Sophocles, Euripides	tragedy	−5	108,386
Aristophanes, Cephisidorus Comicus	comedy	−5	47,547
Aeneas Tacticus	manual	−5	7,207
Herodotus, Thucydides	history	−5	65,494
Xenophon	history	−5/4	142,635
Lysias, Isocrates, Demosthenes, Aeschines, Andocides, Isaeus	oratory	−4	153,088
Aristotle, Plato, Theophrastus	philosophy	−4	51,906
Menandrus	comedy	−4	8,069
Epicurus	philosophy	−4/3	1,523
Theocritus	lyric	−3	304
Septuaginta	Bible	−3	19,235
Polybius	history	−2	105,693
Ezechiel	tragedy	−2	1,939
Batrachomyomachia	poem	−1	2,212
Diodorus of Sicily, Dionysius of Halicarnassus	history	−1	56,004
Chion	epistolary	+1	5,577
Hero of Alexandria	science	+1	10,321
Josephus Flavius	history	+1	24,987
Chariton	romance	+1/2	6,265
Plutarch	biography	+1/2	37,203
Phlegon of Tralles	paradox.	+2	5,642
Apollodorus	mythogr.	+2	1,265
Epictetus	philosophy	+2	7,204
Lucian	novel	+2	11,054
Appianus	history	+2	25,665
Athenaeus	miscellany	+2	45,653
Longus	romance	+2/3	672
Sextus Empiricus	philosophy	+3	16,218
Paeanius	history	+4	6,184
Julian	oratory	+4	1,405

Table 1: Statistics for the works contained in the dataset showing authors, genres, (alleged) centuries of composition (indicated by Arabic numbers, with + meaning CE and − BCE), and token counts (before normalization). Full details in Appendix E.

An often underestimated problem is that of character encoding for the apostrophe: all apostrophe-looking characters were converted into the character MODIFIER LETTER APOSTROPHE (U+02BC), which affected about 50K characters.

While the vast majority of AG graphic words corresponds to morphosyntactic tokens,[10] this is questionable for coordinate conjunctions such as οὐδὲ or εἴτε, which, in the final dataset, were tokenized (therefore, οὐ δὲ and εἴ τε, respectively). Coordination in the AGDT is not only annotated at the level of the syntactic tree but also at that of the syntactic label via use of the suffix `_CO`: to decrease the number of syntactic labels and therefore supposedly improve algorithm performance, this and similar suffixes, such as `_AP`, were deleted.

Another related yet different issue is represented by ellipsis, which poses a representational challenge. The AGDT annotation scheme allows elliptical nodes to be added whenever they are necessary to build a syntactic tree. However, the complexity of the phenomenon and the absence of strict annotation rules on this matter have over time led to the proliferation of various annotation styles: for example, sometimes the word form of an elliptical node is specified, sometimes it is not. The position of elliptical nodes within a sentence is also problematic both on a theoretical and a representational level.

While Keersmaekers and Van Hal (2023) proposed deletion of elliptical nodes, Celano's (2023) ellipsis modeling is followed in the present study: elliptical nodes were added at the end of a sentence (whatever their alleged position was) and, to avoid uncertainties about their word forms, they were always encoded with placeholders such as `[0]`, `[1]`, and so on, depending on their number.[11]

4 Experiment

A total of six model architectures were compared: four (i.e., three + baseline) for morphosyntactic prediction and three (i.e., two + baseline) for lemma prediction. More precisely, the baseline model called **Dithrax**[12] is able to predict both morphosyntax and lemmata, while the other five models can predict either one, in that their modeling for character prediction for lemmatization is kept distinct

[10] Crasis annotation, which is more elaborate to normalize, was left untouched.

[11] Since a model to predict such elliptical nodes is provided at https://git.informatik.uni-leipzig.de/celano/ellipsis_Ancient_Greek, new texts can be made compliant with this ellipsis annotation style.

[12] The name derives from Dionysius Thrax, the author of the first extant AG grammar.

Model	POS	XPOS	Feats	AllTags	UAS	LAS	Lemmas
Dithrax	95.55 (0.23)	90.65 (0.32)	94.40 (0.17)	89.80 (0.39)	77.70 (0.62)	70.81 (0.65)	86.85 (0.18)
Trankit	**96.18** (**0.13**)	**91.55** (**0.21**)	**94.61** (**0.12**)	**91.21** (**0.22**)	**82.28** (**0.27**)	**76.67** (**0.34**)	N/A
GreBERTa	94.12 (0.54)	89.16 (0.73)	93.21 (0.45)	88.31 (0.85)	58.85 (2.04)	53.41 (2.06)	N/A
GreTa	N/A	N/A	N/A	N/A	N/A	N/A	**91.17** (**0.17**)
PhilBERTa	85.34 (24.03)	79.85 (24.3)	86.67 (16.87)	77.8 (27.73)	61.24 (20.64)	54.95 (20.1)	N/A
PhilTa	N/A	N/A	N/A	N/A	N/A	N/A	90.09 (0.24)
UD Perseus Trankit	93.97	87.25	91.66	86.88	83.48	78.56	88.52
UD Perseus GreBERTa	95.83	91.09	N/A	N/A	88.20	83.98	N/A
UD Perseus GreTa + Chars	N/A	N/A	N/A	N/A	N/A	N/A	91.14
UD Perseus PhilBERTa	95.60	90.41	N/A	N/A	86.99	82.69	N/A
UD Perseus PhilTa + Chars	N/A	N/A	N/A	N/A	N/A	N/A	90.66

Table 2: Mean F1 scores + standard deviations in parentheses for the test set results of the 5-fold cross-validation models (training on each split repeated twice with different random seeds). Best scores are in boldface. Results for parsers trained on the UD Perseus data are shown only for loose comparison (see Section 5).

from that for word prediction for morphosyntax.

The performance of each model was evaluated with the official CoNLL 2018 Shared Task script: it outputs F1 scores for UPOS, XPOS, UFeats, AllTags (i.e., UPOS+XPOS+UFeats), UAS (i.e., HEAD match), LAS (i.e., HEAD + DEPREL match), and Lemmas. Since the AGDT tagsets are different, the above-mentioned metrics are conveniently renamed: POS, XPOS, Feats, AllTags, UAS, LAS, and Lemmas.

The original dataset was divided into training, validation, and test sets (60%, 20%, 20%). Each model was trained 10 times, using 5-fold cross-validation, with each training-validation split being used twice: as a result, 10 models (i.e., 5 splits $\times$ 2 random seeds) were trained for each model architecture (therefore, 10 final F1 scores were calculated for each of the above-mentioned metrics). Since the final models were not retrained on the entire dataset (train + validation sets) for time reasons, the mean scores presented in Table 2 are the ones obtained on the test set—the best-performing model was then chosen for use in production (see Table 3).

The training strategy is motivated by the fact that, while cross-validation reduces variance by use of different splits of the dataset, repetition of training on the same split allows experimentation with different random seeds. Final hyperparameters were set after a number of preliminary experiments and are documented in Appendix B.

Model	POS	XPOS	Feats	AllTags	UAS	LAS	Lemmas
Trankit	96.41	91.90	94.77	91.56	82.60	77.10	N/A
GreTa	N/A	N/A	N/A	N/A	N/A	N/A	91.41

Table 3: Scores of the best-performing cross-validation runs evaluated on the test set.

4.1 The statistical framework

The results of the present experiment are interpreted through the Bayesian analysis proposed by Benavoli et al. (2017). More precisely, they propose a Bayesian correlated t-test to compare cross-validation scores of two models on one dataset.

The proposed posterior distribution coincides with the Student distribution used in the frequentist t-test. This means that the probabilities of the Bayesian correlated t-test coincide with the p-values of the frequentist correlated t-test: what changes, however, is the interpretation of such numerical values.

While the frequentist approach returns the probability of data under the assumption that the null hypothesis is true, the Bayesian correlated t-test computes the actual probabilities of the null and alternative hypotheses.

Benavoli et al.'s (2017) Bayesian correlated t-test provides three probability scores concerning the comparison of the models x and y (see Appendix C for the scores):

(i) $P(x = y)$: the probability of model x being practically equivalent to model y: this is the *region of practical equivalence* (**ROPE**) corresponding to an arbitrary interval within which two models are considered not to differ in practice. In the present study, this is $[-1, 1]$, i.e., the posterior probability of the mean difference of F1 scores less than 1% is considered to mean practical equivalence.

(ii) $P(x \ll y)$: the probability that model x is practically worse than model y, i.e., the posterior probability of the mean difference of F1 scores being practically negative.

(iii) $P(x \gg y)$: the probability that model x is practically better than model y, i.e., the posterior probability of the mean difference of F1 scores being practically positive.

The Bayesian approach provides a more straightforward statistical interpretation of data and offers a solution for the well-known pitfalls of the frequentist framework, which include the fact that null hypotheses are always false in practice and sufficiently large datasets can yield statistical significance even if the effect size is very small.

4.2 Dithrax: the baseline model

As shown in Figure 1, Dithrax is a multi-output LSTM model vectorizing morphosyntactic tokens with *randomly initialized* character embeddings, which are used for prediction of both lemmata and, after further processing through LSTM layers, morphosyntax.

The model is inspired by the COMBO parser (Rybak and Wróblewska, 2018), which was among the most accurate parsers at the CoNLL 2018 Shared Task (Zeman et al., 2018).

More precisely, Dithrax proposes a similar modeling strategy for HEAD and DEPREL targets based on adjacency matrices resulting from dot products of two rank-2 tensors representing, respectively, heads and dependents of the same sentence, with each matrix row corresponding to the vector representation of a token.

4.3 Trankit

Trankit (Nguyen et al., 2021) is a state-of-the-art transformer-based toolkit for morphosyntactic analysis and lemmatization. It is designed for UD data, and is also able to process raw documents, in that it comprises a tokenizer and sentence splitter. Key features of Trankit are:

(i) use of the multilingual pretrained transformer XLM-RoBERTa, whose output is fine-tuned on new data.

(ii) adapters: feed-forward networks for each major component of Trankit (six in total), whose weights—together with the specific ones for final predictions—are the only ones updated, while the pretrained transformer weights remain fixed. These make Trankit memory- and time-efficient.

(iii) syntax is modeled via Dozat and Manning's (2017) biaffine attention.

For the purpose of the present experiment, we trained Trankit's joint model for part-of-speech tagging, morphological feature tagging, and dependency parsing (i.e., POS, XPOS, Feats, AllTags, UAS, and LAS scores); the lemmatizer could not be trained because of an internal code error.[13]

4.4 Pretrained models: Gre(BERTa|Ta) and Phil(BERTa|Ta)

The pretrained models GreBERTa and GreTa (for AG) and PhilBERTa and PhilTa (for AG and Latin) were fine-tuned for comparison,[14] in that they have recently been argued to perform better than previous pretrained AG models.

Riemenschneider and Frank (2023) fine-tuned GreBERTa and GreTa on the Greek data of the Open Greek and Latin Project, the CLARIN corpus of Greek Medieval Texts, the Patrologia Graeca, and the Internet Archive (in total, about 185.1M tokens). They fine-tuned PhilBERTa and PhilTa on not only AG but also Latin and English data. The latter come from the Corpus Corporum project (167.5M tokens) and a collection of English texts from different sources (212.8M tokens), whose topics are similar to the ones found in AG and Latin sources (for example, English translations of AG and Latin texts), for a total of 565.4M tokens.

GreBERTa and PhilBERTa are encoder-only transformers providing token embeddings for prediction of word-related targets (i.e., UPOS, XPOS, UFeats, AllTags, HEAD, and DEPREL). Since not the original scripts but only the pretrained models are made available online (see also Section 8), it was not possible to test the former with the AGDT dataset (see Section 6): in the present experiment, therefore, the pretrained token embeddings were just used as inputs to dense layers outputting the final probability scores for each token. However, the parameters of the pretrained models were left trainable. GreTa and PhilTa are encoder-decoder transformers for character prediction, and we fine-tuned them for lemmatization.[15]

[13] See https://github.com/nlp-uoregon/trankit/issues/48.

[14] We use the names GreBERTa, PhilBERTa, GreTa, and PhilTa to also name the models obtained by our fine-tuning: context is sufficient to clarify what these names exactly refer to.

[15] We are grateful to Frederick Riemenschneider, who provided us with a script for lemma prediction similar to the one used for his paper.

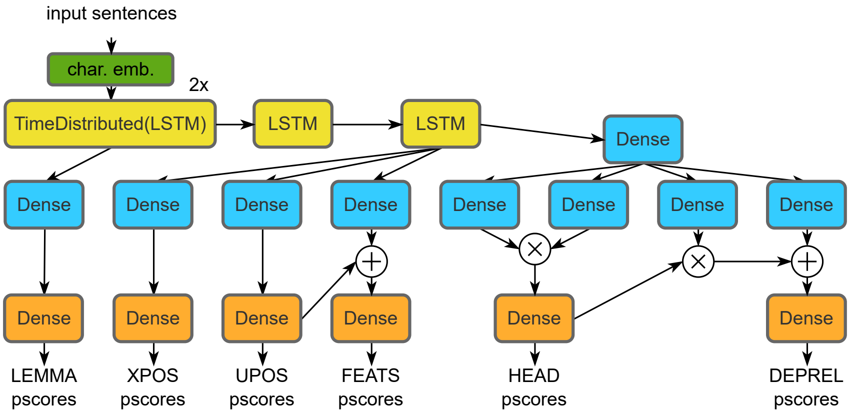

Figure 1: Main layers of Dithrax, the baseline model architecture. Blue color stands for `tanh(linear(x))`, while orange for `softmax(linear(x))` (with $\times$ meaning dot product and $+$ concatenation).

5 Results

Table 2 shows the mean F1 scores and related standard deviations[16] for the models trained with 5-fold cross-validation, with each split being used twice with different random seeds (in total, 10 models for each architecture). The mean scores are based on the F1 scores returned by the evaluation script of the CoNLL 2018 Shared Task applied to the results outputted by each model when tested on the test set. The models created by the runs with the best scores (see Table 3) are made available online.[17]

Table 2 also displays Riemenschneider and Frank's (2023) results for the models trained on UD Perseus data, i.e., a small subset of the dataset used for the present study, which were evaluated using the same CoNLL 2018 Shared Task script.[18] Even if the UD annotation scheme and the AGDT one are similar, there are differences that are likely to impact parsing results. For example, Keersmaekers (2021) argues that UD annotation style of coordination allows one to achieve higher scores for UAS and LAS. Moreover, UD data, unlike the AGDT data used for the present study, do not contain elliptical nodes. This means that comparison of F1 scores between UD models and the ones of the present study can only be loose, especially with reference to UAS and LAS.

The mean scores for PhilBERTa shown in Table 2 are the lowest ones and their related standard deviations are remarkably high (>20) because the model performed very poorly in one of the runs. However, even if that run were not considered, the mean scores would still be lower and the standard deviations would remain rather high in comparison to the values of the other models: POS: 92.87 (3.63); XPOS: 87.42 (4.37); Feats: 91.92 (3.11); AllTags: 86.44 (4.94); UAS: 67.44 (6.9); LAS: 60.94 (7.11).

Figures 2, 3, 4, 5, 6, 7, and 8 show the posterior distributions of the mean differences of F1 scores between all models pairwise returned by Benavoli et al.'s (2017) Bayesian correlated t-test.[19]

In each of the above-mentioned figures except Figure 8, the top-left, top-middle, top-right, bottom-left, bottom-middle, bottom-right plots show, respectively, the posteriors for the pairs Dithrax-Trankit, Dithrax-PhilBERTa, Dithrax-GreBERTa, Trankit-PhilBERTa, Trankit-GreBERTa, and GreBERTa-PhilBERTa. In Figure 8, which visualizes Lemmas scores, the left, middle, and right plots represent the posteriors for Dithrax-PhilTa, Dithrax-GreTa, and GreTa-PhilTa, respectively—as noted above, Trankit could not be trained for lemmatization because of an internal code error. Each above-mentioned Figure is coupled with a table (i.e., Tables 5, 6, 7, 8, 9, 10, and 11 in Appendix C), which reports the values of the areas under the curve.

Each single plot gives information about the probabilities that the mean differences of F1 scores between two models are practically negative, practically equivalent, and practically positive. For

[16]SDs have been calculated using `numpy.std` with `ddof=1`.

[17]See footnotes 2 and 3.

[18]Results for Trankit are taken from `https://trankit.readthedocs.io/en/latest/performance.html` (`Ancient_Greek-Perseus` treebank).

[19]The Python package documented at `https://github.com/janezd/baycomp` was used for the plots and calculations.

example, the bottom-left plot in Figure 4 and the corresponding Table 7 show:

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically negative, i.e., the integral of the posterior over the interval $(-\infty, -1)$, equal to ≈ 0.80. This is the probability that Trankit is practically **better** than PhilBERTa.

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically equivalent, i.e., the integral of the posterior over the ROPE interval $[-1, 1]$, equal to ≈ 0.06. This is the probability that PhilBERTa and Trankit are practically **equivalent**.

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically positive, i.e., the integral of the posterior over the interval $(1, +\infty)$, equal to ≈ 0.14. This is the probability that PhilBERTa is practically **better** than Trankit.

6 Discussion

Table 2 seems to suggest that Trankit is the best model in each morphosyntactic task. This is only *partly* confirmed by the Bayesian statistical analysis.

Even if Trankit's results for POS, XPOS, and Feats are the highest in absolute terms, its performance can be considered to be practically equivalent to that of the baseline model Dithrax with reference to these metrics. Indeed, the corresponding Tables 5, 6, and 7 show that the area under the curve within the ROPE is ≈ 1 for POS and Feats, and ≈ 0.88 for XPOS.[20]

On the other hand, the models PhilBERTa and GreBERTa perform practically worse than both Dithrax and Trankit with respect to these same metrics: there is at least an ≈ 0.79 probability (see Dithrax-PhilBERTa in Table 5)[21] that Dithrax or Trankit performs practically better.

This is an interesting result because, unlike Trankit, PhilBERTa, and GreBERTa, Dithrax does not rely on pretrained (but randomly initialized) character embeddings and its architecture has a lower overall number of parameters (see Table 4):[22] this suggests that classification tasks such as POS, XPOS, and Feats can be successfully addressed without use of more expensive model architectures—however, as shown in Table 4, Dithrax has a longer training time. The AllTags F1 score is a metric for POS+XPOS+Feats. Trankit turns out to perform practically better than any other model (Table 8), including Dithrax.

Syntactic prediction is notoriously more complex, and this is shown in the lower results reported in Table 2 for UAS and LAS. Trankit's performance is clearly superior to that of any other model, even if its scores are much lower than the POS and XPOS ones.

Syntactic analysis is a much more challenging task because HEAD and DEPREL values heavily depend on contextual information. Even if a pretrained transformer such as GreBERTa or PhilBERTa outputs context-aware token embeddings, it turns out to predict syntax poorly without a further modeling strategy.

In the GreBERTa and PhilBERTa models, the pretrained token embeddings were used as input to dense layers outputting probabilities for morphology and syntax in a multi-output model; however, while results for morphology are comparable to those of the other models, those for syntax clearly are not (see also Section 8): as Table 2 shows, UAS and LAS scores for GreBERTa and PhilBERTa are remarkably lower, and there is an ≈ 0.93 or higher probability that Dithrax or Trankit performs practically better than them (Tables 9 and 10).

This can be explained by the fact that, contrary to GreBERTa and PhilBERTa, Dithrax and Trankit employ a modeling strategy on top of embeddings: Dithrax models sentence syntax through adjacency matrices (Rybak and Wróblewska, 2018), while Trankit implements Dozat and Manning's (2017) biaffine attention mechanism, both of which aim to capture the complex relationship between heads and dependents within a sentence.[23]

Lemmatization is performed best by GreTa. While Dithrax simply employs LSTM layers over character embeddings, GreTa and PhilTa are seq2seq models: Table 11 shows that, while the seq2seq models perform practically better than Dithrax (≈ 1.00), there is an ≈ 0.75 probability that

[20] A threshold of 0.80 can be chosen when comparing the models.

[21] 0.79 is actually lower than the threshold of 0.80, but the difference is minimal.

[22] However, Trankit has fewer trainable parameters than Dithrax.

[23] To filter syntactic cycles, the Chu-Liu-Edmonds algorithm is applied to each parser's output.

GreTa performs practically better than PhilTa and an ≈ 0.25 probability that their performance is practically equivalent.

If we compare Trankit's results on the AGDT dataset with those on the UD dataset (see Table 2), scores for POS, XPOS, Feats, and AllTags are considerably higher in absolute terms on the AGDT dataset, with differences of ≈ 2.21, ≈ 4.3, ≈ 2.95, and ≈ 4.33, respectively; UAS and LAS scores, however, are higher on the UD dataset, with differences of ≈ 1.2 and ≈ 1.89, respectively. Interestingly, UAS and LAS scores do not seem to be impacted by the much larger size of the AGDT dataset; however, the model trained on the AGDT data can be expected to generalize much better than that trained on the UD data due to the much larger variety of texts used during training.

7 Conclusions

A comparison of six model architectures (Dithrax, Trankit, PhilBERTa, GreBERTa, PhilTa, and GreTa) was documented to select state-of-the-art models for annotation of morphosyntax and lemmata of literary texts according to the AGDT annotation scheme. A Bayesian statistical analysis was adopted to interpret cross-validation scores, which suggests that Trankit annotates syntax better than the other models do, while GreTa's performance for lemmatization is the best. The study shows that the baseline model Dithrax can also achieve state-of-the-art performance for morphological annotation—it employs randomly initialized character embeddings and a lower overall number of parameters, but its training time is longer.

A noteworthy finding of the study is that, although pretrained embeddings, such as GreBERTa and PhilBERTa, rely on complex model architectures vectorizing tokens with embeddings calculated on a very large collection of AG texts, they do not perform well for syntactic prediction (i.e., UAS and LAS scores), unless a further modeling strategy aimed at capturing syntax information within a sentence is put in place, such as adjacency matrices or biaffine attention.

8 Limitations

The study aimed to document state-of-the-art models for morphosyntactic analysis and lemmatization of Ancient Greek. The dataset used for training contains manual annotations produced over many years by different (single) annotators (some were students, others scholars). Therefore, as is often the case with manual annotations, annotation consistency within the dataset cannot be guaranteed because of either annotation errors or different annotation styles, the first annotation guidelines[24] not being sufficiently specific regarding a number of morphosyntactic phenomena—it should also be noted that the morphosyntactic annotation of Ancient Greek literary texts is arguably much more complex than that of modern texts.

For this reason, the present study set aside the question of how annotation quality/consistency affects parsing results. Similarly, no experiment was conducted with respect to corpus composition, under the assumption that model architectures are powerful enough to capture distinctions between texts of different genres and/or composition dates. Moreover, as stated in Section 1, the focus of the study was to select a morphosyntactic parser and a lemmatizer that performed best overall based on well-known metrics and a statistical analysis: a model error analysis would be of interest, but lies beyond the scope of this study.

The reuse of models and model architectures for comparison was often limited: either they are not released or the provided code is partial. The latter case is that of PhilBERTa and GreBERTa: they achieved state-of-the-art UAS and LAS scores on the UD Perseus treebank, but the original scripts have not been released,[25] and therefore their original model architectures could not be used in the present study.

Acknowledgments

This work has been supported by the German Research Foundation (DFG project number 408121292).

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.

[24] https://github.com/PerseusDL/treebank_data/blob/master/v1/greek/docs/guidelines.pdf; newer annotated texts should follow the much more specific annotation guidelines at https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/.

[25] https://github.com/Heidelberg-NLP/ancient-language-models/tree/main.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. 2017. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36.

Giuseppe G. A. Celano. 2019. The dependency treebanks for Ancient Greek and Latin. *Digital Classical Philology*, pages 279–298.

Giuseppe G. A. Celano. 2023. A neural network approach to ellipsis detection in Ancient Greek. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 151–158, Online. Association for Computational Linguistics.

Giuseppe G. A. Celano. 2024. Opera Graeca Adnotata: Building a 34m+ token multilayer corpus for Ancient Greek. *Preprint*, arXiv:2404.00739.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Vanessa B. Gorman. 2020. Dependency treebanks of Ancient Greek prose. *Journal of Open Humanities Data*, 6(1).

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 and Toon Van Hal. 2023. Creating a large-scale diachronic corpus resource: Automated parsing in the Greek papyri (and beyond). *Natural Language Engineering*, pages 1–30.

Jan Kostkan, Márton Kardos, Jacob Palle Bliddal Mortensen, and Kristoffer Laigaard Nielbo. 2023. OdyCy – A general-purpose NLP pipeline for Ancient Greek. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–134, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

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.

Piotr Rybak and Alina Wróblewska. 2018. Semi-supervised neural system for tagging, parsing and lemmatization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK. Coling 2008 Organizing Committee.

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.

Ivan P. 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.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Model statistics

Model	APar	TPar	TTime
Dithrax	$58,906,077$	$58,906,077$	≈14.6h
Trankit	$283,463,421$	$5,419,773$	≈6.9h
GreBERTa	$127,860,506$	$127,860,506$	≈2.6h
GreTa	$247,539,456$	$247,539,456$	≈11.4h
PhilBERTa	$137,076,506$	$137,076,506$	≈2.6h
PhilTa	$296,691,456$	$296,691,456$	≈12.3h

Table 4: Model statistics consisting of number of all parameters (APar), trainable parameters (TPar), and approximate training time (TTime) calculated on an NVIDIA RTX4500 ADA 24GB GDDR6.

B Model hyperparameters

The present section reports the relevant hyperparameters for the training of the models. Dithrax (TensorFlow/Keras): batch size 28, epochs 100 with early stopping (patience 2, best model saved), and Adam optimizer with clipvalue 4.5, $\beta_1 = 0.9$, $\beta_2 = 0.9$, weight decay 1e-4, and learning rate using piecewise constant decay with boundaries $[15000, 27000]$ and values $[0.001, 0.0001, 0.00007]$.

PhilBERTa and GreBERTa (TensorFlow/Keras/Transformers): batch size 28, epochs 100 with early stopping (patience 2, best model saved), and Adam optimizer with clipvalue 4.5, $\beta_1 = 0.9$, $\beta_2 = 0.9$, weight decay 1e-4, and learning rate using piecewise constant decay with boundaries $[10000]$ and values $[0.001, 0.0001, 0.00007]$.

Trankit (PyTorch/Transformers): token embeddings `xlm-roberta-base`, batch size 16, epochs 100 (best model saved), and a linear scheduler with warmup steps 80, training steps 160, and AdamW optimizer with learning rate 1e-3 and weight decay 1e-4.

PhilTa and GreTa (PyTorch/Transformers): `Seq2SeqTrainingArguments` with batch size 128, epochs 10, learning rate 1e-4, weight decay 1e-3, gradient accumulation steps 1, generation max length 30, and generation number of beams 20.

C Scores from the Bayesian correlated t-tests

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈1.00	≈0.00
Dithrax-PhilBERTa	≈0.79	≈0.04	≈0.17
Dithrax-GreBERTa	≈0.98	≈0.02	≈0.00
Trankit-PhilBERTa	≈0.80	≈0.04	≈0.16
Trankit-GreBERTa	≈1.00	≈0.00	≈0.00
GreBERTa-PhilBERTa	≈0.75	≈0.05	≈0.20

Table 5: Integrals on the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 2 (POS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈0.88	≈0.12
Dithrax-PhilBERTa	≈0.80	≈0.04	≈0.16
Dithrax-GreBERTa	≈0.97	≈0.03	≈0.00
Trankit-PhilBERTa	≈0.82	≈0.04	≈0.14
Trankit-GreBERTa	≈1.00	≈0.00	≈0.00
GreBERTa-PhilBERTa	≈0.76	≈0.05	≈0.19

Table 6: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 3 (XPOS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈1.00	≈0.00
Dithrax-PhilBERTa	≈0.80	≈0.06	≈0.14
Dithrax-GreBERTa	≈0.86	≈0.14	≈0.00
Trankit-PhilBERTa	≈0.80	≈0.06	≈0.14
Trankit-GreBERTa	≈0.98	≈0.02	≈0.00
GreBERTa-PhilBERTa	≈0.75	≈0.07	≈0.18

Table 7: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 4 (Feats).

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈0.00	≈1.00
Dithrax-PhilBERTa	≈0.80	≈0.04	≈0.17
Dithrax-GreBERTa	≈0.95	≈0.05	≈0.00
Trankit-PhilBERTa	≈0.82	≈0.03	≈0.14
Trankit-GreBERTa	≈1.00	≈0.00	≈0.00
GreBERTa-PhilBERTa	≈0.76	≈0.04	≈0.19

Table 8: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 5 (AllTags).

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈0.00	≈1.00
Dithrax-PhilBERTa	≈0.93	≈0.02	≈0.05
Dithrax-GreBERTa	≈1.00	≈0.00	≈0.00
Trankit-PhilBERTa	≈0.97	≈0.01	≈0.02
Trankit-GreBERTa	≈1.00	≈0.00	≈0.00
GreBERTa-PhilBERTa	≈0.36	≈0.08	≈0.56

Table 9: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 6 (UAS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	≈0.00	≈0.00	≈1.00
Dithrax-PhilBERTa	≈0.93	≈0.02	≈0.05
Dithrax-GreBERTa	≈1.00	≈0.00	≈0.00
Trankit-PhilBERTa	≈0.98	≈0.01	≈0.02
Trankit-GreBERTa	≈1.00	≈0.00	≈0.00
GreBERTa-PhilBERTa	≈0.39	≈0.09	≈0.52

Table 10: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 7 (LAS).

Model pair	Left	ROPE	Right
Dithrax-PhilTa	≈0.00	≈0.00	≈1.00
Dithrax-GreTa	≈0.00	≈0.00	≈1.00
GreTa-PhilTa	≈0.75	≈0.25	≈0.00

Table 11: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 8 (Lemmas).

D Posteriors

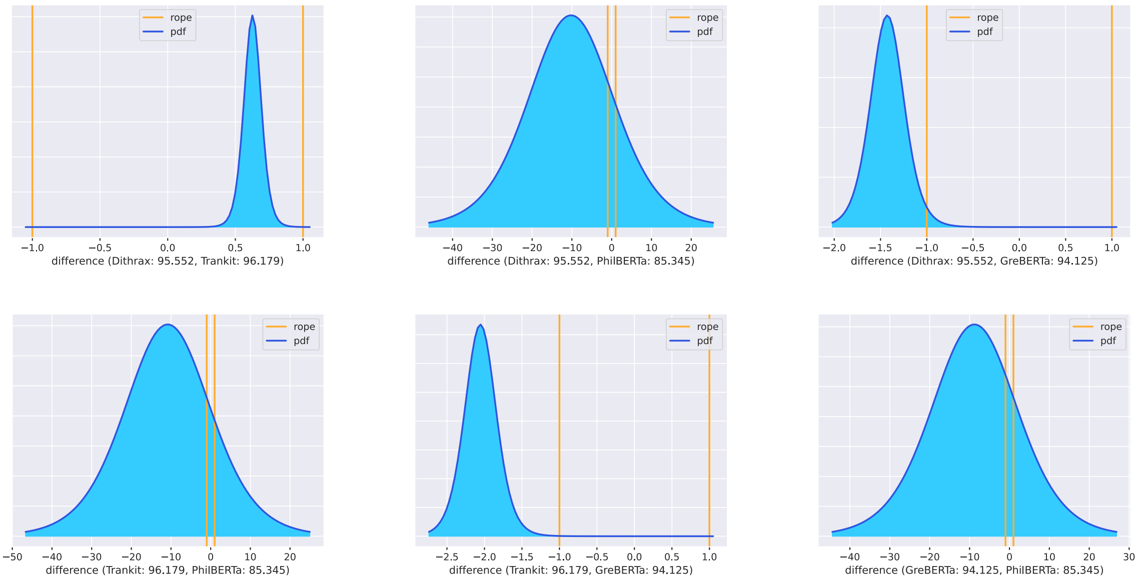

Figure 2: Posteriors of the Bayesian correlated t-test for all model pairs with reference to POS scores.

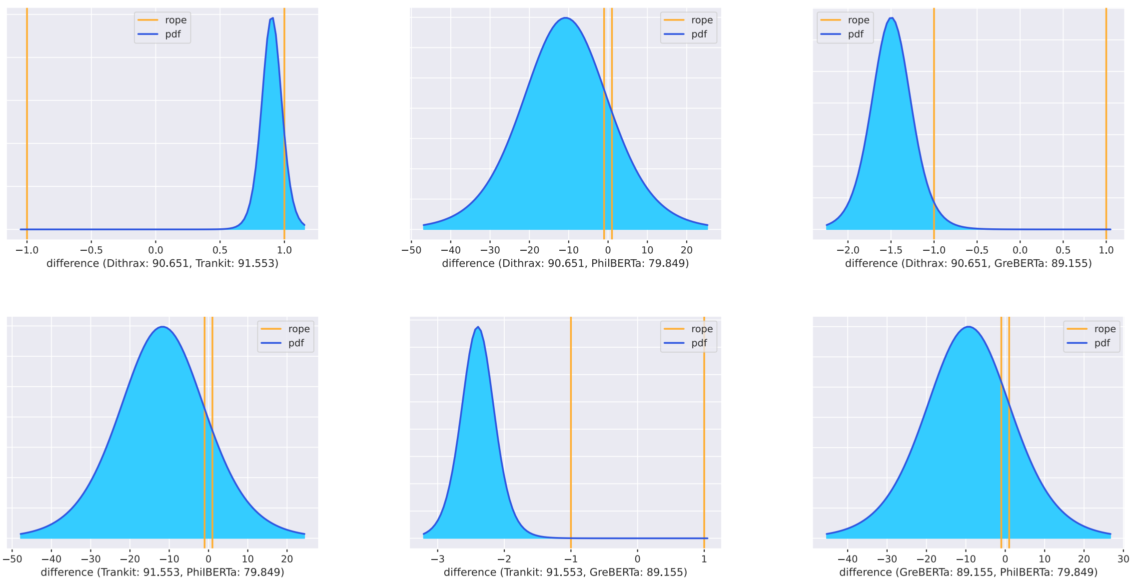

Figure 3: Posteriors of the Bayesian correlated t-test for all model pairs with reference to XPOS scores.

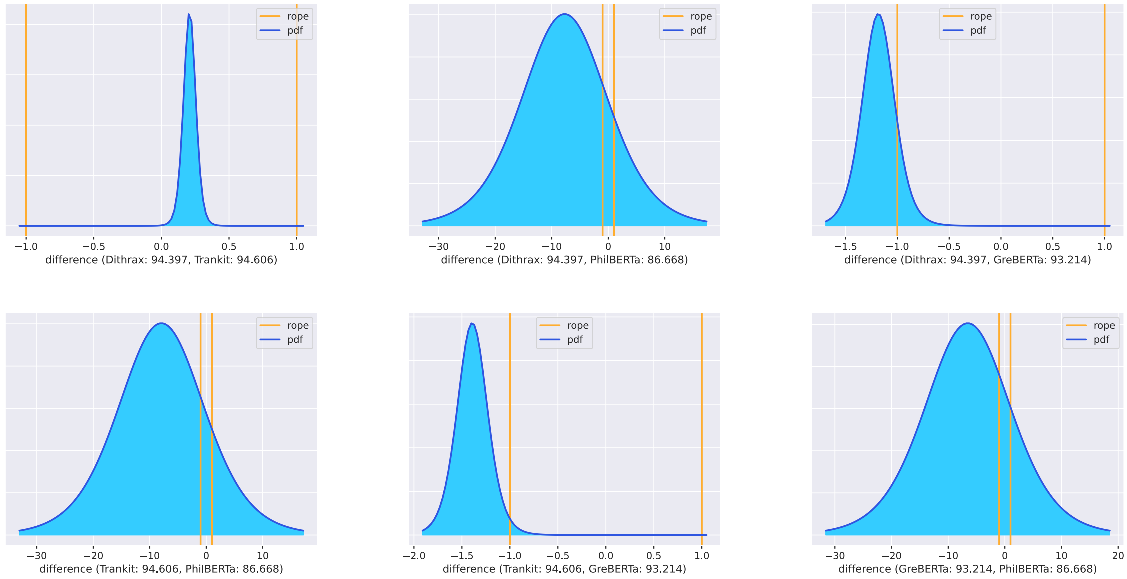

Figure 4: Posteriors of the Bayesian correlated t-test for all model pairs with reference to Feats scores.

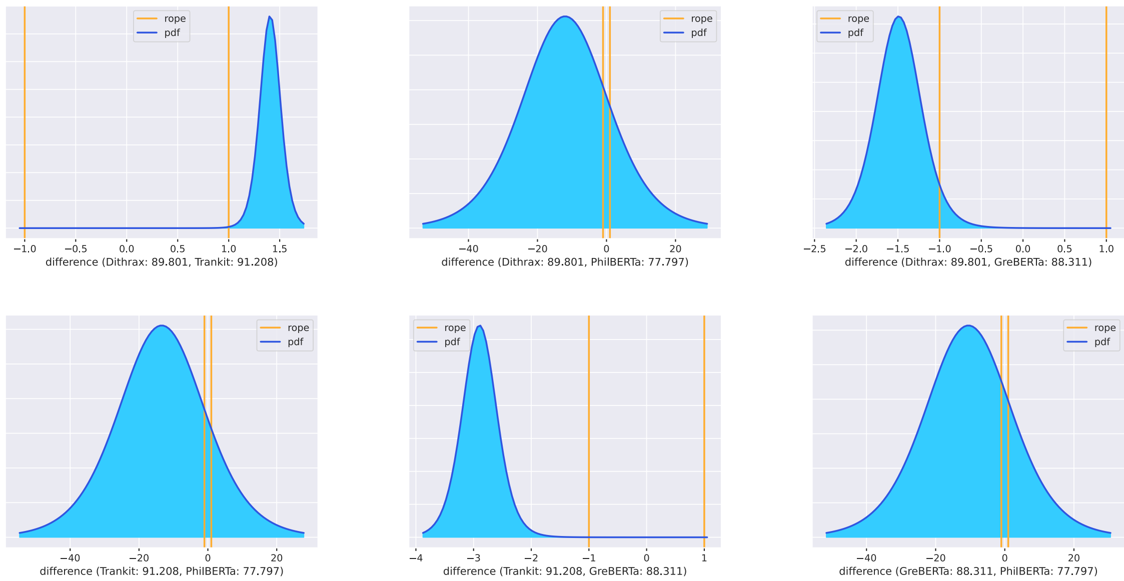

Figure 5: Posteriors of the Bayesian correlated t-test for all model pairs with reference to AllTags scores.

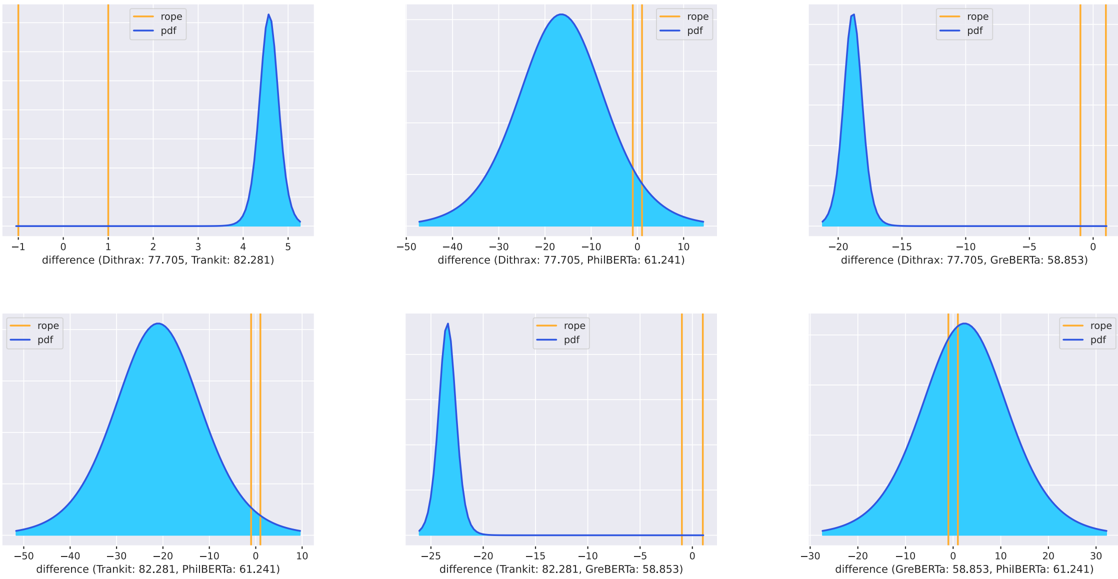

Figure 6: Posteriors of the Bayesian correlated t-test for all model pairs with reference to UAS scores.

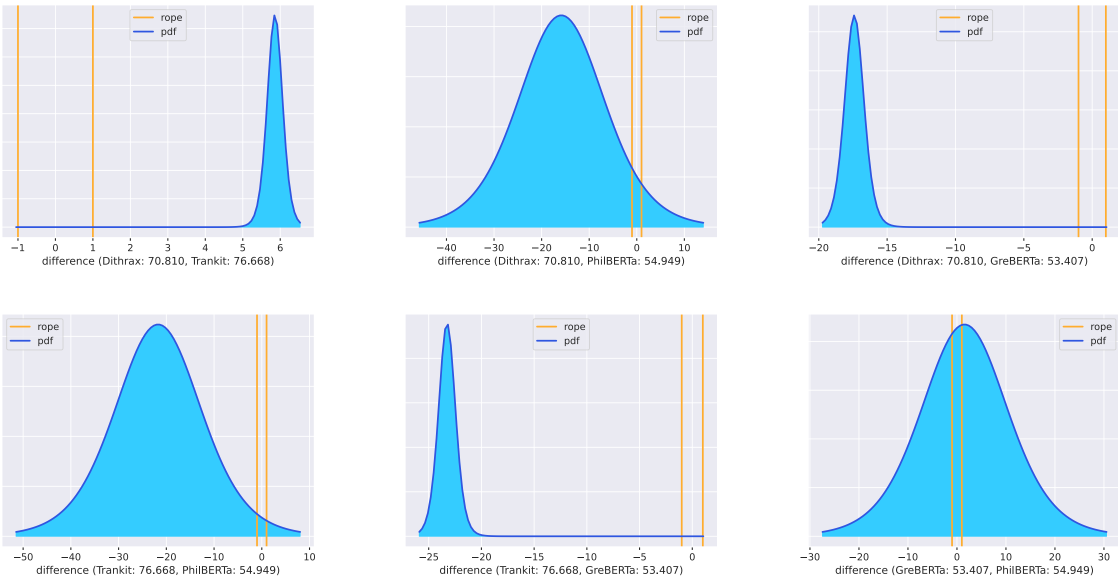

Figure 7: Posteriors of the Bayesian correlated t-test for all model pairs with reference to LAS scores.

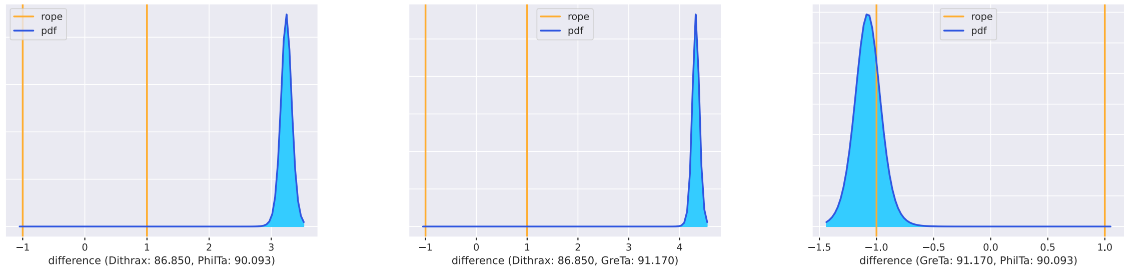

Figure 8: Posteriors of the Bayesian correlated t-test for all model pairs with reference to Lemmas scores.

E Texts

The following tables provide details of the texts used in the training, validation, and test sets (see also Table 3 for a more concise presentation). The authors, titles, and dates of each work were retrieved primarily from the file https://github.com/OperaGraecaAdnotata/OGA/tree/main/work_chronology/texts/chronology_greek_works.xml. This file contains work and title metadata derived from the canonical-greekLit[26] and First1KGreek[27] GitHub repositories, as well as from the Perseus Catalogue.[28] The dates of the works, expressed in ISO 8601 format, were manually annotated by a single annotator,[29] who used reference sources documented in the file mentioned above. All metadata should be regarded as work in progress.

[26] https://github.com/PerseusDL/canonical-greekLit.

[27] https://github.com/OpenGreekAndLatin/First1KGreek.

[28] https://catalog.perseus.org/.

[29] The annotator is an expert in AG literature and was paid fairly in accordance with German law.

CTS	Author	Title	Date	Tokens
tlg0003.tlg001	Thucydides	History of the Peloponnesian War	–0430-01/–0410-12	32,344
tlg0005.tlgxxx	Theocritus	Fragments	–0299-01/–0259-12	304
tlg0006.tlg003	Euripides	Medea	–0430-01/–0430-12	9,845
tlg0007.tlg004	Plutarch	Lycurgus	+0096-01/+0120-12	10,709
tlg0007.tlg015		Alcibiades	+0096-01/+0120-12	11,439
tlg0007.tlg086		On the Fortunes of the Romans	+0060-01/+0065-12	5,232
tlg0007.tlg087		On the Fortune or the Virtue of Alexander I and II	+0096-01/+0120-12	9,823
tlg0008.tlg001	Athenaeus of Naucratis	The Deipnosophists	+0175-01/+0200-12	45,653
tlg0009.tlg001	Sappho	Fragments	–0699-01/–0599-12	4,530
tlg0010.tlg002	Isocrates	Against Callimachus	–0401-01/–0401-12	4,109
tlg0010.tlg020		To Philip	–0345-01/–0345-12	466
tlg0011.tlg001	Sophocles	Trachiniae	–0449-01/–0449-12	9,026
tlg0011.tlg002		Antigone	–0442-01/–0437-12	8,990
tlg0011.tlg003		Ajax	–0438-01/–0435-12	9,751
tlg0011.tlg004		Oedipus Tyrannus	–0418-01/–0415-12	11,521
tlg0011.tlg005		Electra	–0417-01/–0406-12	10,806
tlg0012.tlg001	Homer	Iliad	–0799-01/–0700-12	130,479
tlg0012.tlg002		Odyssey	–0799-01/–0700-12	105,612
tlg0013.tlg002	Homeric Hymns	Hymn 2 to Demeter	–0624-01/–0574-12	3,968
tlg0014.tlg001	Demosthenes	First Olynthiac	–0348-01/–0348-12	2,194
tlg0014.tlg004		First Philippic	–0350-01/–0350-12	3,951
tlg0014.tlg007		On Halonnesus	–0342-01/–0341-12	2,886
tlg0014.tlg017		On the Treaty with Alexander	–0330-01/–0330-12	2,076
tlg0014.tlg018		On the Crown	–0329-01/–0329-12	26,435
tlg0014.tlg027		Against Aphobus I	–0363-01/–0362-12	5,346
tlg0014.tlg036		For Phormio	–0349-01/–0348-12	4,649
tlg0014.tlg037		Against Pantaenetus	–0346-01/–0346-12	4,528
tlg0014.tlg039		Against Boeotus I	–0347-01/–0346-12	3,351
tlg0014.tlg041		Against Spudias	–0363-01/–0358-12	2,333
tlg0014.tlg042		Against Phaenippus	–0329-01/–0329-12	2,624
tlg0014.tlg045		Against Stephanus I	–0349-01/–0348-12	6,839
tlg0014.tlg046		Against Stephanus II	–0349-01/–0348-12	2,168
tlg0014.tlg047		Against Evergus and Mnesibulus	–0354-01/–0354-12	6,235
tlg0014.tlg049		Apollodorus Against Timotheus	–0361-01/–0361-12	5,005
tlg0014.tlg050		Apollodorus Against Polycles	–0359-01/–0359-12	5,306
tlg0014.tlg051		On the Trierarchic Crown	–0359-01/–0357-12	1,580
tlg0014.tlg052		Apollodorus Against Callippus	–0368-01/–0367-12	2,490
tlg0014.tlg053		Apollodorus Against Nicostratus	–0367-01/–0366-12	2,340

CTS	Author	Title	Date	Tokens
tlg0014.tlg054	Demosthenes	Against Conon	–0354-01/–0340-12	3,755
tlg0014.tlg057		Against Eubulides	–0345-01/–0344-12	5,498
tlg0014.tlg059		Theomnestus and Apollodorus Against Neaera	–0342-01/–0339-12	10,489
tlg0016.tlg001	Herodotus	Histories	–0429-01/–0424-12	33,150
tlg0017.tlg003	Isaeus	The Estate of Pyrrhus	–0388-01/–0388-12	4,959
tlg0019.tlg001	Aristophanes	Acharnians	–0424-01/–0424-12	8,984
tlg0019.tlg008		Thesmophoriazusae	–0410-01/–0410-12	9,073
tlg0020.tlg001	Hesiod	Theogony	–0899-01/–0700-12	8,234
tlg0020.tlg002		Works and Days	–0899-01/–0700-12	7,116
tlg0020.tlg003		Shield of Heracles	–0899-01/–0700-12	3,934
tlg0026.tlg001	Aeschines	Against Timarchus	–0345-01/–0344-12	15,971
tlg0027.tlg001	Andocides	On the Mysteries	–0399-01/–0398-12	5,964
tlg0028.tlg001	Antiphon	Against the Stepmother for Poisoning	–0419-01/–0410-12	2,046
tlg0028.tlg002		First Tetralogy	–0479-01/–0410-12	2,915
tlg0028.tlg005		On the Murder of Herodes	–0417-01/–0417-12	7,458
tlg0028.tlg006		On the Choreutes	–0418-01/–0418-12	4,014
tlg0032.tlg001	Xenophon	Hellenica	–0361-01/–0353-12	27,401
tlg0032.tlg002		Memorabilia	–0409-01/–0353-12	27,840
tlg0032.tlg004		Symposium	–0369-01/–0360-12	7,291
tlg0032.tlg006		Anabasis	–0379-01/–0359-12	18,737
tlg0032.tlg007		Cyropaedia	–0368-01/–0365-12	50,690
tlg0032.tlg008		Hiero	–0356-01/–0356-12	6,953
tlg0032.tlg015		Constitution of the Athenians	–0442-01/–0405-12	3,723
tlg0041.tlg001	Chion	Epistulae	+0001-01/+0100-12	5,577
tlg0058.tlg001	Aeneas Tacticus	Poliorcetica	–0374-01/–0349-12	7,207
tlg0059.tlg001	Plato	Euthyphro	–0398-01/–0346-12	6,349
tlg0059.tlg002		Apology	–0398-01/–0389-12	10,457
tlg0059.tlg003		Crito	–0398-01/–0389-12	5,093
tlg0059.tlg029		Cleiphon	–0398-01/–0346-12	1,875
tlg0060.tlg001	Diodorus of Sicily	Historical Library	–0059-01/–0029-12	25,692
tlg0061.tlg001	Lucian of Samosata	Asinus	+0125-01/+0180-12	11,054
tlg0081.tlg001	Dionysius of Halicarnassus	Antiquitates Romanae	–0007-01/–0006-12	30,312
tlg0085.tlg001	Aeschylus	Supplices	–0465-01/–0458-12	6,071
tlg0085.tlg002		Persians	–0471-01/–0471-12	6,381
tlg0085.tlg003		Prometheus Bound	–0459-01/–0455-12	7,222
tlg0085.tlg004		Seven against Thebes	–0466-01/–0466-12	6,372
tlg0085.tlg005		Agamemnon	–0457-01/–0457-12	10,037
tlg0085.tlg006		Libation Bearers	–0457-01/–0457-12	5,846
tlg0085.tlg007		Eumenides	–0457-01/–0457-12	6,518

CTS	Author	Title	Date	Tokens
tlg0086.tlg035	Aristotle	Politics	–0399-01/–0299-12	19, 867
tlg0093.tlg009	Theophrastus	Characters	–0316-01/–0316-12	8, 265
tlg0096.tlg002	Aesop	Aesop's Fables	–0599-01/–0500-12	5, 221
tlg0255.tlg001	Mimnermus of Colophon	Fragmenta	–0699-01/–0599-12	213
tlg0260.tlg001	Semonides of Amorgos	Fragmenta	–0699-01/–0599-12	767
tlg0343.tlg001	Ezechiel	Exagoge	–0199-01/–0099-12	1, 939
tlg0429.tlg001	Cephisodorus Comicus	Fragmenta	–0401-01/–0401-12	29, 490
tlg0526.tlg004	Josephus Flavius	The Jewish War	+0075-01/+0075-12	24, 987
tlg0527.tlg001	Septuaginta	Genesis	–0299-01/–0200-12	19, 235
tlg0537.tlg012	Epicurus	Epistula ad Menoeceum	–0310-01/–0270-12	1, 523
tlg0540.tlg001	Lysias	On the Murder of Eratosthenes	–0402-01/–0401-12	2, 834
tlg0540.tlg012		Against Eratosthenes	–0402-01/–0402-12	5, 638
tlg0540.tlg013		Against Agoratus	–0399-01/–0397-12	5, 641
tlg0540.tlg014		Against Alcibiades 1	–0394-01/–0394-12	2, 801
tlg0540.tlg015		Against Alcibiades 2	–0394-01/–0394-12	688
tlg0540.tlg019		On the Property of Aristophanes	–0386-01/–0386-12	3, 624
tlg0540.tlg023		Against Pancleon	–0399-01/–0398-12	896
tlg0540.tlg024		On the Refusal of a Pension	–0402-01/–0402-12	1, 665
tlg0541.tlg007	Menander of Athens	Dyscolus	–0315-01/–0315-12	8, 069
tlg0543.tlg001	Polybius	Histories	–0167-01/–0117-12	105, 693
tlg0544.tlg002	Sextus Empiricus	Adversus Mathematicos	+0201-01/+0300-12	16, 218
tlg0548.tlg001	Apollodorus	Library	+0101-01/+0200-12	1, 265
tlg0551.tlg017	Appianus of Alexandria	Civil Wars	+0101-01/+0200-12	25, 665
tlg0554.tlg001	Chariton	De Chaerea et Callirhoe	+0075-01/+0125-12	6, 265
tlg0557.tlg001	Epictetus	Discourses	+0108-01/+0108-12	7, 204
tlg0559.tlg002	Hero of Alexandria	De Automatis	+0062-01/+0085-12	10, 321
tlg0561.tlg001	Longus	Daphnis and Chloe	+0101-01/+0300-12	672
tlg0585.tlg001	Phlegon of Tralles	Book of Marvels	+0100-01/+0200-12	5, 642
tlg1220.tlg001	Batrachomyomachia	Batrachomyomachia Homerica	–0099-01/–0029-12	2, 212
tlg2003.tlg001	Julian	Panegyric in Honor of the Emperor Constantinus	+0355-01/+0355-12	1, 405
tlgxxxx.tlgxxx	Paeanius	Brevarium	+0337-01/+0379-12	6, 184