Detecting Semantic Reuse in Ancient Greek Literature: A Computational Approach.

Caterina D'Angelo[1, *], Andrea Taddei[1] , and Alessandro Lenci[1,]

[1] *Università di Pisa, Lungarno Pacinotti 43, 56126 Pisa, Italy*

Abstract

This paper introduces the first step towards a computational method for detecting semantic textual reuse in Ancient Greek literature. While existing tools focus primarily on exact or near-lexical matching, our approach leverages the semantic capabilities of contextual LLMs, aiming to finetune a pretrained encoder via contrastive learning to recognize textual reuse even when expressions are paraphrased and/or morphologically altered.

To build a suitable dataset, we developed an automatic pipeline that generates positive samples by extracting paraphrases for each sentence using the Ancient Greek Wordnet and a custom-trained morphological re-inflection model. Negative samples, or "confounders", are selected through topic modeling to ensure thematic relevance while preserving semantic dissimilarity.

The model is evaluated through a curated case study on Homeric formulae. We retrieve the top ten most similar sentences in a corpus of Ancient Greek authors from the classical age, assessing model outputs using both standard metrics and comparison with established philological studies. The outcomes demonstrate that contrastive fine-tuning, paired with linguistically informed data augmentation, offers promising directions for identifying non-literal textual reuse in historical corpora. This work contributes a framework for philological discovery, combining deep learning with interpretive scholarship in classical studies.

Keywords

Ancient Greek, intertextuality, contrastive learning, paraphrase generation, topic modeling, morphological inflection, synonyms extraction, computational philology.

1. Introduction

Reuse and, more generally, intertextuality have always been peculiar lens through which literary works can be analyzed. It has been the focus of literary critics and philologists such as Gerard Genette (Genette, 1982), Julia Kristeva (Kristeva, 1986), Roland Barthes (Barthes, 1975) and Michael Riffaterre (Riffaterre, 1978) to establish the importance of intertextual allusions as well as "word by word" quotations, with structuralist thinking going as far as to say that «*Intertextuality is…. The mechanism specific to literary reading. It alone, in fact, produces significance, while linear reading, common to literary and nonliterary texts, produces only meaning.* (Genette, 1982, p. 18)». With the present work, our aim is to build a computational tool that can aid in the complex task of identifying instances of re-use in Ancient Greek texts. We start from the definition that Gerard Genette gives us of intertextuality, focusing on its less literal guise: «*it is the traditional practice of quoting […] in a still less explicit and less literal guise, it is the practice of allusion* (Genette, 1982, p. 18).». In the following paper we focus specifically on semantic

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]

* Corresponding author.

c.dangelo6@studenti.unipi.it (C. D'Angelo); andrea.taddei@unipi.it (A. Taddei); alessandro.lenci@unipi.it (A. Lenci)

0009-0008-4500-225X (C. D'Angelo) 0000-0002-8977-5528 (A. Taddei); 0000-0001-5790-4308 (A. Lenci);

reuse by developing methods to detect semantic connections that may indicate shared themes, motifs, or conceptual relationships between texts. Our approach represents a foundational step toward the broader goal of computational intertextuality detection, providing scholars with a tool to identify semantically related passages that merit further philological investigation.

1.1. Related Works

Existing computational tools for reuse detection in classical languages are primarily based on lexical similarity. Among them, the most prominent is the **Tesserae project** (Coffee, et al., 2013), which identifies parallels in Latin and Ancient Greek texts by combining lexical overlap with phonetic and thematic similarity, the latter through topic modeling algorithms. Nonetheless, such thematic similarity does not imply intentional intertextuality, which involves the conscious use of another author's language or ideas.

Another widely used tool is **Diogenes**[3], a desktop application that enables exact lexical searches across a large corpus of classical texts.

Another significant tool in this domain is **TRACER** (Büchler et al., 2014), a flexible framework for automatic detection of text reuse that supports multiple similarity measures.

Despite their usefulness, these systems are focused on surface-level matches and fail to capture semantic paraphrases or allusive reuse.

To detect such deeper forms of intertextuality, recent approaches have turned to distributional semantics. A key challenge, however, is the scarcity of annotated and homogeneous corpora in ancient languages, which makes training large language models (LLMs) difficult (Moritz, Wiederhold, Pavlek, Bizzoni, & Buchler, 2016).

A seminal contribution in this direction is (Burns, Brofos, Li, Chaudhuri, & Dexter, 2021) who uses **Word2Vec** embeddings to measure the semantic similarity between Latin bigrams. Their method computes pairwise cosine similarities between words and averages the results. Although effective, they acknowledge the limitations of static embeddings and propose that contextual embeddings (e.g., BERT-based models) may offer better nuance and generalization.

The paper by Burns et al. also frames intertextuality as a form of anomaly detection, using the embeddings created with the corpus of a specific author (in this case Livy) as input for a **SVM**: with this model, the goal is to predict the "Livianess" of each work, so as to find instances in which the authors have alluded to Livy's works.

Following the parallel between intertextuality and anomaly detection, similar methods have been explored in the context of authorship attribution. In (Yamschikov, Tikhonov, Pantis, Schubert, & Jurgen, 2022) the authors aim to obtain contextual embeddings for Ancient Greek by leveraging transfer learning. Starting from pre-trained models, they fine-tune both a multilingual transformer and one trained on Modern Greek, adapting them to downstream tasks in Ancient Greek.

While this approach demonstrates the feasibility of adapting general-purpose models to low-resource historical languages, it suffers from the limitations of using a tokenizer and vocabulary not optimized for Ancient Greek.

A common obstacle encountered in our research pertains the shortage of digitized Ancient Greek texts. The main source would be the *Thesaurus Linguae Grecae*[4], but its policy is against using the data for machine learning purposes.

Nonetheless, the work by (Yamschikov, Tikhonov, Pantis, Schubert, & Jurgen, 2022) inspired our own application of transfer learning, allowing us to make efficient use of limited annotated data while focusing on semantic reuse detection.

A similar strategy is adopted by (Riemenschneider & Frank, 2023), who leverage pre-trained language models to detect intertextual allusions in a multilingual setting, analyzing sentence-level correspondences across Ancient Greek, Latin, and English. Although their focus lies primarily on cross-lingual reuse, their work further confirms the potential of contextual models in identifying non-literal textual relationships.

1.2. Contributions

This paper makes the following contributions:

- We propose an automated pipeline for generating paraphrases of Ancient Greek sentences, combining resources such as the Ancient Greek WordNet with a custom-trained morphological re-inflection model based on annotated Ancient Greek data.
- We conduct a qualitative assessment of different contextual encoders for

[3] https://d.iogen.es/web/?ver=1.003&user=stud

[4] https://stephanus.tlg.uci.edu/

Ancient Greek, tested on a synonym selection task.
- We introduce a method for automatically generating hard negative samples: sentences with high lexical overlap but low semantic relatedness.
- We fine-tune a domain-specific pretrained language model to capture non-lexical, semantic forms of textual reuse in Ancient Greek literature.
- We evaluate our approach on a curated case study of Homeric formulae, assessing semantic reuse in classical Greek authors through both retrieval metrics and philological validation.

2. Method

To fine-tune a model for semantic reuse detection in Ancient Greek, we first selected a suitable encoder.

We then constructed a contrastive dataset consisting of 11,305 triplets, each composed of a query sentence, a positive sample (paraphrase), and a negative sample (confounder). The query sentences were randomly extracted from a subcorpus of works by Homer, Thucydides, and Herodotus, taken from the *Opera Graeca Adnotata*. Positive and negative samples were generated automatically through the paraphrase and confounder generation pipeline described in Sections 2.1 and 2.2.

2.1. Model Selection

Although Ancient Greek remains a low-resource language, recent years have seen the development of several contextual language models tailored to its linguistic properties. For our task, the encoder must be able to encode semantic contextual information, particularly the similarity between lexically and morphologically varied expressions.

To evaluate model performance in capturing semantic relationships, we designed a synonym retrieval task, which will be described in detail in Section 2.2.

The models considered include:

- **Logion** (Cowen-Breen, Brooks, Haubold, & Graziosi, 2023): A BERT-based architecture pre-trained on modern Greek and fine-tuned on Ancient Greek texts from *First1KGreek*[5], *Perseus Digital Library*[6] and data obtained from fellow scholars. The training corpus comprises approximately 70 million words. In its 50K version, a WordPiece tokenizer was trained on the same corpus, resulting in a vocabulary of 50,000 subword units tailored to Ancient Greek.
- **GreBERTA** (Riemenschneider & Frank, Exploring Large Language Models for Classical Philology, 2023): A RoBERTa-style encoder with dynamic masking, trained on a composite corpus including the *Open Greek and Latin Project*[7] (30M tokens), the *CLARIN Greek Medieval corpus*[8] (3.3M), the *Patrologia Graeca*[9] (28.5M), and the Ancient Greek texts contained in the *Internet Archive*[10] (123.3M). Despite its size, the latter source contains substantial noise and inconsistencies.
- **Word2Vec**: A non-contextual baseline model, included for comparison.

As will be further explained in section 2.2, lemmatization was necessary for synonym extraction. We therefore compared the two main lemmatization libraries available for Ancient Greek: **CLTK**[11] and **greCy**[12].

Table 1 reports the top predicted synonym for the word βαίνω (whose meaning in the context of the selected sentence is “to go up”) across all model and lemmatizer combinations. A broader comparison covering multiple lexical entries is available in the appendix.

Table 1
Top predicted synonyms for the word βαίνω

[5] https://opengreekandlatin.github.io/First1KGreek/
[6] https://www.perseus.tufts.edu/hopper/
[7] https://opengreekandlatin.org/.
[8] https://inventory.clarin.gr/corpus/890.
[9] https://patristica.net/graeca/.
[10] https://archive.org/.
[11] http://cltk.org/
[12] https://github.com/jmyerston/greCy

Predicted synonym	Model and lemmatization	Meaning	Similarity Score
διαβαίνω	Logion 50K with CLTK	“To go up”	0.34
διαβαίνω	Logion 50k with greCy	“To go up”	0.31
στείχω	Logion BASE with CLTK	“To go”	0.48
στείχω	Logion BASE with greCy	“To go”	0.45
στείχω	GreBerta with CLTK	“To go”	0.42
στείχω	GreBerta with greCy	“To go”	0.42
διαβαίνω	Word2Vec with CLTK	“To go up”	0.89
διαβαίνω	Word2Vec with greCy	“To go up”	0.89

We didn’t consider the model described in (Pranaydeep, Rutten, & Lefever, 2021) since the ***Logion*** models are initialized with the same weights and increase the size of the finetuning corpus.

The following example illustrates the full paraphrase generation process, including synonym substitution and morphological re-inflection.

Table 2
Example of the paraphrase generation process

Version	Greek sentence	Translation
Original	είς ταύτην οὖν τὴν **άκτὴν** ἐξ Ἀβύδου **ὁρμώμενοι** ἐγεφύρουν τοῖς προσέκειτο	Towards this shore, then, starting from Abido, they built a bridge, those who had been assigned the task.
Paraphrased	είς ταύτην οὖν τὴν **ἄκραν** ἐξ Ἀβύδου **ἐξιστάμενοι** ἐγεφύρουν τοῖς προσέκειτο	Towards this end, then, moving away from Abido, they built a bridge those who had been assigned the task.

2.2 Positive Samples

Since the objective of our model is to detect semantic reuse, positive samples must exemplify cases of non-literal reuse. For this purpose, we developed an automated pipeline for paraphrase generation through targeted lexical substitution, following data augmentation techniques such as those described in (Bayer, Kaufhold, & Reuter, 2022).

Specifically, we focused on substituting semantically salient tokens—nouns, verbs, and adjectives—with suitable synonyms. To identify these, we combined lexical information from the Ancient Greek WordNet (Bizzoni, et al., 2014) with semantic similarity estimates derived from contextual embeddings.

For each semantically relevant word in a sentence, we queried the WordNet to retrieve its synsets (i.e., sets of synonyms grouped by sense). For each offset (individual sense), we collected a candidate list of synonyms. We then computed the cosine similarity between the contextual embedding of the original word and four contextual embeddings of each synonym, obtained by extracting four different sentence contexts in which that synonym appears. The sentences were extracted from the corpus *Lemmatized Ancient Greek Texts*[13] by Giuseppe Antonio Celano.

This method allowed us to select the most semantically coherent synonym among candidates, accounting for the high degree of polysemy in Ancient Greek vocabulary.

2.1.1. Re-inflection Model

As mentioned above, the synonym selection pipeline outputs the lemma of the best synonym. However, to generate a valid paraphrase within the Ancient Greek sentence, it is necessary to re-inflect the selected lemma according to the morphological features of the word it replaces.

To this end, we developed a morphological re-inflection model, which takes as input the lemma and a set of morphological features (e.g., case, number, tense) and returns the inflected form.

The model was trained on a corpus constructed by merging and normalizing data from multiple resources:

- ***SIGMORPHON 2023 – UniMorph Shared Task***[14]: 5,572 inflected forms annotated with morphosyntactic features.

[13] https://github.com/gcelano/LemmatizedAncientGreekXML
[14] https://github.com/sigmorphon/2023InflectionST

- ***Perseus Project***: A dataset of 1,290,544 linguistically annotated forms originally produced by the Morpheus parser and generator of Ancient Greek inflected forms (Crane, 1991).

- ***Opera Graeca Adnotata***[15] (Celano, 2024): A morphologically annotated corpus curated by G. A. Celano, from which we extracted 589,105 forms.

After removing defective entries and applying standard normalization procedures (e.g., Unicode harmonization, feature unification), we trained a sequence-to-sequence model composed of an LSTM layer, a dropout layer, and a Bidirectional LSTM decoder. This architecture was chosen for its balance between simplicity and effectiveness in character-level morphological generation tasks.

2.2. Negative Samples

To create negative samples for the contrastive learning task we introduced the notion of lexical confounders: these are sentences that share semantically relevant words with the target sentence but express a different meaning. This technique allows us to create “hard negatives”, capable of aiding the model in identifying sentences with no lexical overlap but semantically similar, teaching it to disentangle lexical similarity from semantic equivalence.

To automatically select these confounders, we applied topic modeling with the goal of identifying sentences that differ in thematic content. The underlying assumption is that sentences on distinct topics are unlikely to convey the same meaning, even if they share lexically similar elements.

The topic modeling process was carried out on the *Opera Graeca Adnotata* corpus, leveraging lemmatized tokens to improve generalization. We first applied the **Hierarchical Dirichlet Process (HDP)** to estimate the optimal number of latent topics (resulting in k = 10), and then trained a **Latent Dirichlet Allocation (LDA)** model accordingly. The resulting LDA model achieved an average UMass topic coherence score of −0.68, indicating a moderate level of interpretability suitable for the identification of semantically distinct negative samples.

3. Results

In this section, we present the results obtained from the evaluation of the two main components of our pipeline: the re-inflection model and the contrastive sentence encoder.

3.1. Re-inflection Model Evaluation

To generate grammatically coherent paraphrases, we trained a sequence-to-sequence model to perform morphological inflection from lemma + features to surface form. The architecture consists of a single-layer LSTM followed by dropout and a bidirectional LSTM.

The model was trained for a maximum of 120 epochs with early stopping (patience = 10), halting at epoch 77. We used the Adam optimizer with a learning rate of 0.001.

The learning curves of accuracy and loss for the training and validation set can be seen in Figure 1 and Figure 2.

The model reached 0.90 accuracy on both the validation and test set. While performance on frequent forms is consistent, rare accented forms remain problematic. For instance, characters such as “ΐ” and “ΰ”, which appear only 398 and 48 times respectively in the validation set, obtained F1-scores as low as 0.39 and 0.21. This imbalance affects the macro average, which is significantly lower than the weighted average, as shown in Table 3 (test set results).

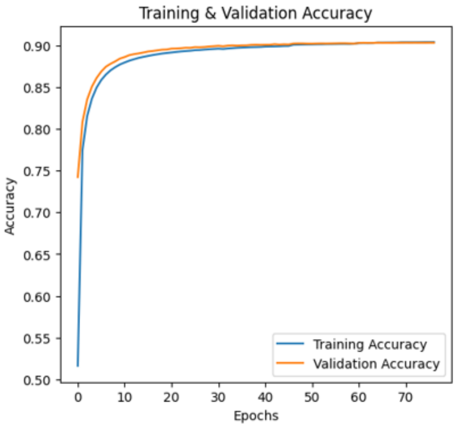


Figure 1: Training and Validation curves for loss over 77 epochs for the re-inflection model.

[15] https://github.com/OperaGraecaAdnotata/OGA

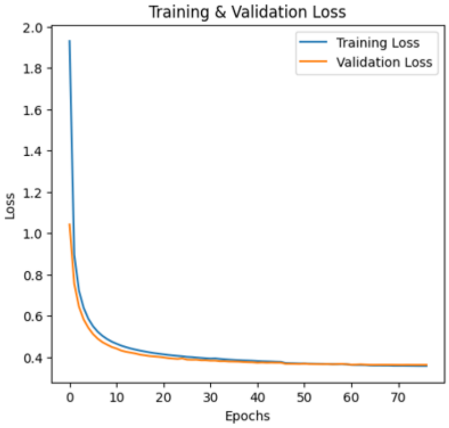


Figure 2: Training and Validation curves for accuracy over 77 epochs for the re-inflection model.

Table 3
Test set metrics for re-inflection model.

	Precision	Recall	F1-score
Accuracy			0.90
Macro avg	0.62	0.52	0.54
Weighted avg	0.90	0.90	0.90

Nonetheless, performance on frequent cases is sufficient to support the generation of realistic paraphrastic samples.

3.2. Contrastive Model Evaluation

To fine-tune the Logion 50k model, we used the HuggingFace SentenceTransformers library, representing each sentence with its [CLS] embedding. The model was trained for 7 epochs, reaching its optimal performance at epoch 6.18. We used the AdamW optimizer with a learning rate of 5e-6 and a weight decay of 0.01.

The contrastive dataset was split into 80% training, 10% validation, and 10% testing, with sentence triplets shuffled prior to the split to ensure distributional uniformity across subsets.

Figures 3 and 4 illustrate the training and validation curves for loss and accuracy, showing a stable convergence pattern.

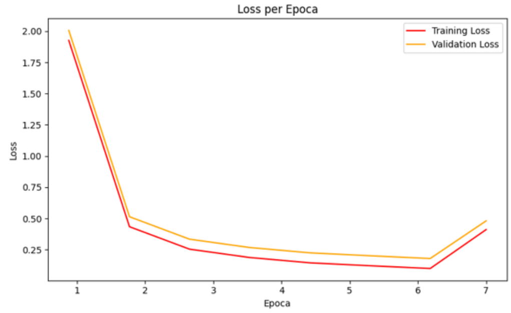


Figure 3: Training and Validation curves for loss over 7 epochs for the contrastive model.

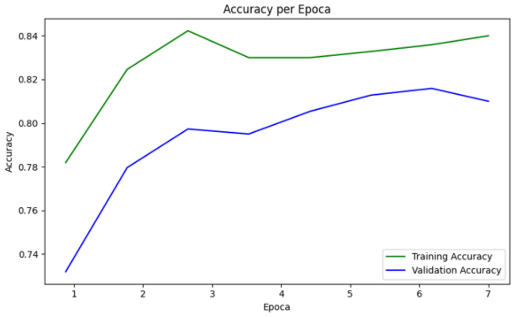


Figure 4: Training and Validation curves for accuracy over 7 epochs for the contrastive model.

The final accuracy on the test set is 0.81, marking a notable improvement over earlier experiments. In a preliminary run using only 5,000 triplets, the model reached an accuracy of 0.71, highlighting its sensitivity to the amount of training data.

Due to the computational complexity of the pipeline used to generate positive and negative samples, we limited the dataset to ~11,000 triplets. However, we hypothesize that a larger dataset—enabled by scaling the paraphrasis and confounder generation—would likely lead to further performance improvements. The model shows strong generalization capabilities despite the relatively limited dataset size.

3.3. Case Study: Homeric Formulae

To evaluate the model's ability to detect semantic reuse, we selected Homeric formulas from the *Odyssey* and retrieved their most similar counterparts from the prose corpora of Herodotus and Thucydides using cosine similarity.

We first performed a general comparison by encoding all sentences from Homer, Herodotus, and Thucydides. For each Homeric sentence, we computed the most similar sentence from both historians. Figure 5 reports how often the most similar match came from each author. Herodotus consistently emerged as the "most Homeric" in style.

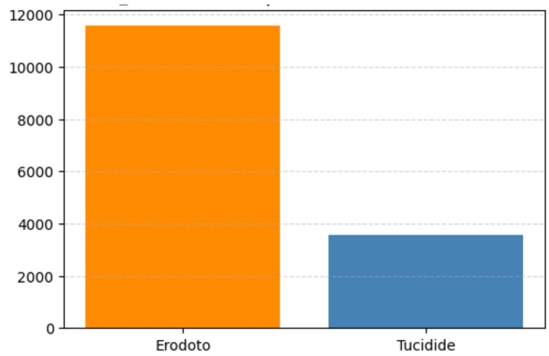


Figure 5: Bar chart showing the number of sentences from Thucydides and Herodotus most similar to Homeric sentences based on cosine similarity.

We then zoom in on the top matches for a handful of Homeric formulas. Table 4 reports the top-3 most similar matches (with cosine similarity) from Herodotus and Thucydides.

In Herodotus, the top match for "ἄσμενοι ἐκ θανάτοιο, φίλους ὀλέσαντες ἑταίρους" ("Glad to have escaped death, having lost dear companions") is: κομισθεὶς ἄρα ἐς τὰς Ἀθήνας ἀπήγγελλε τὸ πάθος (V.87) "Back in Athens, he reported the terrible news." (CosSim: 0.73)

Though the sentences are lexically unrelated, the narrative context aligns: both recount survival from disaster followed by the emotional burden of reporting it. In the Herodotean passage, the warrior coming home is the only survivor: he, too, has "lost dear companions". The model appears to capture these semantic and narrative parallels, ignoring surface forms.

On the other hand, the matching Thucydidean phrase "καὶ τροπαῖον στήσαντες ἀνεχώρησαν ἐς τὸ Ῥήγιον" (IV.25) refers to a commemorated but marginal victory: as noted by (Graves, 1884), the use of fixed epic-like expressions for minimal accomplishments may reflect a form of ironic intertextuality.

Table 4
Top 3 most similar matches for the sentence "ἄσμενοι ἐκ θανάτοιο, φίλους ὀλέσαντες ἑταίρους" (Od. X.134) ("Glad to have escaped death, having lost dear companions") in Thucydides and Herodotus.

Herodotus	Thucydides
κομισθεὶς ἄρα ἐς τὰς Ἀθήνας ἀπήγγελλε τὸ πάθος Transl: Back in Athens, he reported the terrible news. Sim: 0.74	καὶ τροπαῖον στήσαντες ἀνεχώρησαν ἐς τὸ Ῥήγιον. Transl: Erecting a trophy, they withdrew to Reggio. Sim: 0.74
(οὔ τε γὰρ ὕπεστι οίκήματα ὺπὸ γῆν...) Transl: There are no underground dwellings, nor does any canal from the Nile reach it... Sim: 0.73	ού γὰρ ήγγέλθη αύτοῖς ὅτι τεθνηκότες εἶεν. Transl: They had not been told that they were dead. Sim: 0.73
νῦν τε ὅδε έστί. Transl: And here it is now. Sim: 0.72	πολέμιος οὖν ἦν. Transl He was therefore an enemy. Sim: 0.73

Across both historians, the model demonstrates sensitivity to semantic and narrative similarities even in absence of direct verbal overlap. This reinforces the notion that the contrastive objective, paired with linguistically-informed data, enables detection of non-literal textual reuse. Herodotus tends to reuse Homeric motifs to elevate the narrative or align with epic tradition, while Thucydides may repurpose similar forms to subvert or problematize epic conventions.

4. Discussion

The results of our evaluation show that the proposed model is capable of identifying semantic similarity in Ancient Greek texts with a significant degree of accuracy. The performance of the contrastive model—reaching 0.81 accuracy on the test set—suggests that even with a relatively limited dataset, it is possible to fine-tune contextual embeddings for a low-resource language such as Ancient Greek.

Importantly, our qualitative case study demonstrates that the model does not rely solely on lexical overlap, but is able to capture semantic connections grounded in context. This capability is particularly relevant for supporting scholarly analysis of textual relationships, where surface variation and thematic connections require careful interpretation.

Our analysis of Herodotus' proximity to Homer in the similarity distributions aligns with established literary hypotheses about thematic continuity and shared motifs between these authors. However, it is important to note that the semantic similarities detected by our model represent connections that merit further philological investigation rather than definitive instances of literary allusion. The distinction between shared themes, common literary topoi, and intentional intertextual references requires expert scholarly judgment that goes beyond computational analysis.

The matches found in Thucydides, while semantically related to Homeric passages, illustrate this distinction clearly: while our model identifies thematic connections, determining whether these represent ironic reuse, coincidental similarity, or genuine allusion requires deeper interpretive knowledge of the historical and literary context. The contrastive learning objective appears well-suited to identifying such semantic connections as potential candidates for scholarly investigation.

5. Limitations

Our approach faces several important limitations that should be acknowledged:

- Methodological limitations: The generation of paraphrastic and confounding samples, while linguistically motivated, is computationally expensive and depends on the quality of available lexical resources. The method relies heavily on the accuracy of synonym lists from Ancient Greek WordNet and morphological re-inflection models.
- Evaluation constraints: Our evaluation remains primarily qualitative and impressionistic. A more rigorous assessment would require comparison with known allusions identified in scholarly literature, which represents a significant challenge for future work.
- Scope of detection: Our model identifies semantic similarities and thematic connections, but cannot distinguish between coincidental similarity, shared literary tradition, and intentional allusion. This distinction requires expert philological knowledge and cultural context that computational methods cannot currently provide.
- Dataset limitations: The relatively small dataset limits the model's generalizability, and further work is needed to expand coverage across different genres, time periods, and authors to explore cross-genre or diachronic reuse phenomena.

These limitations do not invalidate our approach but rather define its appropriate scope: as a tool for identifying semantically related passages that warrant scholarly attention, rather than as an autonomous detector of literary allusions.

6. Conclusion and Future Work

This paper presented a novel approach to the detection of semantic reuse in Ancient Greek literature through the use of contrastive learning and contextual language models. We developed a pipeline for generating paraphrastic sentence pairs and lexically confounding negatives, enabling the fine-tuning of an encoder model specifically trained for Ancient Greek.

Our method demonstrates the feasibility of identifying thematic connections and semantic relationships in ancient texts, providing a foundation for future work in computational intertextuality detection.

While promising, this system is not meant to replace human judgment. In many cases, interpretation requires close reading and contextual insight that go beyond the scope of automated retrieval. Rather, our model should be seen as an exploratory aid, offering novel perspectives and candidate matches for scholarly validation.

Looking ahead, our goal is to scale the dataset by including larger portions of Herodotean, Thucydidean, and Homeric corpora, and to refine the model further through application to other authors and genres. In particular, we aim to focus on specific thematic domains such as the lexicon of the sacred. Future work should also include more rigorous evaluation against annotated corpora of known literary allusions identified in scholarly literature as well as an evaluation of the paraphrases and confounders by scholarly experts,

Ultimately, this study shows that the intersection of artificial intelligence and philology is not only feasible, but capable of generating innovative and promising contributions to the study of ancient textual reuse.

7. Acknowledgements

We thank Federico Boschetti for sharing annotated corpora and the Ancient Greek WordNet, Sebastian Padò for his contribution to the initial phases of this work, and Barbara Graziosi for her kind willingness to discuss the use of large language models for Ancient Greek.

The implementation code is available from the corresponding author.

References

[1] Barthes, R. (1975). Il Piacere del Testo. Torino: Einaudi.

[2] Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. ACM Computing Surveys, vol. 55, Issue 7, 1-39.

[3] Bizzoni, Y., Boschetti, F., Del Gratta, R., Diakoff, H., Monachini, M., & Crane, G. (2014). The making of Ancient Greek WordNet. LREC 2014. European Language Resources Association ELRA (p. 1140-1147). Paris, France: European language resources association (ELRA).

[4] Boschetti, F. (2019). Semantic Analysis and Thematic Annotation. In M. Berti (Ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution* (pp. 321-340). Berlin, Boston: De Gruyter Saur.

[5] Büchler, M., Burns, P.R., Müller, M., Franzini, E., Franzini, G. (2014). Towards a Historical Text Re-use Detection. In: Biemann, C., Mehler, A. (eds) Text Mining. Theory and Applications of Natural Language Processing. Springer, Cham

[6] Burns, P. J., Brofos, J. A., Li, K., Chaudhuri, P., & Dexter, J. P. (2021). Profiling of Intertextuality in Latin Literature Using Word Embeddings. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (p. 4900-4907). Online: Association for Computational Linguistics.

[7] Celano, G. G. (2024). Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek. ArXiv abs/2404.00739.

[8] Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R., & Jacobson, S. L. (2013). The Tesserae Project: intertextual analysis of Latin poetry. Literary and Linguistics Computing, 221-228.

[9] Cowen-Breen, C., Brooks, C., Haubold, J., & Graziosi, B. (2023). Logion: Machine Learning for Greek Philology. Proceedings of the Ancient Language Processing Workshop (p. 170-178). Varna, Bulgaria: INCOMA Ltd.

[10] Crane, G. R. (1991). Generating and Parsing Classical Greek. Literary and Linguistic Computing, vol. 6, 243-245.

[11] Genette, G. (1982). Palimpsests. Lincoln and London: University of Nebraska Press.

[12] Graves, C. E. (1884). Commentary on Thucydides. London: MacMillan & Company.

[13] Kristeva, J. (1986). Word, Dialogue and Novel. The Kristeva Reader.

[14] Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y., & Buchler, M. (2016). Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (p. 1849-1859). Austin, Texas: Association for Computational Linguistics.

[15] Pranaydeep, S., Rutten, G., & Lefever, E. (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, (p. 128-137).

[16] Riemenschneider, F., & Frank, A. (2023). Exploring Large Language Models for Classical Philology. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (p. 15181-15199). Toronto, Canada: Association for Computational Linguistics.

[17] Riemenschneider, F., & Frank, A. (2023). Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature. Proceedings of the Ancient Language Processing Workshop (p. 30-38). Varna, Bulgaria: INCOMA Ltd.

[18] Riffaterre, M. (1978). Semiotics of poetry. Bloomington and London: Indiana University Press.

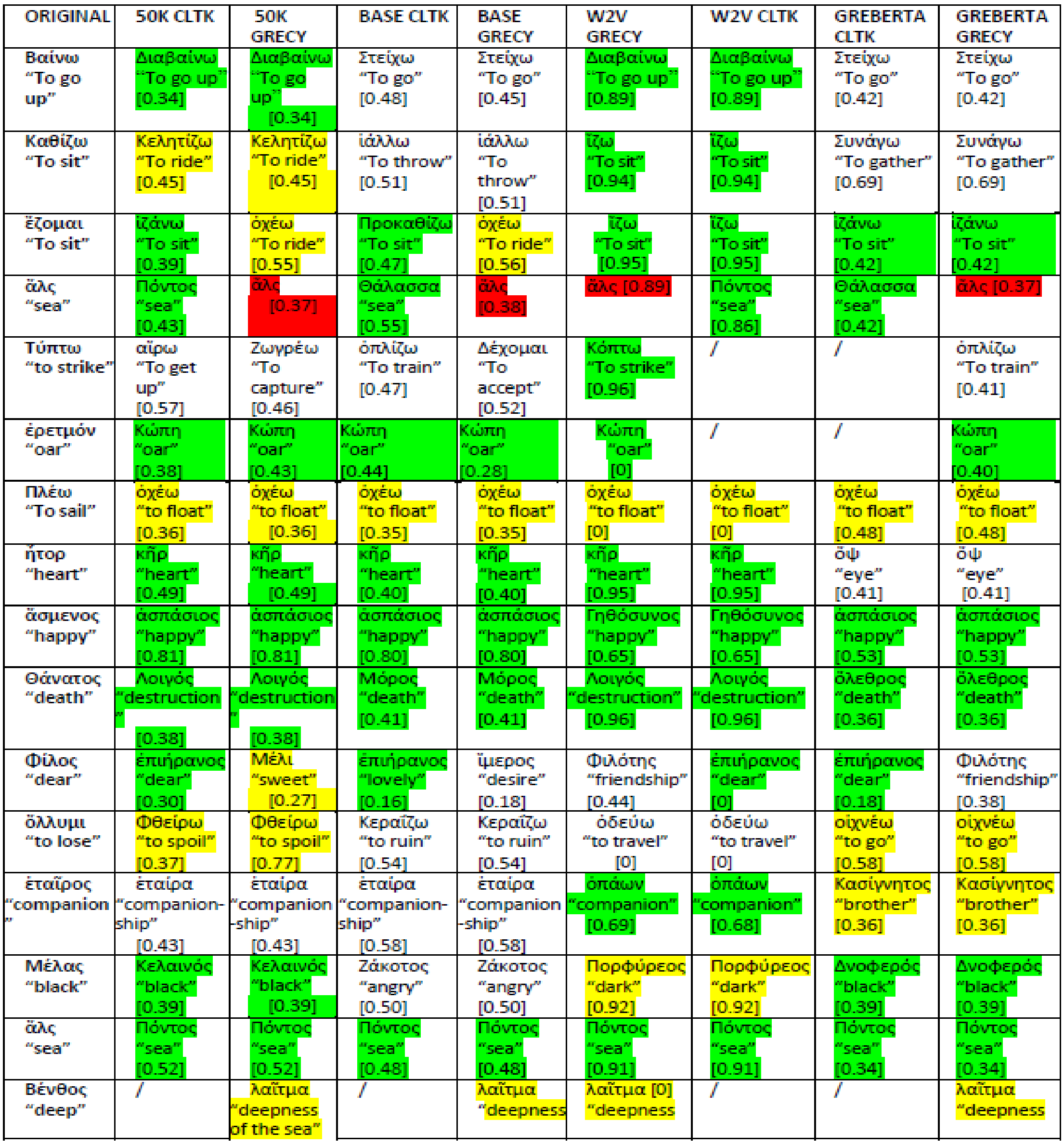
[19] Rodda, M. A., Probert, P., and McGillivray, B. (2019). Vector Space Models of Ancient Greek Word Meaning, and A Case Study on Homer. Traitement Automatique des Langues (TAL). (p. 63-87).

[20] Stopponi, S., Peels-Matthey, S., Nissim, M. (2024). AGREE: a new benchmark for the evaluation of distributional semantic models of ancient Greek. Digital Scholarship in the Humanities. (p. 373-392).

[21] Yamschikov, I. P., Tikhonov, A., Pantis, Y., Schubert, C., & Jurgen, J. (2022). BERT in Plutarch's Shadows. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (p. 6071-6080). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

A. Appendix

In the following image, in green the correct synonyms, in yellow those semantically similar to the original word and in red the results stemming from an incorrect lemmatization, while the synonyms considered wrong are not underlined.



ORIGINAL	50K CLTK	50K GRECY	BASE CLTK	BASE GRECY	W2V GRECY	W2V CLTK	GREBERTA CLTK	GREBERTA GRECY
Βαίνω “To go up”	Διαβαίνω “To go up” [0.34]	Διαβαίνω “To go up” [0.34]	Στείχω “To go” [0.48]	Στείχω “To go” [0.45]	Διαβαίνω “To go up” [0.89]	Διαβαίνω “To go up” [0.89]	Στείχω “To go” [0.42]	Στείχω “To go” [0.42]
Καθίζω “To sit”	Κελητίζω “To ride” [0.45]	Κελητίζω “To ride” [0.45]	ἰάλλω “To throw” [0.51]	ἰάλλω “To throw” [0.51]	ἵζω “To sit” [0.94]	ἵζω “To sit” [0.94]	Συνάγω “To gather” [0.69]	Συνάγω “To gather” [0.69]
ἕζομαι “To sit”	ἱζάνω “To sit” [0.39]	ὀχέω “To ride” [0.55]	Προκαθίζω “To sit” [0.47]	ὀχέω “To ride” [0.56]	ἵζω “To sit” [0.95]	ἵζω “To sit” [0.95]	ἱζάνω “To sit” [0.42]	ἱζάνω “To sit” [0.42]
ἅλς “sea”	Πόντος “sea” [0.43]	ἅλς [0.37]	Θάλασσα “sea” [0.55]	ἅλς [0.38]	ἅλς [0.89]	Πόντος “sea” [0.86]	Θάλασσα “sea” [0.42]	ἅλς [0.37]
Τύπτω “to strike”	αἴρω “To get up” [0.57]	Ζωγρέω “To capture” [0.46]	ὁπλίζω “To train” [0.47]	Δέχομαι “To accept” [0.52]	Κόπτω “To strike” [0.96]	/	/	ὁπλίζω “To train” [0.41]
ἐρετμόν “oar”	Κώπη “oar” [0.38]	Κώπη “oar” [0.43]	Κώπη “oar” [0.44]	Κώπη “oar” [0.28]	Κώπη “oar” [0]	/	/	Κώπη “oar” [0.40]
Πλέω “To sail”	ὀχέω “to float” [0.36]	ὀχέω “to float” [0.36]	ὀχέω “to float” [0.35]	ὀχέω “to float” [0.35]	ὀχέω “to float” [0]	ὀχέω “to float” [0]	ὀχέω “to float” [0.48]	ὀχέω “to float” [0.48]
ἦτορ “heart”	κῆρ “heart” [0.49]	κῆρ “heart” [0.49]	κῆρ “heart” [0.40]	κῆρ “heart” [0.40]	κῆρ “heart” [0.95]	κῆρ “heart” [0.95]	ὄψ “eye” [0.41]	ὄψ “eye” [0.41]
ἄσμενος “happy”	ἀσπάσιος “happy” [0.81]	ἀσπάσιος “happy” [0.81]	ἀσπάσιος “happy” [0.80]	ἀσπάσιος “happy” [0.80]	Γηθόσυνος “happy” [0.65]	Γηθόσυνος “happy” [0.65]	ἀσπάσιος “happy” [0.53]	ἀσπάσιος “happy” [0.53]
Θάνατος “death”	Λοιγός “destruction” [0.38]	Λοιγός “destruction” [0.38]	Μόρος “death” [0.41]	Μόρος “death” [0.41]	Λοιγός “destruction” [0.96]	Λοιγός “destruction” [0.96]	ὄλεθρος “death” [0.36]	ὄλεθρος “death” [0.36]
Φίλος “dear”	ἐπιήρανος “dear” [0.30]	Μέλι “sweet” [0.27]	ἐπιήρανος “lovely” [0.16]	ἵμερος “desire” [0.18]	Φιλότης “friendship” [0.44]	ἐπιήρανος “dear” [0]	ἐπιήρανος “dear” [0.18]	Φιλότης “friendship” [0.38]
ὄλλυμι “to lose”	Φθείρω “to spoil” [0.37]	Φθείρω “to spoil” [0.77]	Κεραΐζω “to ruin” [0.54]	Κεραΐζω “to ruin” [0.54]	ὁδεύω “to travel” [0]	ὁδεύω “to travel” [0]	οἰχνέω “to go” [0.58]	οἰχνέω “to go” [0.58]
ἑταῖρος “companion”	ἑταίρα “companionship” [0.43]	ἑταίρα “companionship” [0.43]	ἑταίρα “companionship” [0.58]	ἑταίρα “companionship” [0.58]	ὀπάων “companion” [0.69]	ὀπάων “companion” [0.68]	Κασίγνητος “brother” [0.36]	Κασίγνητος “brother” [0.36]
Μέλας “black”	Κελαινός “black” [0.39]	Κελαινός “black” [0.39]	Ζάκοτος “angry” [0.50]	Ζάκοτος “angry” [0.50]	Πορφύρεος “dark” [0.92]	Πορφύρεος “dark” [0.92]	Δνοφερός “black” [0.39]	Δνοφερός “black” [0.39]
ἅλς “sea”	Πόντος “sea” [0.52]	Πόντος “sea” [0.52]	Πόντος “sea” [0.48]	Πόντος “sea” [0.48]	Πόντος “sea” [0.91]	Πόντος “sea” [0.91]	Πόντος “sea” [0.34]	Πόντος “sea” [0.34]
Βένθος “deep”	/	λαῖτμα “deepness of the sea”	/	λαῖτμα “deepness	λαῖτμα [0] “deepness	/	/	λαῖτμα “deepness