The Evolution of Distributional Semantics for Ancient Greek: From Vector Models to Transformer-Based Approaches – State of the Art and Perspectives on Statement Retrieval

Federico Iezzi*†
Università di Modena e Reggio Emilia

Elia Scapini*‡
Università di Modena e Reggio Emilia

Transformer-based Pre-trained Language Models (PLMs) for Ancient Greek are rapidly advancing in both number and performance. Among the opportunities enabled by these models is semantic statement retrieval, but significant efforts to refine the technique and make it accessible to the scholarly community are still in their early stages. This survey reviews advancements in computational approaches to Ancient Greek semantics, describing the transition from traditional vector-based models to state-of-the-art transformer-based architectures, with a particular focus on their application to statement retrieval. It examines advantages and limitations of applying transformer-based PLMs to tasks such as the study of semantic change, lexical analysis, and statement retrieval. Although these models provide innovative methodologies for addressing semantics in the analysis of Ancient Greek sources, challenges such as limited training data, the absence of shared evaluation standards, and the "black box" nature of deep learning remain significant obstacles. The paper highlights the potential of these technologies to enhance digital scholarship on ancient texts and calls for further refinement and benchmarking to overcome existing limitations.

1. Introduction

1.1 Purpose

This paper aims to explore the application of computational linguistics techniques to semantic tasks in Ancient Greek, presenting an investigation of the state of the art in this field, with a particular focus on PLMs based on the transformer architecture. Although these models are inherently multitasking and thus adaptable to various Natural Language Processing (NLP) tasks, their application to semantic analyses of Ancient Greek texts remains underexplored. By critically examining existing transformer-based PLMs, this study also aims to observe their current capabilities and explore their potential to contribute to semantic research in the context of Ancient Greek.

*These authors contributed equally.
†Dipartimento di Educazione e Scienze Umane (DESU) - Viale Timavo 93, 42121 Reggio nell'Emilia RE, Italy. E-mail: federico.iezzi@unimore.it. (https://orcid.org/0009-0005-5488-1426).
‡Dipartimento di Educazione e Scienze Umane (DESU) - Viale Timavo 93, 42121 Reggio nell'Emilia RE, Italy. E-mail: elia.scapini@unimore.it. (https://orcid.org/0009-0001-6698-0457).

1.2 Structure

We focus here on those PLMs that are deep, multi-layered neural networks specifically trained for language understanding tasks. These models are generally classified as Large Language Models (LLMs) when they exceed 6–10 billion parameters (Liu et al. 2025). Depending on the specific language processing tasks, these models can employ either an encoder-only architecture or a combined encoder-decoder architecture[1]. According to this standard, we can outline the structure of our contribution as follows. After the introduction (sec. 1), we describe the rationale behind distributional semantics and sentence embedding, showing how it has led to the development of Vector Space Models (VSMs) for Ancient Greek (sec. 2). Then, we illustrate the main Distributional Models in Semantic Scholarship on Ancient Greek (sec. 3) firstly numbering context-count, context-predict and bayesian projects, switching secondly to transformer-based models and finally opening the issue of the benchmark. We then discuss advantages and limitations of statement retrieval performed with cosine similarity (sec. 4). We conclude by cautiously suggesting that transformer-based models and statement retrieval using cosine similarity, despite their limited transparency to human users, still provide valuable tools for scholarly work.

2. Distributional Semantics and Semantic Embedding: from Theory to VSMs

The core principle that semantically related words tend to co-occur in shared contexts inspires all the techniques of distributional semantics. This discipline aims to represent the relatedness of terms through spatial proximity, where geometric closeness reflects their semantic similarity. Such representations enable NLP systems to tackle a variety of tasks. "Similarity" must be understood in the broadest sense possible; it encompasses not only synonymy and antonymy but also the broader, self-implicating relationships between words. The rationale behind this approach is to consider terms not as isolated units but as dense tensors. This idea traces back to the principle famously summarized by Firth (1957) as «you shall know a word by the company it keeps» and to the view of Harris (1954) that language is organized within a distributional structure[2].

Efforts to apply this principle to concrete algorithms for measuring semantic similarity often lead to the use of vectors, matrices, and higher-order tensors. According to Turney and Pantel (2010)[3], the intimate connection between the distributional hypothesis and VSMs was a strong motivation to closely examine VSMs. In their survey, they classified previous work on VSMs based on the type of matrix involved: term–document, word–context, and pair–pattern. Although these three types of matrix cover most of the existing work, Turney and Pantel did not believe they exhausted the possibilities. They anticipated that future research would introduce new types of matrices and higher-order tensors, expanding the potential applications of distributional semantics. They concluded their paper by openly stating that they expected the introduction of new types of matrices and more complex tensors to open up new applications for VSMs. In

[1]Since this review focuses only on encoders and does not include attempts to apply GPT-like models to ancient texts we will not discuss solutions recently proposed such as the `https://humanitext.ai/` project introduced by Iwata, Tanaka, and Ogawa (2024) .

[2]Quoting Firth and Harris has become almost a ritual practice. For a deeper discussion and comparison of their views, see Brunila and LaViolette (2022), as well as the additional references provided in their bibliography.

[3]For a comprehensive overview of the development of VSMs for semantic purposes, from their origins in the SMART information retrieval system (Salton 1971) to 2010, we refer the reader to the survey by Turney and Pantel (2010).

this sense, their prediction was partially realized. In fact, just a few years after the publication of their work, one of the most promising techniques in the field of distributional semantics was developed: embeddings. With "embedding", a technique is meant that transforms words (and sentences) into multi-dimensional vectors, statistically encoding information about grammar, morphology, syntax and context in a tensor made of a fixed quantity of real numbers. Embeddings are usually distinguished between "static" and "dynamic"[4].

Static embeddings, either calculated with context-count models or predicted with neural networks (Baroni, Dinu, and Kruszewski 2014), are unequivocally assigned to a vocabulary word, regardless of its contextual behavior. These can be generated using tools like Word2Vec proposed by Mikolov et al. (2013a, 2013b). Word2Vec is a neural network-based model that maps words to a continuous vector space, where semantically similar words are represented by similar vectors. The two main architectures of Word2Vec are as follows:

- Skip-gram, which predicts the context words given a target word;
- CBOW (Continuous Bag of Words), which predicts a target word given its surrounding context words.

This ability to "predict" often allows the model to outperform so-called context-count models in many NLP tasks, particularly those related to semantics (Baroni, Dinu, and Kruszewski 2014). However, this is not always the case, as shown by Lenci et al. (2022), who provide a comparative evaluation of approaches. For Ancient Greek specifically, see also Stopponi et al. (2023) for an assessment of distributional semantic models. It is also true that, the outputs of context-count models remain more interpretable and transparent (Keersmaekers and Speelman 2023). Word2Vec has nonetheless revolutionized computational linguistics by enabling highly precise word embeddings in large corpora and facilitating the use of machine learning models for complex language tasks.

The year after the launch of Word2Vec, Stanford researchers introduced GloVe (Pennington, Socher, and Manning 2014). GloVe, which stands for Global Vectors for Word Representation, is similar to Word2Vec but relies on a word co-occurrence matrix rather than a neural network. This matrix captures how words co-occur within a corpus of text, making GloVe particularly effective at encoding global information about the relationships between words.

In 2016, FastText (Bojanowski et al. 2017), a model that represents words as sequences of character subsets (*n*-grams), was introduced. This approach enables FastText to handle rare or unknown words (out-of-vocabulary) more effectively by generating embeddings for unseen words based on their constituent parts. Thanks to this capability, FastText has significantly influenced the development of models for languages with complex morphologies or numerous compound words.

Dynamic embeddings, on the other hand, provide different vector representations for the same word depending on its context and are most effectively generated using PLMs based on Deep Neural Networks. For this reason, dynamic embeddings are often referred to as "contextual" embeddings. These embeddings are closely associated with the

[4] An overview on the history of the development of these tools to state-of-the-art transformers is provided by Russell and Norving (2021). An introduction to vector semantics specifically designed for an humanist audience is also offered by Gavin (2018). More recently, Barbara McGillivray (2022) provided a useful introduction to distributional semantics.

transformer architecture introduced in *Attention is All You Need* (Vaswani et al. 2017). The transformer employs self-attention mechanisms that allow it to evaluate the importance of each word in relation to others, regardless of their order in a sentence. This represents a major advancement over older models like RNNs (Recurrent Neural Networks) and LSTMs (Long short-term memory), which process words sequentially and struggle with long-range dependencies. By processing words in parallel, the transformer produces rich, context-sensitive word representations. This architecture, first implemented by Devlin et al. (2019) in models like Bidirectional Encoder Representations from Transformers (BERT), has been pivotal for generating dynamic embeddings that adapt to surrounding words, significantly outperforming traditional word embeddings such as Word2Vec on many NLP tasks. Moreover, these models have advanced not only general language tasks but also research in the study of ancient languages. The implications of these developments for semantic tasks in Ancient Greek language processing will be examined in the following sections.

3. Distributional Models in Semantic Scholarship on Ancient Greek [5]

3.1 Context-count, Context-predict and Bayesian models

It is interesting to note that the first use of NLP for Ancient Greek involved a Part-of-Speech (PoS) tagger created to assist students in learning the language (Packard 1973). Only later was the potential of computational resources for advanced research recognized, with applications extending to historical linguistics and philology. In any case, it should be noted that research applying and investigating semantics in this area remains relatively scarce, leaving many opportunities for further exploration within computational linguistics. Some studies have focused on the identification of hidden topics within a corpus of documents using statistical models like LDA (Latent Dirichlet Allocation)[6]. However, the primary application of distributional models in the semantic field for Ancient Greek has been the study of semantic change in words[7].

[5] Although this section primarily focuses on studies related to distributional semantics for Ancient Greek, it is important to highlight the existence of a significant tool with an ontology-based structure: WordNets. The WordNet architecture, introduced for the English language by Fellbaum (1998) and adapted to ancient languages by Biagetti, Zanchi, and Short (2021) for Sanskrit, Ancient Greek, and Latin, builds on earlier attempts in the field, such as those by Bizzoni et al. (2014) and Boschetti (2019) for Ancient Greek, and Minozzi (2009) for Latin. Biagetti, Zanchi, and Short (2021) expanded these works by integrating the framework with theories of meaning from Cognitive Linguistics. WordNets are lexical databases that organize meanings relationally. Lemmas are represented as nodes associated with one or more synsets (sets of cognitive synonyms with short definitions) and are connected through lexical and semantic relationships, forming networks of words and concepts interlinked by meaning. Although the challenge of adapting concepts and methodologies from modern languages to ancient languages remains unresolved (Biagetti et al. 2024), WordNets offer a robust representation of polysemy in Indo-European languages and fosters research of interlinked meanings (Mambrini et al. 2021). A notable limitation, however, is that creating WordNet requires extensive manual work on large corpora by annotators, who must follow standard labeling for the representation of meanings.

[6] Wishart and Prokopidis (2017) adapted a PoS tagger and lemmatizer for Hellenistic Greek and analyzed the texts using an LDA topic modeling to identify the most significant words for each topic. The paper highlights the importance of lemmatization for automated semantic analyses, as errors in the automatic lemmatizer can confuse the LDA model, dispersing semantically related terms across different topics, thereby necessitating human intervention for better results. Similarly, Köntges (2020a) applied an LDA model to philosophical texts in Ancient Greek. Based on the identified topics, the author distilled three numerical scores for these texts: one measuring the concept of "good and virtue", another measuring "scientific inquiry", and a third, which combined the two, assessing the degree of "philosophicalness" in a given text.

[7] However, it is important to highlight the study by Keersmaekers (2020), which integrated the use of word vectors not for semantic change, but for the task of semantic role labeling. Moreover, he recognizes the vectors as «the most helpful features» for the supervised learning model employed.

To our knowledge, the first attempt to apply distributional semantics to Ancient Greek (specifically using a context-count model) for the study of semantic changes in the language over time was made by Boschetti (2010)[8]. In the third part of his doctoral thesis, the author focused on the study of semantic spaces. Using a varied corpus, both in terms of genre and chronology (8th century B.C. – 15th century A.D.), Boschetti examined the meanings of words across different literary genres and observed semantic changes over time through the calculation of cosine distance between the vectors associated with words. His study deserves recognition for pioneering the application of distributional semantics to Ancient Greek. However, the considerable variety of literary genres in the corpus introduces a significant amount of noise. Similarly, Rodda, Lenci, and Senaldi (2017) employed distributional models that utilized Positive Pointwise Mutual Information (PMI) to analyze the semantic lexical change in Ancient Greek. Their study used a smaller corpus than Boschetti's (7th century B.C. – 5th century A.D.), divided into two sub-corpora: one representing the pre-Christian era and the other the Christian era. Applying representational similarity analysis (RSA) to measure semantic changes, the study identified the extent to which certain terms shifted in meaning over time, shedding light on the influence of Christianity on the lexicon. Also the doctoral thesis of Grewcock (2018) falls within this framework, aiming – through the use of a context-count model – to identify patterns, structures and variations in the syntax and semantics of movement verbs. On the other hand, Homeric formulas were the focus of the analysis of Rodda, Probert, and McGillivray (2019). This paper described the creation of a vector space model and its validation against a new benchmark[9] introduced by the authors. The benchmarks were derived from scholarship from the ancient world, modern lexicography, and an NLP resource. Finally, to demonstrate the applicability of the model, authors use the model for a study on epic formulae to track semantic variation.

On the other hand, regarding context-predict VSMs, Burns (2019) must be recognized as a forerunner. He presented an experimental model based on Word2Vec in version 0.1 of the CLTK (Classical Language Toolkit), which would later merge into version 1.0 of the CLTK (Johnson et al. 2021), offering a standard API (application programming interface) and pre-configured processing pipelines for several ancient languages. Another example of the application of Word2Vec is List (2022), who used vector spaces to examine kinship terms, demonstrating how the distributional analysis of words can contribute to the creation of lexical (dictionary) entries by suggesting relevant meaning associations to the target word. The author tests the method on three kinship terms. The embedding technology, along with other resources, plays also a central role in the recent project funded by CLARIN *A New CLARIN Resource Family for Lexical Semantic Change Research*[10] (McGillivray, Khan, and Marongiu 2023). The project's aim is to centralize essential resources for lexical semantic change research. These resources include datasets annotating word meanings, word embeddings derived from diachronic corpora, automated algorithms to detect semantic changes and lexical resources. In January 2024, Marongiu, McGillivray, and Khan (2024) published a paper describing a workflow example for the application of the tools provided by CLARIN to English, Ancient Greek and Latin. A few months later, Stopponi et al. (2024) stated that they were following a workflow that largely overlaps with that defined by Marongiu and

[8]The writing of the thesis was completed in 2009, but it was discussed in 2010. For this reason, some refer bibliographically to this study as Boschetti (2009).

[9]https://zenodo.org/records/3552763#.YfAItOrMKWA

[10]https://www.clarin.eu/blog/workflows-semantic-change-research-clarin-resource-families.

her colleagues. First, they divided the Diorisis corpus (Vatri and McGillivray 2018) into five time windows. They then applied two measures of lexical semantic change detection to Word2Vec embeddings trained on the selected corpus: VC (Vector Coherence) and J (Jaccard). The analysis revealed that both measures were effective at identifying stable words, with Vector Coherence being more reliable at detecting changed words. However, a low Jaccard score also indicated semantic change when combined with low Vector Coherence. Both measures were affected by lemmatization errors in the training corpus, which hindered the detection of changed words[11]. Other attempts to study semantic change in Ancient Greek have been made using Bayesian models, which, in some papers, intersect with embedding technology. Perrone et al. (2019) developed GASC (Genre-Aware Semantic Change), a Bayesian model that uses metadata about the genre of texts to improve the inference of semantic change in Ancient Greek corpora. This approach helped to distinguish between polysemy and semantic change. McGillivray et al. (2019)[12], starting from the paper just described, applied the SCAN model – originally developed for English by Frermann and Lapata (2016) – to Ancient Greek, comparing it with the GASC model, which performed better. The authors, in conclusion of their paper, underlined the difficulty of the challenges related to the evaluation of computational models, particularly when comparing automatic analyses with those of experts. The paper by Perrone et al. (2019) is explicitly considered the starting point for the study of Zafar and Nicholls (2022), which further refined this framework by integrating concepts of sense and time as additive effects into their DiSC model (Diachronic Sense Change), thus achieving a significant improvement in the performance of the model. Two years later, the same authors presented EDiSC, a model that extends DiSC, combining it with word embeddings to improve the analysis of sense change (Zafar and Nicholls 2024). The main novelty consists in having integrated topic-based models and embedding-based models to optimize performance. As a final example of the use of a Bayesian model – which in this case was applied to both Greek and Latin – we report the study of Perrone et al. (2021), in which GASC was compared with embedding-based models.

The authors of this latest work showed that Bayesian models offer explicit and easy-to-interpret representations, while embedding-based methods, despite being extremely sophisticated and widely used, are less transparent. As we will discuss later, interpretability remains a pivotal feature to negotiate with when evaluating which methodology best suits the scholar's research purpose.

3.2 Transformer-based Models for Ancient Greek

Being it a language that is more computationally studied, provided with the largest textual corpus available and currently associated with a significantly larger number of resources, English naturally served as the basis for the development of transformer-based models. Techniques and solutions were often adapted to Ancient Greek[13]. As

[11] The authors' work has produced the software AGALMA (Ancient Greek Accessible Language Models for linguistic Analysis), accessible at the following link: `https://huggingface.co/spaces/GroNLP/agalma`.

[12] Some of the authors of this paper have created a collection containing five objects: two Python scripts and three datasets (Vatri, McGillivray, and Lähteenoja 2019). (https://doi.org/10.6084/m9.figshare.c.4445420). The dataset is based on the Diorisis corpus (Vatri and McGillivray 2018) and contains the manual annotation - of a semantic nature - of sentences in which the words are needed *mus*, *harmonia* and *kosmos*.

[13] This is a natural process in the field of digital humanities, which often adapts techniques and models originally developed for business applications to the needs of humanistic research (Thaller 1987). It is also worth noting that, although our investigation focuses exclusively on the semantics of Ancient Greek, advances

a result, there is a physiological latency period that separates the appearance of tools designed for English and the correspondent attempt to apply the same technique to ancient languages, including Ancient Greek[14]. Indeed, there was a three-year gap between the release of the first BERT (Devlin et al. 2019) and the first Ancient Greek BERT (AGB) model, released by Singh, Rutten, and Lefever (2021). Between the two models, a key intermediate step was represented by a Modern Greek BERT model (Koutsikakis et al. 2020). This tool has to be mentioned because, before training it with data from corpora such as First1KGreek Project, Perseus Digital Library, PROIEL and fine-tuning it to perform PoS and morphological tagging, Singh, Rutten, and Lefever (2021) based their model on that of Koutsikakis et al. (2020). As BERT models, they share several common features: their base structure consists of twelve transformers layers, each with its attention head[15]. The output representation they provide is a vector with 768 dimensions[16]. One distinctive aspect of their operation is the use of special tokens, among which the most relevant for the purpose of this review is the [CLS] (Classification) token. This token is added at the beginning of each input sentence and serves as a context aggregator for the entire sequence. The output vector corresponding to each [CLS] token can be used for various tasks, having captured relevant information from the entire sequence.

The first BERT model by Devlin et al. (2019) was initially released in both an UNCASED and a CASED version. In the former, the corpus was fully converted to lowercase while in the latter uppercase letters were taken into account by the model. In contrast, the initial AGB models did not include this distinction. In fact, not only did the AGB model inherit its lack of diacritics from the Modern Greek model (Koutsikakis et al. 2020) on which it was based, but the same approach was also adopted by the models released by Yamshchikov et al. (2022) and Spanopoulos (2022).

Yamshchikov et al. (2022) released a BERT model for Ancient Greek specifically optimized to predict authorship attribution, with a validation accuracy of around 80%. As in Singh, Rutten, and Lefever (2021), the model was not created from scratch for Ancient Greek but was derived from Modern Greek BERT[17] and fine-tuned using transfer learning on a Masked Language Modeling (MLM) task. The goal in this case was not PoS tagging, but rather to clarify the attribution of three documents to Pseudo-Plutarch[18]. Moreover, this model was initialized with Modern Greek (Koutsikakis et al. 2020), which raises the issue of the omission of certain diacritics in Ancient Greek. Indeed, authorship attribution, like statement retrieval, is not the task for which these

in AI and ML have enabled analyses on a scale and with a level of detail that are reshaping the field of humanities, facilitating various NLP tasks across many ancient languages. For a state-of-the-art overview of other ancient languages, see the survey by Sommerschield et al. (2023).

[14] Appendix A provides a summary table of the models discussed in this section.

[15] It is common to consider the first layers as representing lower-level information, particularly related to grammar, morphology, and syntax, while higher layers are viewed as corresponding to more semantic information (Jawahar, Sagot, and Seddah 2019).

[16] Dimensions raise to 1024 in larger BERT models.

[17] `https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1`.

[18] Yamshchikov et al. (2022) attempted to overcome the so-called "quantitative authorship attribution" approach (Grieve 2007). Following this same approach, Manousakis and Stamatatos (2018) used an SVM classifier to analyze the character *n*-grams of Ancient Greek texts, with the aim of detecting authorial variability in the tragedy *Rhesus*. Similarly, Köntges (2020b) combined the analysis of *n*-grams with philological arguments to examine the attribution of the *Menexenus*, concluding that the *Menexenus* has not been written by Plato. Likewise, Pavlopoulos and Konstantinidou (2023) used character-level statistical language models to analyze the linguistic proximity between each book of the *Iliad* and the *Odyssey* with all other books of the same poem. They further examined correlations between pairs of books, assessing how strongly each book is associated with every other book in the corresponding poem, in order to identify outlying or atypical passages.

models were initially designed, nor is it the one where they are expected to perform best. In fact, among the tasks for which BERT encoders were originally designed, 'fill in the gap' stands out. Typically, training data are divided into a training set and a validation set, with an approximate ratio of 80% to 20%. The validation corpus is then randomly masked by replacing some words with an anonymous tag (e.g., [MASK]), which the model has the task to predict. This technique is known as MLM. A number of training epochs are then iterated to adjust the internal weights of the neural network in order to improve the accuracy of the model's representations.

We have mentioned the model from Spanopoulos (2022): this is a RoBERTa (Robustly Optimized BERT Pre-training Approach) model that was released one year after the first AGB. Similarly, the first RoBERTa model was released around one year after BERT. RoBERTa is an encoder-only PLM first introduced by Liu et al. (2019). RoBERTa and BERT share a similar structure; however, the former often outperforms the latter due to specific adjustments in the training process, including dynamic masking and the removal of the next sentence prediction (NSP) task. Disclosing the performance of their model, Spanopoulos noted that although it performed reasonably well on PoS tagging, it showed signs of underfitting, likely caused by insufficient training data[19]. The lack of an adequate amount of training and validation data is one of the key challenges in developing state-of-the-art transformer-based models. This issue is particularly pronounced in the case of Ancient Greek, since the available corpus is relatively small and confined to a fixed body of texts, with no possibility of new material being produced. Furthermore, Ancient Greek poses additional difficulties due to its high degree of inflection, its considerable diachronic variation, and its substantial genre variation (Perrone et al. 2019, 2021).

Yousef et al. (2022a, 2022b) published an XLM-RoBERTa-based model optimized for multilingual text alignment. This model was developed using aligned material from UGARIT[20] (Palladino, Foradi, and Yousef 2021), a translation alignment editor. The development was informed by observations on the application of translation alignment in ancient language courses conducted at Tufts University and Furman University between 2017 and 2019. With this work, the acronym XLM, which stands for "Cross-Lingual Model", entered the field of transformer-based models applied to Ancient Greek. In fact, models can also be trained in multiple languages (Artetxe and Schwenk 2019; Lample and Conneau 2019), opening up new possibilities thanks to their ability to

[19] Access to digitized and interconnected data related to historical languages plays a crucial role in driving advances in machine learning. Ancient Greek, in particular, poses significant challenges in this area. Although a substantial amount of material is available, two main problems persist: (a) not all resources are freely accessible—most notably the bility of data, resources remain scarce. The *Thesaurus Linguae Graecae* (TLG), the largest literary corpus (110M tokens), which requires a paid subscription and does not allow the download of texts for direct reuse in Machine Learning projects—and (b) in comparison with some modern languages, the overall size of Ancient Greek corpora remains relatively small. In terms of numbers, the freely available corpus *Opera Graeca Adnotata* (Celano 2024) contains about 40M tokens, while GLAUx contains 20M (Keersmaekers 2021). Nevertheless, these resources are confined to the literary domain (as noted in footnote 31). Large quantities of non-literary material also exist—such as papyri, inscriptions, or lamellae—but with limitations that still represent a barrier to the large-scale and democratic application of machine learning technologies to Ancient Greek. However, competitions and conferences have stimulated the creation of specialized datasets that have fueled the development of advanced models, such as those based on transformer architectures. These models have surpassed traditional techniques in terms of ability and precision, reaching milestones previously inaccessible without human input (Sommerschield et al. 2023). Despite this, the absence of freely accessible resources for languages such as Ancient Greek constitutes a barrier to the widespread and large-scale use of these technological innovations, slowing down the process of democratization of these technologies.

[20] `https://ugarit.ialigner.com/`.

represent concepts in a language-agnostic way and to transfer knowledge, adapting it to new tasks. This is particularly interesting because multilingual models sometimes outperform single-language models in zero-shot configuration. This means that if a model is pre-trained to perform a specific downstream task in one language, it can adapt to a second language without requiring fine-tuning to accomplish the same task. For example, a model trained in English and fine-tuned on a downstream task for that language may, after multilingual adaptation, outperform a native Ancient Greek model on the same task. Following m(ultilingual)BERT, which supports more than a hundred languages, Lample and Conneau (2019) introduced Translation Language Modeling (TLM) and proposed training a XLM model. Later, other models such as XLM-R(oBERTa) (Conneau et al. 2020) were developed based on this architecture. The XLM-RoBERTa-based model by Yousef et al. (2022a) was created with the assumption that multiple fine-tuning steps can enhance the model's performance on a downstream target. More specifically, they have fine-tuned multilingual model XLM-R on 12M Ancient Greek tokens with MLM and subsequently tailored the improved model on language pairs taken from UGARIT. These models, while designed for alignment and translation, can also be used for other NLP tasks, such as improving Named Entity Recognition (NER), as shown by Yousef, Palladino, and Shamsian (2023) and Yousef (2023).

Since PLMs are trained to model patterns in the use of a language, they can be exploited to suggest how to complete missing portions of text[21]. This approach has been applied to both inscriptions and manuscripts. Assael, Sommerschield, and Prag (2019), Assael et al. (2022), within the framework of the PithiaPlus project[22], worked on a tool designed to restore gaps in inscriptions caused by physical damages. The first (LSTM-based) tool named PYTHIA (2019) was capable of predicting how to fill missing parts of inscription. It was then followed by ITHACA (2022)[23], a model that leverages a transformer-based architecture to restore inscriptions and predict their location and date. Not only was the training performed on the largest pre-modern Greek corpus to date, but, interestingly, the authors also trained ITHACA with augmented data, i.e. by generating multiple artificial variants of the same inscription through segmentation and masking. This procedure increased the number of training examples and simulated the fragmentary nature of the sources, allowing the model to improve its ability to restore damaged texts. Serving as a bridge between ITHACA and PYTHIA was the Blank Language Model (BLM) by Shen et al. (2020), a transformer-based architecture evaluated on the dataset by Assael, Sommerschield, and Prag (2019). This model achieved similar performance to PYTHIA but stood out for its ability to generate arbitrary sequences without needing to define the target length. By the side of manuscripts, Logion BERT, announced in the concept-paper of Graziosi et al. (2023) and released by Cowen-Breen et al. (2023), was specifically customized for Greek philology. This BERT model was trained to perform a fill-gap task on errors in a long hand-copy transmission queue to achieve correction of scribal errors in manual transcription. It should be noted that none of the studies reported in this paragraph evaluated their model for semantic purposes.

[21] For the sake of completeness in our survey, we point out the existence of the Ancient-Greek-Char-Bert project (`https://github.com/brennannicholson/ancient-greek-char-bert`), developed as part of a Bachelor's thesis at the University of Leipzig. The project uses artificial intelligence models based on BERT technology, operating at the character level, to predict missing symbols in damaged Ancient Greek texts. Built with the FARM framework, it aims to serve as a tool for the restoration of Ancient Greek texts. We relegate it to a footnote as it is currently not supported by a scientific publication.

[22] `https://pric.unive.it/projects/pythiaplus/home`.

[23] `https://ithaca.deepmind.com/`.

Up to this point, it is important to note that no transformer-based model among those discussed above has specifically been created from scratch for Ancient Greek. The first model trained directly on Ancient Greek was introduced by Riemenschneider and Frank (2023a) from the Computational Linguistics department of the University of Heidelberg. Their work presented four PLMs for Classics: two RoBERTa encoder-only models, GrεBERTa and PhilBERTa, and two T5 encoder-decoder models, GrεTa and PhilTa. It is the first time that we see the construction of T5 models for Ancient Greek. Raffel et al. (2020) firstly suggested this technique: aiming to adapt one model to several tasks, authors introduced Text-to-text Transfer transformer architecture (5 "T"s). These models consist of an encoder-decoder structure so that they can create intermediate representations of the input before sending it to the decoder that generates the output. This approach enables the transformation of each task into a text-to-text problem, giving generalization capacity to the model. Riemenschneider and Frank's GrεBERTa and GrεTa are, respectively, encoder-only RoBERTa and T5 encoder-decoder trained exclusively on Ancient Greek texts. In contrast, PhilBERTa and PhilTa are the trilingual counterparts of the previous models, pre-trained on datasets in Greek, Latin and English. Since all these models were trained on a significantly larger corpus of Ancient and Medieval Greek (185.1 million tokens; see Table 1) compared to previous BERT models, and incorporate a text processing system that retains diacritics, the result is that GrεBERTa outperformed all existing models in morphological and syntactic tasks.

Table 1
The four models trained by Riemenschneider and Frank.

	Greek	*Multilingual*
Encoder only	GrεBERTa	PhilBERTa
Encoder-decoder	GrεTa	PhilTa

In the same year, Riemenschneider and Frank (2023b) presented SPhilBERTa, a cross-trilingual model designed for Classical Philology, "distilled" from a multilingual model called all-mpnet-base-v2. Here, two new concepts are introduced: the distillation technique and the sentence model. The first term, distillation, refers to a method for adapting a model trained on a highly attested language (e.g., English) to work on less-resourced languages (e.g. Ancient Greek and Latin). This technique, described by Reimers and Gurevych (2020), involves transferring knowledge from the original model to improve performance in languages with limited data. This method operates teaching a "learning" model to assign the same vector representations to a sentence in the target language as those assigned by a "teaching" model to the same sentence in the more attested language. To achieve this, authors need to compose (and sometimes synthesize) a dataset of parallel sentences in English and Ancient Greek, and set the learning function to implement sharing of representation. Precisely because the smaller model is trained to replicate the predictions of the large model, this method has proven particularly useful for languages with limited available data. Regarding the second concept (sentence PLMs), BERT models were primarily trained to encode individual words after tokenizing them. Although it is possible to represent a sentence by summing or averaging its word vectors, or by leveraging the [CLS] token, this approach often performs poorly, especially for sentence retrieval tasks in languages with limited available data, such as Ancient Greek. To address this limitation, more recent BERT models have been enhanced to perform sentence embedding, as first suggested by

Reimers and Gurevych (2019). Sentence-BERT (sBERT) models[24] integrate a siamese and triple network structure into the BERT architecture, making them more efficient and less time consuming for sentence embedding tasks, such as similarity, clustering and information retrieval. SPhilBERTa, designed for Classical Philology, has been specifically tailored for cross-linguistic semantic understanding and the identification of identical sentences across Ancient Greek, Latin and English. Riemenschneider and Frank (2023b) also presented a case study on the *Aeneid* and the *Odyssey*, demonstrating the ability of SPhilBERTa to facilitate the automated detection of intertextual parallelisms[25].

In the same year, another sentence transformer-based model for Ancient Greek, SHLM-grc-en, was introduced by Krahn, Tate, and Lamicela (2023). The accompanying paper describes their distillation method, which involves adapting a larger multilingual model. They created aligned vector spaces for Ancient Greek and English, following the alignment approach introduced by Liu and Zhu (2023), creators of Bertalign. The authors then evaluated their models on STS tasks (Semantic Textual Similarity) and SR (Semantic Retrieval).

Finally, Riemenschneider and Krahn (2024) jointly submitted a paper for the SIGTYP conference (Special Interest Group on Typology), where they presented a model that combines the hierarchical tokenization technique suggested by Sun et al. (2023) with a DeBERTa-V3 model (He, Gao, and Chen 2023). The core elements combined here that require to be introduced are two: HLM and DeBERTa-V3. We will begin detailing the latter.

After the introduction of BERT and RoBERTa, many other models can be numbered: the architecture of DeBERTa suggested by He et al. (2021) improved RoBERTa models adding in the pre-training process the disentangled attention to improve the relative-position understanding in the model. Here, two separate vectors provide respectively the content and the position of the represented word. Still, it is pre-trained with MLM. Proposing ELECTRA, Clark et al. (2020) suggested another pre-training technique: Replaced Token Detection (RTD). Here, ambiguous corruptions in the tokens are automatically generated and the model is pushed to spot them. Therefore, ELECTRA models are trained with two transformer models (Generative Adversarial Network - GAN style), a generator that creates ambiguous tokens (MLM) and a discriminator that has to predict whether the tokens are original or were replaced (RTD). DeBERTa was further enhanced in DeBERTa-V3 with a combination of two techniques as detailed by He, Gao, and Chen (2023). The authors implemented changes to the pre-training process, switching from MLM to RTD (basically adapting it from ELECTRA models) while maintaining the relative position encoding of DeBERTa models. He, Gao, and Chen (2023) introduced a new method called Gradient-Disentangled Embedding Sharing (GDES) that enables the generator (MLM) and the discriminator (RTD) to share embeddings more efficiently. In doing so, they addressed the "tug-of-war" dynamic between generator and discriminator. "Tug-of-war" conveys the idea that, having the generator and the discriminator two

[24] `https://sbert.net/`.

[25] Research on intertextuality in ancient texts is a field that has become more and more involved in historical-religious disciplines. Especially Lee (2007) used a quantitative model to identify similarities between the Gospel of Luke and the Gospel of Mark, demonstrating accuracy in replicating scholarly hypotheses on the reuse of texts. Moritz et al. (2016) analyzed non-literal translations of biblical verses in Ancient Greek and Latin, showing that pre-processing techniques such as stemming and lemmatization are not sufficient to capture the complexity of textual reuse. Finally, although not related to the field of historical-religious research, it is worth mentioning the study of Büchler et al. (2012), who studied text reuse in the *Deipnosophistai*, identifying nearly all references to the Homeric poems annotated by publishers through the analysis of uni- and bi-gram frequencies.

different purposes, they end up shaping the embeddings according their own scope, endangering the quality. GDES allows both components to perform their respective task (to generate and detect false tokens) exploiting the same embeddings without pulling them in the opposite way.

The second element, HLM, introduced for Ancient Greek by Riemenschneider and Krahn (2024) for the SIGTYP 2024 consists in the Hierarchical Language Model (HLM) tokenization technique suggested by Sun et al. (2023) that aims to fix the vocabulary issue in the tokenization phase, i.e. the problem that arises when subword tokenizers—especially in multilingual models—split words inconsistently across languages and scripts, producing a noisy vocabulary. The hierarchical tokenization technique proposed by Sun et al. (2023) addresses this challenge by introducing a two-level approach: it first segments the text into smaller intra-word units, applying self-attention to represent them into initial embeddings, and then refines these representations with further integration of inter-word contextual information. This approach not only enhances the handling of rare words, but also ensures better alignment of tokens across languages, which is crucial for multilingual models aiming to perform well across diverse linguistic contexts.

3.3 Challenges in Semantic Evaluation of Transformer Models for Ancient Greek

English, a widely attested language, has several benchmark datasets available for evaluating PLMs, such as GLUE (Wang et al. 2018) for sentence understanding tasks, SQUAD (Rajpurkar et al. 2016), RACE (Lai et al. 2017), and XNLI (Conneau and Kiela 2018), which is tailored for cross-lingual evaluation. On the other hand, despite the growing number of transformer-based PLMs for Ancient Greek — whether BERT, SBERT, RoBERTa, or multilingual models distilled from larger tools — much work remains to be done in creating one or more reference standards shared by scholars, which would enable a homogeneous and transparent evaluation. It would be desirable that a benchmark campaign to evaluate tools tailored for Ancient Greek started, following the positive example opened by EvaLatin, an evaluation event that has seen, in 2024, its third edition[26]. This necessity is accentuated by the fact that a more recent model is not necessarily a better performing model, especially since the fine-tuning adapts it to specific downstream tasks. Therefore, digital humanists might invest considerable effort and time testing and implementing models to identify those that best suit their research project. Nonetheless, as mentioned above, some projects, such as Krahn, Tate, and Lamicela (2023), provide their internal standards of self-evaluation. In a recent paper, Giuseppe Celano (Celano 2025) compared six PLMs (including GreBERTa, PhilBERTa, GreTa and PhilTA from Riemenschneider and Frank (2023a)) to evaluate which performs better regarding automatic prediction of morphosyntax and lemmatization.

As authors of this contribution, we highlight the lack of a fair and standardized benchmark for statement retrieval across different models, i.e. the task of retrieving the most correlated texts from a corpus given a query sentence, across different transformer-based models. To date, the only attempt to establish a benchmark for PLMs in Ancient Greek is the work of Krahn, Tate, and Lamicela (2023). Their study provides an

[26]In each occurrence, EvaLatin poses shared tasks to scholars and teams to obtain and improve state-of-the-art quantitative results. In 2024, besides Dependency Parsing, also Emotion Polarity Detection has been addressed, in line with a trend of interest in semantics. For further reference, see Sprugnoli, Iurescia, and Passarotti (2024), where authors outline their intention for upcoming semantic challenges, including Semantic Role Labeling and Word Sense Disambiguation.

evaluation of the recall and mean average precision (mAP) for each of the models developed. In their repository[27], 100 queries have been paired with corresponding relevant passages. However, we express reservations about whether this repository can be regarded as a gold standard for information retrieval in Ancient Greek. Specifically, the methodology associates each query with only a single passage, which we consider highly unlikely to exhaustively represent all semantically related passages within the corpus.

Additionally, our efforts to advance the state-of-the-art in benchmarking statement retrieval with transformer-based PLMs for Ancient Greek and Latin are published within the proceedings of a domain-specific conference (Toyin et al. 2026).

4. Statement Retrieval Process and Its Impact on the Study of Classical Texts

We can now outline the technique that allows for statement retrieval with transformer based PLMs. Scholars can already refer to rule-based tools to query an Ancient Greek corpus to retrieve statements from sources they are focusing on in their work. TLG[28] is the most famous one, while *Tracer*[29], despite being more flexible, requires more technical preparation to be used. Besides direct word retrieval, either exact matches or lemmatized forms, both TLG and *Tracer* offer scholars to use an *n*-grams retrieval tool which allows them to construct a single query by combining multiple if clauses, specifying in which shape words must appear in sequence within a defined span of *n*-grams.

In this review, we focus on semantic retrieval for Ancient Greek sentences performed using PLMs. Kevin Krahn has implemented a transformer-based PLM in a retrieval interface for Ancient Greek[30], which, although still in its nascent stage, provides a useful insight of the essential expected outcome of our research project. Therefore, we need to briefly explain how this approach works and how it differs from the rule-based ones.

As a retrieval method, this approach enables users to extract information from a database which, in this case, is an Ancient Greek corpus of written sources[31]. Scholars can query this textual material by specifying one or more words or sentences relevant

[27] https://github.com/kevinkrahn/ancient-greek-datasets/blob/master/sr_search.txt.
[28] https://stephanus.tlg.uci.edu/.
[29] https://www.etrap.eu/research/tracer/.
[30] https://semantic-search.kevinkrahn.com/?lng=grc.
[31] The most complete corpus for Ancient Greek literary text is the TLG https://stephanus.tlg.uci.edu/index.php\#login=true. It is a digital collection of Greek literature from antiquity to the present era. The research program was founded in 1972, is based at the University of California, Irvine, and is nowadays directed by Professor Maria Pantelia. TLG allows users to navigate texts by several search features (author, date, location, genre, etc.), to inquire the corpus through basic tools (e.g. *n*-grams), to visualize textual statistics, and to dynamically call vocabulary tools on the target words. An abridged version is publicly available but to unlock all the features of the TLG a subscription is required; it is not open data. For this reason, digital humanists dedicated to the NLP for Ancient Greek tend to prefer open access databases, even if they contain far less texts than TLG. Among them one can count The Perseus Digital Library (PDL) (Crane 1996), the PROIEL treebank (Haug and Jøhndal 2008), the Ancient Greek Dependency Treebank (AGDT) (Bamman, Mambrini, and Crane 2009), the Diorisis Corpus (Vatri and McGillivray 2018), the Gorman treebanks (Gorman 2020), the GLAUx corpus (Keersmaekers 2021), First1KGreek (https://www.opengreekandlatin.org/) and the Opera Graeca Adnotata (OGA) (Celano 2024). We also noticed that Riemenschneider and Frank (2023a) confirmed to have produced a much larger corpus of Greek text using additional sources, including the Ancient Greek texts conserved in the Internet Archive (https://archive.org/), but the material has still to be published. Finally, we signal that Perseus and First1KGreek material flows into the Open Greek and Latin Project (https://opengreekandlatin.org/). Depending on how they have been processed, these corpora are available either as morpho-syntactically annotated and lemmatized texts or as raw texts, with varying levels of preparation.

to their research. Both the corpus and the queries are encoded into multidimensional vector representations by the transformer-based model to capture morphosyntactic and semantic features. A program calculates the similarity between the query vector and each sentence vector in the corpus. Among the existing similarity measures, the most common is cosine similarity.

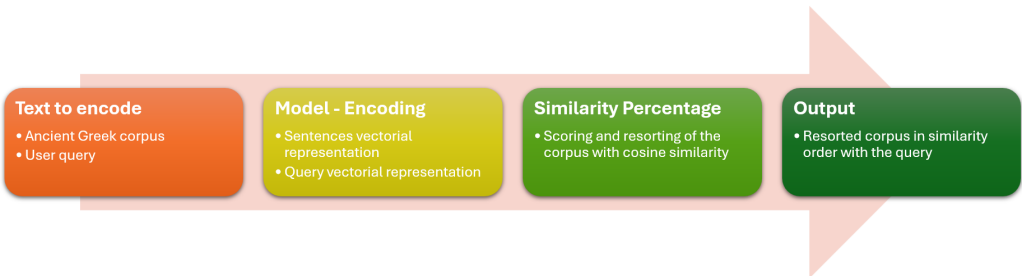

Figure 1
Example of a workflow for retrieving statements semantically similar to the user's query using a transformer-based model.

This process generates a similarity score for each sentence that can also be expressed as a percentage. Based on these scores, the sentences in the corpus are sorted in descending order of similarity, allowing researchers to prioritize the most relevant results.

We may number some key differences between statement retrieval performed through transformer-based and rule-based *verbatim* retrieval tools: Firstly, corpus re-sorting. One can easily tell that the higher the n becomes in a n-grams query, the less results the software will retrieve since the number of if rules makes the query always more restrictive and the possible matches will decrease. This does not happen with statement retrieval performed with transformer-based PLMs since these tools do not select *verbatim* matches within the corpus relying on if-rules. Instead, they perform corpus resorting depending on the (cosine) similarity of the sentences to the query. The corpus is thus reordered in a list of sentences based on similarity decreasing value compared to the sentence in query. Since the user does not select statements across the corpus according to specific rules the number of results will always remain equal to the number of statements in the corpus.

Secondly, non-restrictive search with neglection of sentence length and word order. Since sentence retrieval performed through transformer-based PLMs allows for a less restrictive and non-*verbatim* display of results, the length of the query sentence and the order of the words do not really influence the retrieval process, making this kind of query much more flexible and opening up the possibility to surf the corpus in a semantically related way. This is usually achieved through specifically tailored tokens ([CLS] tokens) that represent the entire sentence, regardless of its length.

Thirdly, representational and semantic retrieval. If n-grams impose an exact word or lemma match, retrieving words and sentences with PLMs offers a much greater range of possibilities, since these tools operate at the distributional representation level of such statements. With the method we outline here, the comparison between statements is performed at the level of the numerical representation, rather than the lexical level. This means that semantic closeness of terms in the vectorial space will significantly influence the retrieval process, resorting the corpus according to similarity and, therefore, considering shared meaning elements.

However, some downsides must be taken into account.

Firstly, there is an inherent uncertainty regarding the completeness of the retrieved texts. The scholar will inevitably face the possibility that the retrieval process is not exhaustive. Even if the corpus is fully reordered according to its similarity with the query, it cannot be guaranteed that all the meaningful sentences related to the query will appear among the top-ranked results. While rule-based search returns all the explicit occurrences of the query in the corpus according to the retrieval parameters, semantic retrieval performed through transformer-based PLMs and cosine similarity produces a continuum of decreasing similarity scores across the entire corpus. Consequently, the user must personally decide when to stop examining the ranked list of statements, being aware that relevant items may remain hidden among the lower-ranked results.

Secondly, unrelated results. The retrieval approach we detail here involves the possibility of having unexpected and unrelated sentences in the resorted results. In fact, *verbatim* retrieval exhausts all the possibilities offered by the corpus, while sentence retrieval performed confronting numerical representations might suggest unrelated elements. Therefore, the scholar will have to carefully evaluate the resorted sentences to make sure they are really related to his query.

Thirdly, black box. This likely remains the core issue, that of the "black box": even if, once the training process is complete, the model is crystallized and operates in a (deterministic) fixed way always returning the same representation per word or sentence, we still cannot fully explain the rationale behind the distribution of the meaning of the model. The researcher does not have control over the elaboration performed by the distributional model, this happens entirely at a sub-word level and cannot be oriented through rules since it is completely data-driven. The developer's space of action remains, in any case, absolutely important and details can radically change the results provided by the model: the workflow that surrounds the embedding process, especially the training phase, must be carefully designed, as minimal nuances, scaled to tens of millions of words, deeply impact the results. The pipeline includes core steps like composition and pre-processing of the training and validation dataset, training or distillation of the model, fine-tuning for specific tasks, exploitation and evaluation. This is the reason why researchers, in the training phase are expected to carefully establish hyperparameters such as the loss function (difference between predicted and actual outputs), the batch size (number of samples processed per iteration), the max sequence length of tokens in a sentence, learning rate (how quickly the optimizer updates the model's weights), number of warm-up steps (in which the learning rate is gradually increased), mask ratio (for MLM training) and training epochs. This parameters are to be carefully designed and significantly impact the performances, but, in the end, scholars will have to face a margin they cannot deal with, the black box.

In sum, statement (cosine) similarity retrieval pursued confronting numerical representations, the method we outline here, is a technique that may offer advancements in the scholarship with the caveat of considering the tool as a co-pilot instead of an oracle. Therefore, a good training should be offered with the distribution of this kind of tools.

5. Conclusion

We have shown how the application of distributional semantics and transformer-based models has significantly advanced the study of Ancient Greek, transforming both semantic analysis and historical linguistic research. The development of PLMs highlights the growing potential of these methods to uncover semantic changes of words, semantic relationships and intertextual connections within ancient texts. Despite these advances, challenges remain, including the need for comprehensive evaluation benchmarks and

the difficulty of training models for an ancient language. About statement retrieval, although PLMs offer more flexible and semantically nuanced retrieval methods compared to traditional rule-based tools, issues such as uncertainty about retrieval completeness, irrelevant results, and the "black box" nature of these models remain significant obstacles. As digital humanists continue to refine these models and explore new applications, it is clear that only careful evaluation and a deep understanding of the limitations of the models can fully unlock their potential. In other words, some caution is necessary to temper some *positivistic* tendencies that have surrounded the use of AI in language studies over the past 70 years.

In 1955, McCarthy et al. (1955) submitted a whitepaper containing their proposal for a two-months research program to be held at Dartmouth College the following summer. In this document, they introduced the expression "artificial intelligence" for the first time and suggested laying the groundwork to replicate human intelligence and language, based on the conjecture that

> every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problem now reserved for humans, and improve themselves.

A lot has changed since this first step: instead of having human instructors expliciting all the rules that describe a certain reality to train a domain-expert system, nowadays, with language models, information is implicitly and automatically distributed among weights: no single weight corresponds to a specific piece of information. Instead, it is the collective activation of weights that represents complex features and relationships from the input data. This method is effective, but at the same time, it makes it difficult to put into words what kind of linguistic features the model has learned, entangling the challenge of explainability and tenability of scientific gains. Automatically learning from data with sub-symbolic representation of features means, as suggested by McGillivray (2022), exchanging interpretability for performance or, in other words, giving priority to the results promised by PLMs relieving control over the work process. We should also be cautious about anticipating results: more than half a century later the white paper of McCarthy et al., Turney and Pantel (2010) say that

> it seems possible to us that all of the semantics of human language might one day be captured in some kind of VSM

The two quotations share the hope of the authors for a drastic development of AI's performance toward and beyond the human level, a hope that all the features of human-level subjects will one day be fully grasped by computers so as to obtain a complete representation of the data. As the authors of this contribution, we suspect that the sub-linguistic distribution of features without its logical understanding will not grant models to properly appreciate meaning. In any case, even if models will always have to asymptotically approximate concepts from the outside, the craft of improving and tailoring them for specific tasks may be a beneficial aid to the community of scholars who daily engage with the challenge of semantics.

Acknowledgments

§1: Elia Scapini and Federico Iezzi; §2: Federico Iezzi; §3,1: Federico Iezzi; §3,2: Elia Scapini e Federico Iezzi; §3,3-4: Elia Scapini; §5: Elia Scapini and Federico Iezzi.

This review was performed in the framework of the Italian Strengthening of the ESFRI RI RESILIENCE (ITSERR) project (cod. Progetto IR0000014 - CUP B53C22001770006) – Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4, "Istruzione Ricerca", Componente 2, "Dalla ricerca all'impresa", Investimento 3.1, "Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione", finanziato dall'Unione Europea – NextGenerationEU – rif. Avviso MUR 3264/2021. ITSERR is an interdisciplinary and distributed Research Infrastructure for Religious Studies that aims to strengthen the RESILIENCE RI project in Italy. Elia Scapini and Federico Iezzi are involved in ITSERR as memebers of the fourth Work Package (WP4) that devotes its efforts to study the semantics of the Nicene-Constantinopolitan Symbolo and to build informatic tools for scholars in Religious Studies. Federico Iezzi and Elia Scapini are also members of the Fondazione per le Scienze Religiose (FSCIRE) in Bologna.

References

Artetxe, Mikel and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Assael, Yannis, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375, Hong Kong, China, November. Association for Computational Linguistics.

Assael, Yannis, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Maria Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

Bamman, David, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8). 4-5 December 2009, Milan, Italy*, pages 5–15.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Biagetti, Erica, Martina Giuliani, Silvia Zampetta, Silvia Luraghi, and Chiara Zanchi. 2024. Combining neo-structuralist and cognitive approaches to semantics to build wordnets for ancient languages: Challenges and perspectives. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING*, pages 151–161, Torino, Italia, May. ELRA and ICCL.

Biagetti, Erica, Chiara Zanchi, and William Michael Short. 2021. Toward the creation of WordNets for ancient Indo-European languages. In Piek Vossen and Christiane Fellbaum, editors, *Proceedings of the 11th Global Wordnet Conference*, pages 258–266, Online and Pretoria, South Africa, January. Global Wordnet Association.

Bizzoni, Yuri, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boschetti, Federico. 2010. *A Corpus-based Approach to Philological Issues*. Ph.D. thesis, University of Trento. https://hdl.handle.net/20.500.14242/92918.

Boschetti, Federico. 2019. Semantic analysis and thematic annotation. In Monica Berti, editor, *Digital Classical Philology*. De Gruyter, Berlin, pages 321–339. https://doi.org/10.1515/9783110599572-018.

Brunila, Mikael and Jack LaViolette. 2022. What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 4403–4417, Seattle, United States, July. Association for Computational Linguistics.
Burns, Patrick. 2019. Building a text analysis pipeline for classical languages. In Monica Berti, editor, *Digital Classical Philology*. De Gruyter, Berlin, pages 159–176. https://doi.org/10.1515/9783110599572-010.
Büchler, Marco, Gregory Crane, Maria Moritz, and Alison Babeu. 2012. Increasing recall for text re-use in historical documents to support research in the humanities. In *Theory and Practice of Digital Libraries*, pages 95–100, Berlin, Heidelberg. Springer.
Celano, Giuseppe G. A. 2024. Opera Graeca Adnotata: Building a 34m+ token multilayer corpus for Ancient Greek. http://arxiv.org/abs/2404.00739.
Celano, Giuseppe G. A. 2025. A state-of-the-art morphosyntactic parser and lemmatizer for Ancient Greek. In Isuri Nanomi Arachchige, Francesca Frontini, Ruslan Mitkov, and Paul Rayson, editors, *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities*, pages 48–65, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.
Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. https://arxiv.org/abs/2003.10555.
Conneau, Alexis and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
Conneau, Alexis, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online, July. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.536.
Cowen-Breen, Charlie, Creston Brooks, Barbara Graziosi, and Johannes Haubold. 2023. Logion: Machine-learning based detection and correction of textual errors in greek philology. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 170–178, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.
Crane, Gregory. 1996. Building a digital library: the perseus project as a case study in the humanities. In *Proceedings of the First ACM International Conference on Digital Libraries*, New York, NY, USA, March. Association for Computing Machinery.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, USA, June. Association for Computational Linguistics.
Fellbaum, Christiane. 1998. A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32(2):209–220.
Firth, John Rupert. 1957. *Studies in Linguistic Analysis*. Blackwell, Oxford.
Frermann, Lea and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.
Gavin, Michael. 2018. Vector Semantics, William Empson, and the Study of Ambiguity. *Critical Inquiry*, 44(4):641–673.
Gorman, Vanessa. 2020. Dependency treebanks of Ancient Greek prose. *Journal of Open Humanities Data*, 6.
Graziosi, Barbara, Johannes Haubold, Charlie Cowen-Breen, and Creston Brooks. 2023. Machine learning and the future of philology: A case study. *TAPA - Transactions of the American Philological Association*, 153(1):253–284. Publisher: Johns Hopkins University Press.
Grewcock, Rachel. 2018. *Computational Semantics and the Syntax of Motion in Ancient Greek*. MPhil thesis, University of Cambridge, Cambridge.
Grieve, J. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Harris, Zellig S. 1954. Distributional structure. *WORD*, 10(2):146–162.
Haug, D. and Marius L. Jøhndal. 2008. Creating a parallel treebank of the old indo-european BibleTranslations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*, Marrakech, Morocco, June.
He, Pengcheng, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv. http://arxiv.org/abs/2111.09543.
He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced BERT with disentangled attention. https://arxiv.org/abs/2006.03654.
Iwata, Naoya, Ikko Tanaka, and Jun Ogawa. 2024. Improving semantic search accuracy of classical texts through context-oriented translation. In *Proceedings of IPSJ SIG Computers and the Humanities Symposium*, Sendai, Japan, 11-18. Information Processing Society of Japan (IPSJ).
Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.
Johnson, Kyle P., 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 Heng Ji, Jong C. Park, and Rui Xia, editors, *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, August. Association for Computational Linguistics.
Keersmaekers, Alek. 2020. Automatic semantic role labeling in Ancient Greek using distributional semantic modeling. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 59–67, Marseille, France, May. European Language Resources Association (ELRA).
Keersmaekers, Alek. 2021. The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In Nina Tahmasebi, Adam Jatowt, Yang Xu, Simon Hengchen, Syrielle Montariol, and Haim Dubossarsky, editors, *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 39–50, Online, August. Association for Computational Linguistics.
Keersmaekers, Alek and Dirk Speelman. 2023. Applying distributional semantic models to a historical corpus of a highly inflected language: the case of Ancient Greek. *Glottometrics*, 55:17–43.
Koutsikakis, John, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting sesame street. http://arxiv.org/abs/2008.12014.
Krahn, Kevin, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence embedding models for Ancient Greek using multilingual knowledge distillation. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.
Köntges, Thomas. 2020a. Measuring Philosophy in the First Thousand Years of Greek Literature. *Digital Classics Online*, pages 1–23.
Köntges, Thomas. 2020b. The un-platonic *Menexenus*: A stylometric analysis with more data. *Greek, Roman, and Byzantine Studies*, 60(2):211–241.
Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.
Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. http://arxiv.org/abs/1901.07291.
Lee, John. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic, June. Association for Computational Linguistics.
Lenci, Alessandro, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, 56:1–45, 03.

List, Nicholas. 2022. How can we investigate Ancient Greek categories without the influence of our own? Exploring kinship terminology using Word2Vec. *International Journal of Lexicography*, 35(2):137–152.

Liu, Lei and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Liu, Yiheng, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, Yi Pan, Shaochen Xu, Zihao Wu, Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, Tianming Liu, and Bao Ge. 2025. Understanding llms: A comprehensive overview from training to inference. *Neurocomputing*, 620:129190.

Liu, Yinhan, 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. http://arxiv.org/abs/1907.11692.

Mambrini, Francesco, Marco Carlo Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking valency frames and wordnet synsets in the lila knowledge base of linguistic resources for latin. In *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems (SEMANTiCS 2021)*, Amsterdam, The Netherdalnds, September.

Manousakis, Nikos and Efstathios Stamatatos. 2018. Devising rhesus: A strange 'collaboration' between Aeschylus and Euripides. *Digital Scholarship in the Humanities*, 33(2):347–361.

Marongiu, Paola, Barbara McGillivray, and Anas Fahad Khan. 2024. Multilingual workflows for semantic change research. *Journal of Open Humanities Data*, 10(1).

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 27(4):12.

McGillivray, Barbara. 2022. How to use word embeddings for natural language processing. SAGE Research Methods: Doing Research Online. Publisher: SAGE Publications Ltd.

McGillivray, Barbara, Simon Hengchen, Viivi Lähteenoja, Marco Palma, and Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907.

McGillivray, Barbara, Fahad Khan, and Paola Marongiu. 2023. A new CLARIN resource family for lexical semantic change - final report. Publisher: Zenodo.

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

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Minozzi, Stefano. 2009. The Latin WordNet Project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*. Innsbruck, AUT, pages 707–716.

Moritz, Maria, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to Bible reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas, November. Association for Computational Linguistics.

Packard, David W. 1973. Computer-assisted morphological analysis of Ancient Greek. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*, Pisa, Italy, August.

Palladino, Chiara, Maryam Foradi, and Tariq Yousef. 2021. Translation alignment for historical language learning: a case study. *Digital Humanities Quarterly*, 15(3).

Pavlopoulos, John and Maria Konstantinidou. 2023. Computational authorship analysis of the homeric poems. *International Journal of Digital Humanities*, 5(1):45–64.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Perrone, Valerio, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. Lexical semantic change for Ancient Greek and latin. http://arxiv.org/abs/2101.09069.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy, August. Association for Computational Linguistics.

Raffel, Colin, 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.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November. Association for Computational Linguistics.

Riemenschneider, Frederick and Anette Frank. 2023a. Exploring large language models for classical philology. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada, July. Association for Computational Linguistics.

Riemenschneider, Frederick and Anette Frank. 2023b. Graecia capta ferum victorem cepit. detecting Latin allusions to Ancient Greek literature. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 30–38, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Riemenschneider, Frederick and Kevin Krahn. 2024. Heidelberg-boston @ SIGTYP 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers. In Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, and Ekaterina Vylomova, editors, *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 131–141, St. Julian's, Malta, March. Association for Computational Linguistics.

Rodda, Martina Astrid, Alessandro Lenci, and Marco S. G. Senaldi. 2017. Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1).

Rodda, Martina Astrid, Philomen Probert, and Barbara McGillivray. 2019. Vector space models of Ancient Greek word meaning, and a case study on homer. *Traitement Automatique des Langues*, 60(3):63–87. Place: France Publisher: ATALA (Association pour le Traitement Automatique des Langues).

Russell, Stuart J. and Peter Norving. 2021. *Intelligenza artificiale: un approccio moderno*, volume 2. Pearson, Milano, 4 edition. OCLC: 1288701545.

Salton, Gerard. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall.

Shen, Tianxiao, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198,

Online, November. Association for Computational Linguistics.
Singh, Pranaydeep, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval greek. In Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *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), November. Association for Computational Linguistics.
Sommerschield, Thea, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747, September.
Spanopoulos, Andreas. 2022. *Language Models for Ancient Greek*. BSc thesis, National and Kapodistrian University of Athens, Athens.
Sprugnoli, Rachele, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia, May. ELRA and ICCL.
Stopponi, Silvia, Nilo Pedrazzini, Saskia Peels, Barbara McGillivray, and Malvina Nissim. 2023. Evaluation of distributional semantic models of Ancient Greek: Preliminary results and a road map for future work. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 49–58, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.
Stopponi, Silvia, Nilo Pedrazzini, Saskia Peels-Matthey, Barbara Mcgillivray, and Malvina Nissim. 2024. Natural language processing for ancient greek: Design, advantages and challenges of language models. *Diachronica*, 41, 07.
Sun, Li, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From characters to words: Hierarchical pre-trained language model for open-vocabulary language understanding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620, Toronto, Canada, July. Association for Computational Linguistics.
Thaller, Manfred. 1987. Methods and techniques of historical computation. *History and Computing*, 1:147–156.
Toyin, Hawau Olamide, Federico Iezzi, Elia Scapini, Giulio Federico, and Giovanni Puccetti. 2026. Gretino: A Greek and Latin dataset to benchmark retrieval systems in classical languages. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 919–928, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).
Turney, P. D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
Vatri, Alessandro and Barbara McGillivray. 2018. The Diorisis Ancient Greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65.
Vatri, Alessandro, Barbara McGillivray, and Viivi Lähteenoja. 2019. Ancient Greek semantic change - annotated datasets and code. Publisher: figshare. https://doi.org/10.6084/m9.figshare.c.4445420.
Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.
Wishart, Ryder and Prokopis Prokopidis. 2017. Topic modelling experiments on Hellenistic corpora. In *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, Bloomington, Indiana, USA, January.
Yamshchikov, Ivan P., Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in plutarch's shadows. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

6071–6080, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.
Yousef, Tariq. 2023. *Translation Alignment Applied to Historical Languages: Methods, Evaluation, Applications, and Visualization*. Ph.d thesis, Leipzig University, Leipzig, June.
Yousef, Tariq, Chiara Palladino, and Farnoosh Shamsian. 2023. Classical philology in the time of AI: Exploring the potential of parallel corpora in ancient language. In Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti, editors, *Proceedings of the Ancient Language Processing Workshop*, pages 179–192, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.
Yousef, Tariq, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. An automatic model and gold standard for translation alignment of Ancient Greek. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France, June. European Language Resources Association.
Yousef, Tariq, Chiara Palladino, David J. Wright, and Monica Berti. 2022b. Automatic translation alignment for Ancient Greek and Latin. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France, September. European Language Resources Association.
Zafar, Schyan and Geoff K. Nicholls. 2022. Measuring diachronic sense change: New models and monte carlo methods for bayesian inference. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1569–1604, 09.
Zafar, Schyan and Geoff K. Nicholls. 2024. An embedded diachronic sense change model with a case study from Ancient Greek. *Computational Statistics & Data Analysis*, 199.

Appendix A: Transformer-based PLMs for Ancient Greek

Table 2

Model	*Author, Year*	*Type*	*Language*	*Purpose*
Ancient Greek BERT	Singh, Rutten, and Lefever (2021)	BERT	AG	Morpho/PoS tagging
AG-BERT	Yamshchikov et al. (2022)	BERT	AG	Authorship attribution
Ancient Greek RoBERTa	Spanopoulos (2022)	RoBERTa	AG	PoS tagging
GRC-ALIGNMENT	Yousef et al. (2022a, 2022b)	XLM-RoBERTa	Multi	Text alignment
Ithaca	Assael et al. (2022)	Transformer-based	AG	Inscriptions restoration
ELECTRA-grc	Mercelis and Keersmaekers (2022)	ELECTRA	AG	
Logion BERT	Cowen-Breen et al. (2023)	BERT	AG	Scribal errors correction
GrεBERTa	Riemenschneider and Frank (2023a)	RoBERTa	AG	Morpho/PoS tagging, Dependency Parsing
PhilBERTa	Riemenschneider and Frank (2023a)	RoBERTa	AG, Lat, En	Morpho/PoS tagging, Dependency Parsing
GrεTa	Riemenschneider and Frank (2023a)	T5	AG	Morpho/PoS tagging, Dependency Parsing, Lemmatization
PhilTa	Riemenschneider and Frank (2023a)	T5	AG, Lat, En	Morpho/PoS tagging, Dependency Parsing, Lemmatization
SPhilBERTa	Riemenschneider and Frank (2023b)	Distilled RoBERTa	AG, Lat, En	Sentence understanding
mBERT, XLM-R	Krahn, Tate, and Lamicela (2023)[32]	Distilled sBERT	AG, En	Semantic similarity
HLM-DeBERTa-V3	Riemenschneider and Krahn (2024)	HLM-DeBERTa-V3	AG	Tokenization, morpho/PoS tagging

[32] A modified version of this model, coupled with a sentence tailored version of HLM, runs at `https://semantic-search.kevinkrahn.com/?lng=grc`.