Applying Distributional Semantic Models to a Historical Corpus of a Highly Inflected Language: the Case of Ancient Greek

Alek Keersmaekers[1*], Dirk Speelman[1]

[1] University of Leuven
[*] Corresponding author's email: alek.keersmaekers@kuleuven.be

DOI: https://doi.org/10.53482/2023_55_410

ABSTRACT

So-called "distributional" language models have become dominant in research on the computational modelling of lexical semantics. This paper investigates how well such models perform on Ancient Greek, a highly inflected historical language. It compares several ways of computing such distributional models on the basis of various context features (including both bag-of-words features and syntactic dependencies). The performance is assessed by evaluating how well these models are able to retrieve semantically similar words to a given target word, both on a benchmark we designed ourselves as well as on several independent benchmarks. It finds that dependency features are particularly useful to calculate distributional vectors for Ancient Greek (although the level of granularity that these dependency features should have is still open to discussion) and discusses possible ways for further improvement, including addressing problems related to polysemy and genre differences.

Keywords: distributional semantics, Ancient Greek, word similarity

1 Introduction

So-called "distributional" language models (also "vector space models", "semantic spaces" or "word embeddings") have become dominant in research on the computational modelling of lexical semantics. These techniques start from the long-held assumption that you can "know a word by the company it keeps" (Firth 1957) and try to model the semantic relatedness among different words based on their occurrence in shared contexts. While there is plenty of literature on the application of such models to modern languages, historical languages such as Ancient Greek have received less attention so far (although this is increasing, see Section 2.2). Yet there are several challenges that make Ancient Greek an interesting case study.

Many of these challenges have to do with the size and nature of the available corpus materials. First of all, we have far less data for Ancient Greek than for a modern language such as English: in the order of millions rather than billions for the whole corpus, and only on average 2 million words per century.

Since distributional language models require large amounts of data, making a selection in the already rather limited corpus material we have would inevitably lead to data sparsity. Yet the Ancient Greek corpus also spans a large period of time, and its genres are rather unevenly distributed (see Section 3), giving us a far less homogenous dataset to start from in comparison to e.g. modern language distributional models trained on Wikipedia or newspaper prose. Additionally, most of the data are of a literary or technical nature, including several genres such as epic poetry or scientific prose with a rather idiosyncratic language, while the non-literary, everyday language parts of the corpus, e.g. papyrus letters, are rather limited. But it is not just the precarious text transmission that stands in the way of a smooth application of distributional language models: the nature of the Greek language itself also presents some additional problems. We mentioned above that distributional language models measure word similarity on the basis of shared contexts: this notion of "context" typically refers to the lexical and syntactic context of a word, i.e. the words it combines with, either based on the words that precede or follow the target word (so called "bag-of-words"-models), or on more sophisticated measures such as syntactic dependency relationships. This works well for isolating languages, but it is not immediately obvious that such approaches would work equally well with a language such as Ancient Greek, which expresses much information by relying on morphological rather than syntactic means. A Greek finite verb, for instance, is inflected for person, number, tense and aspect, mood and voice. Of these features, English only expresses number and tense morphologically. Furthermore, the word and constituent order of Ancient Greek is notoriously free (see Dik 1995), which might complicate distributional bag-of-words models that only take the direct environment of a word into account.

This paper aims to test the validity of distributional semantic models on the Ancient Greek language, by evaluating how well these models are suited to retrieve semantically similar words to a given target word. While language-external issues such as genre imbalance will be addressed to some extent, the focus is first and foremost on language-internal issues, i.e. which contextual information works best to model word similarity for Ancient Greek (and other typologically related languages). It is structured as follows: Section 2 will give a broad technical background of distributional semantic models in general, and discuss previous approaches to distributional semantic modeling of Ancient Greek. Section 3 will give an overview of the corpus we used, and Section 4 will describe the specific parameters of the distributional models we compared in more detail. Section 5 will analyze the results of the word similarity task, and Section 6 will summarize and analyze the main results of this study.

2 Models of distributional semantics

2.1 Calculating distributional vectors

While it goes beyond the scope of this paper to give a full overview of the broad field of distributional semantic modelling (see Erk 2012, Lenci 2018 for some recent surveys), this section will give a concise

presentation of the terminology and techniques used in this paper. First of all, as for distributional techniques in general, a distinction can be made between so called context-counting and context-predicting models (the latter also known as "neural language models") (Baroni et al. 2014). Both types of models represent a word as a vector of real numbers, so that the vectors of words that are semantically similar are also mathematically similar. However, they differ with respect how these vectors are calculated: the vectors of context-count models directly contain the co-occurrence frequencies (either weighted or not, see below) of the context words with which the target word occurs. The weights of context-predict models, in contrast, are calculated in such a way (on the basis of a supervised machine learning approach, using neural networks) to predict the contexts in which the target word tends to appear. Such an approach has been found to outperform context-count models on a wide range of tasks (Baroni et al. 2014). However, one of the main advantages of using context-count models is their greater transparency: the individual elements of these vectors directly refer to the contexts in which the target word appears, while the elements of vectors calculated with a context-predict approach do not have any obvious meaning. This paper aims to compare and understand the underlying reasons why certain models are better suited to perform a number of specific tasks than others. Since the focus is not on achieving state-of-the-art performance for these tasks, we will stick to a context-count approach, although a comparison with context-predict models is certainly a desideratum for the future.

An appropriate starting point for explaining the procedure behind the creation of context-count vectors is Turney and Pantel (2010). The first step consists in counting for each target word how often certain other words occur in its context, for example a window of N preceding and following words (see Section 4 for alternative ways of determining the context).[1] Next, the elements on the matrix are typically weighted to give more weight to more "surprising"[2] co-occurrences. This paper will use the Pointwise Positive Mutual Information (PPMI) measure to do so, which has been shown to outperform other weighting approaches (Bullinaria and Levy 2007).[3] Function words and/or stop words are often removed from the matrix. However, as their removal has been shown to have no significant positive or negative effect on performance for English data (Bullinaria and Levy 2012), we refrained from removing them in the context of this paper (although we left out tokens indicating punctuation or "gaps" in the text): our early experiments suggested that removing them does not have an effect for Ancient Greek either.

[1] The target and context words can be either lemmas or word forms. Since Greek is a highly inflectional language (a Greek participle, for instance, has more than 150 possible forms), using word forms would lead to data sparsity, so all the models described in this paper are based on word forms.
[2] The term "surprising" is used here in a statistical context, to refer to co-occurrences that appear more than we would expect from random chance.
[3] The PPMI is calculated by first log-transforming the observed frequency of a co-occurrence pattern divided by its expected frequency (i.e. the PMI measure), which has a negative value when the observed frequency is lower than the expected frequency and a positive value when it is higher than the expected frequency. Subsequently, all negative PMIs are set to 0 (i.e. all patterns with an observed frequency that is lower than the expected frequency). See Turney and Pantel (2010: 157-158) for more information.

Subsequently, a dimension reduction technique such as Singular Value Decomposition (SVD) is often applied to the co-occurrence matrix in order to reduce the context information to a smaller number of latent dimensions, which often improves the performance of context-count models (Bullinaria and Levy 2012). However, we will refrain from doing so in the context of this paper, in order to gain a better insight in the specific context features that cause semantic similarity (see Section 5).

To detect semantic similarity, we next need to calculate by some measure how similar the vectors of the different target words are. We will use the cosine similarity measure for this purpose, which has been found to outperform other measures to detect semantic similarity in the vector space (Bullinaria and Levy 2007, Lapesa and Evert 2014). The cosine similarity (as is obvious from its name) captures the cosine of the angle between the two vectors that are compared, and is 1 when they are completely similar and 0 when they are completely dissimilar (see Turney and Pantel 2010: 160-161 for the calculation).

2.2 Related work

This section will give an overview of the relevant literature: more details about the model parameters of the main studies discussed here can be found in Table 1. Most studies investigating distributional models for Ancient Greek are applied in nature, in particular using them in order to track lexical semantic change. As for context-count models, the first study was Boschetti (2010), who used a context-count model to examine the Greek lexicon in various ways, including the diachronic development of specific words, their polysemy structure in different genres and the taxonomical relations among them. Additionally, Boschetti argues that such models can also be used for text-critical ends, i.e. to evaluate the appropriateness of editorial conjectures. Rodda et al. (2017) use distributional models trained on a part of the *TLG* corpus (36 million tokens in total) to evaluate the hypothesis whether Christianity had a significant effect on the Greek lexicon. Their results confirm the crucial role of Christianity on lexical semantic change in Greek, and also show that distributional models can bring unexpected patterns of change to light. Rodda et al. (2019) have developed distributional models in order to study linguistic variation in Ancient Greek epic formulae. They are one of the only studies that compare several (context-count, SVD-reduced) distributional models against independent benchmarks from various sources (ancient scholarship – the *Onomasticon* by Julius Pollux – modern lexicography – Schmidt's dictionary of synonyms – and an NLP resource – the *Open Ancient Greek WordNet*). These models vary with regard to the context window (1, 5 and 10 words to the left and right) and frequency threshold (including all words, words that occur at least 20, 50 and 100 times in the corpus). They find that context windows of 5 words and frequency thresholds of 20 or 50 words achieve the best results on their benchmarks (with the *Onomasticon* and Schmidt's dictionary matching the semantic spaces of the distributional models better than *Ancient Greek WordNet*).

There have also been some studies on context-predict models for Ancient Greek: an experimental *word2vec* model has been implemented in the Python Classical Language Toolkit (Burns 2019), although their results

have not been evaluated yet. Perrone et al. (2021) compare the results of two context-predict models to a dynamic Bayesian mixture model for the task of detecting semantic change, but conclude that the latter approach delivers superior results over the context-predict models. List (2022) investigates how Word2Vec models can be used for lexicographic purposes. Finally, recently various transformer models have been trained for Ancient Greek, including Singh et al. (2021) (BERT), Yamshchikov et al. (2022) (BERT) and Spanopoulos (2022) (RoBERTa). These studies did not evaluate their models for semantic purposes, however, making them less relevant for this paper. In contrast, Riemenschneider and Frank (2023) evaluate RoBERTa models on various non-semantic and semantic tasks, including their ability to distinguish synonyms from antonyms. Additionally, Mercelis et al. (Forthcoming) evaluate the results of an ELECTRA model for word sense disambiguation, comparing both unsupervised and supervised techniques.

Stopponi et al. (2023) compare the performance of both context-count and context-predict models trained on the Diorisis corpus. The evaluation is done on the AGREE benchmark, containing evaluations of word similarity rated by experts. For the context-count models, the authors compare both dimensionally reduced vectors (with SVD) and non-reduced vectors, while for the context-predict models, they compare a SGNS model to a Continuous Bag-of-Words model (CBOW), in all cases using a window size of 5 words. They conclude that context-count models perform better than context-predict models against this benchmark, with the non-reduced vectors performing the best of all 4 models.

While interest in distributional models for Ancient Greek is clearly increasing, in all of these studies only bag-of-words models are investigated,[4] and dimension reduction or neural networks are generally employed, making the resulting vectors difficult to interpret. The main contribution of this paper is therefore the following: it will compare various ways to incorporate syntactic context as well (see Section 4), and offer a thorough investigation of the resulting semantic spaces and the various context features that cause semantic similarity. Additionally, it will employ the GLAUx corpus (see Section 3), the largest openly available Greek corpus so far, allowing for higher quality semantic spaces than the previous studies.

Table 1: Previous studies on distributional semantic modeling for Ancient Greek.

Study	Architecture	Application	Corpus
Boschetti (2010)	Count, SVD (window 100)	Describing the lexicon	TLG
Rodda et al. (2017)	Count, SVD (window 5)	Lexical semantic change	TLG
Rodda et al. (2019)	Count, SVD (varying window)	Model comparison, epic formulae	Diorisis
Perrone et al. (2021)	Predict, word2vec (SGNS/TR)	Lexical semantic change	Diorisis
List (2022)	Predict, word2vec (SGNS)	Lexicography	Diorisis
Riemenschneider & Frank (2023)	Predict, transformer (RoBERTa)	Model comparison	Custom
Mercelis et al. (forthcoming)	Predict, transformer (ELECTRA)	Word Sense Disambiguation	GLAUx
Stopponi et al. (2023)	Count, SVD/Non-SVD; Predict, word2vec (SGNS/CBOW)	Model comparison	Diorisis

[4] However, a future investigation into the performance of syntactic models has been announced by Stopponi et al. (2023).

3 The corpus

As mentioned in the introduction of this paper, the Ancient Greek corpus is quite small as compared to some modern language corpora. What is more, the largest collection of Greek text – the corpus of the *Thesaurus Linguae Graecae* (TLG) – has not made its data publicly available. However, there have been some recent large-scale open initiatives: the Diorisis corpus (Vatri and McGillivray 2018), containing 10.2M tokens from the 8th century BC to the 5th century AD, and the GLAUx corpus (Keersmaekers 2020), containing 27.7M tokens from the 8th century BC to the 8th century AD. Since the Diorisis corpus is much smaller and does not contain syntactic annotation, which was essential for the experiments described in the next sections, we made use of the latter corpus. More specifically, we used an earlier version of GLAUx, which was larger (37.2M tokens) but also noisier, containing several texts with OCR problems. The accuracy is about 0.95 for part-of-speech/morphological tagging and 0.98 for lemmatization, while syntactic parsing accuracy (Labeled Attachment Score) ranges between 0.75 and 0.88 depending on text genre (see Keersmaekers 2020).

The literary texts are quite diverse with respect to texts genre, ranging from epic poetry to drama, philosophy, historical narrative, scientific prose and so on. Previous studies have already indicated that text genre has an important effect for the computational modelling of semantics for Ancient Greek (Boschetti 2010, McGillivray et al. 2019). Since we did not want to further reduce the corpus to avoid data sparsity, we used the full corpus for the construction of distributional vectors. However, in our analysis we will also consider how genre and diachrony may influence the resulting semantic spaces.

4 Construction of context models

As mentioned in Section 2, all techniques discussed in this paper make use of some notion of "context". In traditional collocational and distributional semantic approaches, this context is simply defined as a window of preceding and/or following words – a so-called "bag-of-words" approach. This context window can be as wide or small as the researcher wants to define it, but in general it has been found that larger context windows leads to a more associative, topical similarity (e.g. "soldier"/"war") while smaller context windows lead to cosine similarities that indicate relationships that are more taxonomic (e.g. "soldier"/"warrior") (Peirsman et al. 2008; Kolb 2009).

Another way to define "context" is to use the *syntactic* context of a word as features, in particular involving syntactic dependencies (Lin 1998, Padó and Lapata 2007). This approach has been shown to return even tighter taxonomic syntactic relationships than small-window bag-of-words approaches (e.g. Heylen et al. 2008, see also Levy and Goldberg 2014 for context-predict models). In such an approach context features typically look like *child/OBJ* (as in *child* is the object of the target word X, e.g. of *raise* in *he raised the child*), although it is in principle possible to include less or more detailed information (see below).

Finally, in the context of a highly inflectional language such as Ancient Greek, it also makes sense to consider the *morphological* context of a word. Greek dictionaries such as Liddell-Scott-Jones (Jones et al. 1996), for instance, typically list what cases, moods etc. a given word frequently combines with. In fact, one could wonder whether language-internal categories such as case are in fact not better suited to model the semantics of Ancient Greek than categories that are considered to be more language-general such as "object" (i.e. by replacing "child is the object of X" by e.g. "child is a dative dependent on X") – see in this context Croft's (2013) skepticism on defining such language-general categories. Particularly with context-predict models, there have been several approaches that integrated morphological or other formal characteristics of the target word itself in its vector embedding, i.e. to assign similar vectors to formally similarly looking words (e.g. Luong et al. 2013; Botha and Blunsom 2014, Bojanowski et al. 2017), but the use of morphological features as context features has, to the best of our knowledge, not been explored yet.

To test the role of the type of context model in detecting Ancient Greek word similarity, we have constructed five types of context models, as summarized in Table 2 below. All models use PPMI weighting and require a context feature to occur at least 150 times, so as to avoid features that are too infrequent as well as noise in the data. The first context model is a simple bag-of-words model (model *BOW*). We used a context of 4 preceding and following words, since this window size turned out to be the most optimal to detect word similarity for Ancient Greek without bringing in too much noise in some early (unpublished) experiments. The four other models make use of syntactic information, using the automatically parsed data described in Section 3. The first (which we will style *DepMinimal*) simply states the frequency of lemmas that have a direct dependency link with the target word, i.e. when the context word occurs as the head or as a child of the target word, without adding information about syntactic relation or whether the context word occurs as the head or child (i.e. the direction of the arc). The second (*DepHeadChild*) enhances this with the information whether the given context word occurs as the target word's head or child, i.e. in ἡ θυγάτηρ τῆς μητρὸς "the mother's daughter", the relevant feature for μήτηρ "mother" would be θυγάτηρ/head ("daughter"), while in ἡ μήτηρ τῆς θυγατρός "the daughter's mother" the feature would be θυγάτηρ/child. In the third model (*DepSyntRel*) a syntactic label is added, e.g. θυγάτηρ/head/ATR for "μήτηρ is an attribute of θυγάτηρ" or θυγάτηρ/child/ATR for "θυγάτηρ is an attribute of μήτηρ". Finally, in a fourth model (*DepMorph*) we use morphological information instead of syntactic labels. Instead of using the full morphology of the context words (which can be quite extensive for words such as participles and as a result increases data sparsity) we only include two features that we considered to be most relevant in a word's combinatorial behavior (and are therefore often mentioned in dictionaries such as Jones et al. 1996): case (nominative, accusative, dative, genitive, vocative) and mood (indicative, subjunctive, optative, imperative, infinitive, participle). In such a case a feature would look like θυγάτηρ/child/gen for "θυγάτηρ is a genitive with μήτηρ" (see Table 2 below for an illustration).

These syntactic models required us to implement a special treatment of prepositions and conjunctions on the one hand, and coordination structures on the other hand. In a sentence such as ἔρχομαι εἰς πόλιν "I go to a city", εἰς ("to") is treated in our syntactic corpus as a prepositional group with ἔρχομαι ("I go") and πόλιν ("city", accusative of πόλις) as the "object" of εἰς (which is in fact the relation that εἰς πόλιν has to ἔρχομαι). When it comes to determining the syntactic context of ἔρχομαι, one has four options: (1) εἰς, (2) πόλις, (3) both εἰς and πόλις, or (4) a single feature "εἰς πόλιν". Since we considered both εἰς and πόλις to be relevant for the meaning of ἔρχομαι, and since adding a single feature "εἰς πόλιν" would considerably reduce the influence of πόλις to the vector — there are many other prepositional groups with the same noun possible, such as ἀπὸ πόλεως "from the city", ἐκ πόλεως "out of the city" etc. — we preferred to count two context features in such a case, respectively "εἰς" and "πόλις". Secondly, the use of dependencies implies that coordination structures are somewhat awkwardly annotated: in a hierarchical representation it is much more straightforward to annotate subordination than coordination. In our representation, one coordinate has been made dependent of the other: i.e. in a sentence such as ἀκούω φωνὴν καὶ βοήν "I hear a voice and a scream" φωνή ("voice") is annotated as the object of ἀκούω "to hear", while βοή ("scream") is annotated as a conjunct of φωνή. Since we considered both the fact that βοή is an object of ἀκούω and that φωνή is coordinating with βοή to be relevant for the meaning of βοή, we again added two features for βοή in such a case, its technical head "φωνή" and the head of the whole group "ἀκούω".

Finally, since our corpus contains many proper names which would be less useful as either context features (the specific name would not matter except for some rare cases such as "Zeno's paradox") or target words (a vector for specific names, which are shared by several people who have little in common, would make little sense) we chose to replace all words starting with a capital letter simply by the lemma "NAME" (although in the future, it would be preferable to distinguish personal names such as "Socrates" from place names such as "Greece").

Table 2: Distributional models constructed for this study.

	Context	Head/child	Extra info	Example features
BOW	Window (size 4)	N/A	NO	μήτηρ, δίδωμι
DepMinimal	Dependencies	NO	NO	μήτηρ, δίδωμι
DepHeadChild	Dependencies	YES	NO	μήτηρ/child, δίδωμι/head
DepSyntRel	Dependencies	YES	Syntactic label	μήτηρ/child/ATR, δίδωμι/head/OBJ
DepMorph	Dependencies	YES	Morphology	μήτηρ/child/genitive, δίδωμι/head/dative

5 Evaluation of the context models

5.1 Main benchmark

Various benchmarks exists for the evaluation of distributional semantic models for Ancient Greek, described in Section 2.2. However, since they did not exist when the main research for this paper was carried out, and generally only contain lists of semantically related words without specifying in which way they are related, we decided to create our own benchmark, offering more detailed information about semantic relatedness (nevertheless, we will also offer results evaluated on these other benchmarks in Section 5.6). More concretely, we examined a sample of 100 lemmas – 50 nouns and verbs each – divided into 5 frequency bands, with 10 randomly chosen verbs or nouns in each band. We only selected lemmas with a frequency of at least 50 and chose to divide the frequency ranges for each band in such a way that the first group contains the 50% most frequent noun or verb tokens, the second group the next 25% most frequent tokens, the third group the next 12.5%, the fourth group the next 6.7% and the final group the remaining 6.7% tokens.[5] This resulted in the randomly chosen lemmas in Table 3.

Table 3: Words evaluated for the similarity task.

Band	Type	Freq.	Lemmas
1	Nouns	3600+	ἀλήθεια “truth”, πέρας “boundary”, ὄνομα “name”, πόλις “city”, ἀπορία “difficulty”, μάχη “battle”, ἀδελφός “brother”, αἰτία “cause”, ἡδονή “pleasure”, καρδία “heart”
1	Verbs	8000+	δοκέω “seem”, συμβαίνω “agree”, καλέω “call”, φημί “say”, ὁράω “see”, μένω “stay”, ἵστημι “stand”, πάρειμι “be present”, κρίνω “judge”, μανθάνω “learn”
2	Nouns	850-3600	συμφορά “accident”, ὀδούς “tooth”, κῦμα “wave”, σιωπή “silence”, ἔρις “strife”, ἄγαλμα “statue”, πλοῖον “ship”, ὗς “pig”, νεανίσκος “young man”, οὐλή “scar”
2	Verbs	1900-8000	ἀπαντάω “meet”, ἀφίημι “let go”, κατασκευάζω “equip”, ἀποκρίνω “answer”, τέμνω “cut”, συντίθημι “put together”, οἴχομαι “be gone”, γαμέω “marry”, βιάζω “force”, φιλέω “love”
3	Nouns	300-850	λοχαγός “commander”, ἄχος “distress”, ἶρις “iris”, ψάμμος “sand”, ἀνάμνησις “reminiscence”, προσευχή “prayer”, κωμῳδία “comedy”, ταμιεῖον “treasury”, ἠιών “shore”, δελφίς “dolphin”
3	Verbs	650-1900	χαρίζω “please”, ἀποστερέω “rob”, δανείζω “lend”, φορέω “wear”, ἀείρω “lift up”, ἀποτίθημι “put away”, μετέρχομαι “pursue”, ἀποτίνω “pay”, περιαιρέω “remove”, ἀπελαύνω “expel”
4	Nouns	150-300	παραφυλακή “guard”, ἱππόδρομος “chariot-road”, οἶστρος “frenzy”, ῥαφή “seam”, καλοκἀγαθία “nobleness”, πολεμιστής “warrior”, θήκη “case”, ἑστίασις “feasting”, σκοπιά “hill-top”, πέδιλον “sandal”
4	Verbs	250-650	εὐδαιμονέω “be prosperous”, ἀνασκευάζω “remove”, εὐθύνω “make straight”, κρούω “strike”, λῃΐζομαι “carry of as booty”, σκεπάζω “cover”, κατακρύπτω “hide”, ποιμαίνω “herd”, ἀναδείκνυμι “display”, δεξιόομαι “greet”
5	Nouns	50-150	ἀκρόαμα “anything heard”, ἅρπαγμα “booty”, στρύχνον “winter cherry”, γάρος “sauce”, πρόβασις “advance”, ἔλασις “driving away”, εὔπλοια “fair voyage”, εἰδωλολατρία “idolatry”, ὀποβάλσαμον “balsam”, ἰμάσθλη “whip”
5	Verbs	50-250	ἐναπολαμβάνω “intercept”, αὔω “shout”, προλείπω “abandon”, ἐπιβοηθέω “come to aid”, προκατασκευάζω “prepare beforehand”, ἐξισόω “make equal”, προαπαντάω “go forth to meet”, ἐπισυντίθημι “add successively”, ἐκθειάζω “deify”, ἐξοδιάζω “scatter”

[5] This seemed a good compromise to us instead of dividing the groups into five groups of an equal number of types (which would result in a first group consisting of several highly frequent and averagely frequent words, and the other groups consisting of only lowly frequent words), or an equal number of tokens (which would result in the first groups containing only a few very frequent items and the other groups containing all other items).

For each lemma, we calculated the cosine distance with all other remaining nouns/verbs of the full dataset, using the PPMI vectors of the models described in Section 4. Next, we examined the 10 nearest neighbors (i.e. the lemmas with the highest cosine similarity) of each lemma and annotated them with the following labels, which we considered to be useful to distinguish some very basic distinctions of semantic relatedness:

- **Synonym**: has a synonymous or near-synonymous meaning with the target lemma. E.g. *νεανίσκος – νεανίας* (both "young man") or *κρούω – τύπτω* (both "strike, knock").
- **Related**: while the words are not strictly synonymous, they are closely semantically and syntactically related, for instance because they share a hypernym or one word is the hypernym of the other (i.e. there is a taxonomical relationship between the two words). E.g. *νεανίσκος – παρθένος* ("young man" – "young woman") or *κρούω – ὠθέω* ("strike" – "thrush").
- **Distantly-related**: there is a vague resemblance between the two words, but they share a hypernym higher up in the ladder, and as a result they will still frequently occur in the same syntactic environments. E.g. *νεανίσκος – στρατιώτης* ("young man" – "soldier") or *κρούω – ἀείδω* ("strike (often musically)" – "sing").
- **Same domain**: while there is no shared hypernym between the two words, they still often occur in the same thematic contexts (the relation is more associative). E.g. *νεανίσκος – ἡλικία* ("young man" – "youth") or *κρούω – ὀρχέομαι* ("strike (often musically)" – "dance").
- **Unrelated**: there is no overlap in syntactic or thematic contexts. E.g. *νεανίσκος – δῆμος* ("young man" – "populace") or *κρούω – ἵστημι* ("strike" – "stand").

The data were annotated by an independent researcher on Ancient Greek linguistics, starting from the meanings described in the LSJ lexicon of Greek (Jones et al. 1996). Since in most cases there is only partial overlap in meaning between words, overlap with any meaning was checked, e.g. when there was synonymy with at least one meaning (even though the two words might not be synonymous in all meanings) the label "synonym" was used (and similarly for "related" and so on).[6] Since the training data of the distributional model contains a very long time span (16 centuries) and various text genres, polysemy was considered for the full ranges of uses of a word over time and genre: i.e. two words were also called 'synonymous' if they had one meaning that was synonymous, even if this meaning was only present in certain periods or text genres.

[6] For comparative purposes, we also annotated the data ourselves to evaluate how much of the differences described in this section are simply due to the subjectivity of the annotation. Our labeling only overlapped with the independent one in 45.5% of all cases (1012/2226), Cohen's kappa = 0.312 (although in an additional 36% of cases the difference was only one level). Nevertheless, the general tendencies described in this section still hold, although the effect of frequency (see 5.3) was a little stronger in our annotation.

5.2 Main results

The following tables detail the general results we found with each syntactic model. For the top 10 we looked at 500 nearest neighbors in total for each model (the 10 nearest neighbors of 10 verbs per frequency band, with 5 frequency bands in total) and for the top 5 the 250 nearest neighbors.

Table 4: Classification of 10 nearest neighbors among verb distributional models.

Top 10 - Verbs	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.142	0.184	0.178	0.192	0.304
DepMinimal	0.160	0.192	0.214	0.186	0.248
DepHeadChild	0.162	0.188	0.216	0.200	0.234
DepSyntRel	0.140	0.192	0.222	0.214	0.232
DepMorph	0.164	0.192	0.226	0.176	0.242

Table 5: Classification of 10 nearest neighbors among noun distributional models.

Top 10 - Nouns	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.088	0.255	0.335	0.244	0.078
DepMinimal	0.108	0.296	0.356	0.166	0.074
DepHeadChild	0.104	0.318	0.336	0.166	0.076
DepSyntRel	0.102	0.324	0.324	0.160	0.090
DepMorph	0.090	0.326	0.316	0.170	0.098

Table 6: Classification of 5 nearest neighbors among verb distributional models.

Top 5 - Verbs	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.180	0.212	0.180	0.176	0.252
DepMinimal	0.212	0.228	0.204	0.164	0.192
DepHeadChild	0.180	0.208	0.244	0.172	0.196
DepSyntRel	0.188	0.212	0.224	0.212	0.164
DepMorph	0.212	0.232	0.192	0.176	0.188

Table 7: Classification of 5 nearest neighbors among noun distributional models.

Top 5 - Nouns	Synonym	Related	Distantly-related	Same domain	Unrelated
BOW	0.104	0.284	0.356	0.180	0.076
DepMinimal	0.148	0.312	0.356	0.120	0.064
DepHeadChild	0.148	0.356	0.300	0.140	0.056
DepSyntRel	0.148	0.380	0.276	0.124	0.072
DepMorph	0.120	0.384	0.304	0.112	0.080

These data first and foremost reveal that there is a clear difference between the bag-of-words model on the one hand and the syntactic models on the other hand: syntactic models prove to be better suited to return synonyms and closely related words than the former. Although the number of totally unrelated words does not differ that much for nouns, the bag-of-words model returns several more words that are only tangentially or associatively related ("same domain"), which corroborates the findings mentioned in

Section 4. For verbs there were no real differences for the "same domain" label, but it is more difficult to say when a verb belongs to the same domain as another verb (since the meaning of a verb tends to be more abstract and/or vague than that of a noun). Consequently, this might simply be an effect of the annotation: the annotator might have been more disposed to say that two nouns belong to the same domain than in the case of verbs. On the other hand, the number of totally unrelated words is clearly higher for BOW in the verb category than for the syntactic models. Within the four syntactic models, however, there is far less differentiation, with only a one or two percent difference for most categories, and no consistent best performing model. We will analyze the reason for this lack of clear differences below.

5.3 Effect of frequency

The following plots detail the effect of frequency by counting the percentage of **synonymous** and **related** words in the 10 nearest neighbors (N=100 per frequency band) – since many words do not have direct synonyms, it makes more sense to consider both in the evaluation of the performance of the different models.

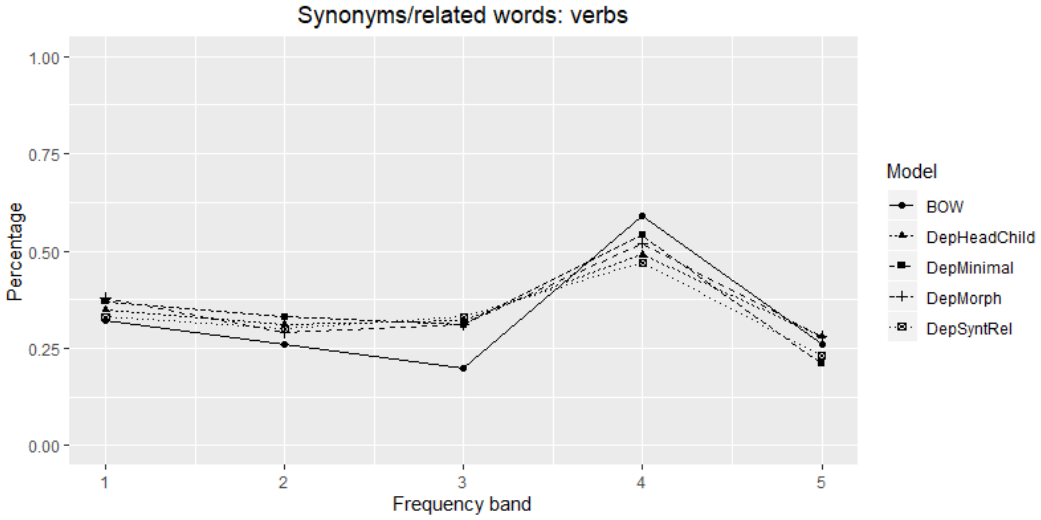

Figure 1: Percentage of synonyms/related words in 10 nearest neighbors by frequency band (verbs).

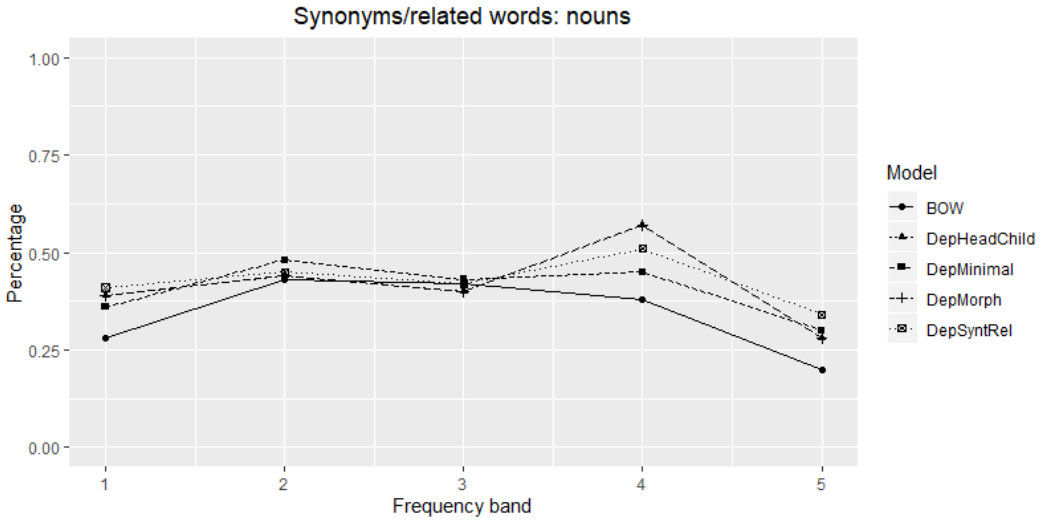

Figure 2: Percentage of synonyms/related words in 10 nearest neighbors by frequency band (nouns).

For almost each model (except the verbs *BOW* model) frequency band 5, containing the lexical items with the lowest frequencies, returns the least number of synonymous/related words in the nearest neighbors. Interestingly, however, the words in the highest frequency band do not seem to substantially outperform the ones in the second to fourth frequency band (or perform even worse, in the case of the nouns). This might possibly suggest a diminishing effect of frequency, i.e. as long as the distributional vectors contain enough observations, adding more data would not have a large effect anymore. Another factor to take in mind is that the highest frequency band contains several words with a quite general and/or abstract meaning, which makes their meaning more difficult to model (see below). These frequency effects seem to be relatively consistent across all 5 distributional models, and any differences are probably caused by random fluctuations.

5.4 Causes of the differences between the various context models

There are two reasons that may explain the limited differentiation between the syntactic models: either these models return the same types of words, or they do not, but the drawbacks of a certain model cancel out its benefits. In order to establish which of these two situations applies, we investigated the degree of overlap of the words that are in the 10 nearest neighbors, as shown in Table 8 (since the numbers for nouns and verbs were almost identical, we did not separate them).

Table 8: Degree of overlap between 10 nearest neighbors returned by each model.

	BOW	**DepMinimal**	**DepHeadChild**	**DepSyntRel**	**DepMorph**
BOW		54%	52%	43%	42%
DepMinimal	54%		73%	53%	51%
DepHeadChild	52%	73%		61%	56%
DepSyntRel	43%	53%	61%		64%
DepMorph	42%	51%	56%	64%	

This table demonstrates that there is not a high degree of overlap between the nearest neighbors returned by the bag-of-words models on the one hand and the syntactic models on the other hand, with especially the models with syntactic or morphological specification (i.e. *DepSyntRel* and *DepMorph*) returning rather different words. Secondly, there is quite a big degree of overlap between *DepMinimal* and *DepHeadChild*, but far less so with *DepSyntRel* and *DepMorph*. In other words, the lack of quantitative differences between the performance of the different models seems to mask the fact that they do in fact return quite different words in their nearest neighbors.

To further investigate the differences among the vector models, we examined the vectors of the nearest neighbors as compared to the ones of the target words, and identified which features have a high PPMI value in both vectors: these features would have a high influence on the cosine metric. More precisely, we selected a number of pairs of target words and nearest neighbors that were not synonymous or related

(to gain a deeper understanding on why these "erroneous" nearest neighbors words were retrieved). Next, we listed a number of features that were in the top 5% of highest PPMI values for both vectors. Table 9 summarizes these results, containing a (random) selection of these high-ranking features. For comparative purposes, we kept the target word constant.

Table 9: Features in top 5% of PPMI values for target words and their 'erroneous' nearest neighbors.

Model	Target word	Neighbor	Example features
BOW (Nouns)	σιωπή "silence"	δικαστής "judge"	καθέζομαι "sit down", συκοφαντία "sycophancy", ἐνθυμέομαι "desire", φρίκη "shivering", ἡρωικός "heroic", ἀκροάομαι "listen to", μητρυιά "stepmother", ἀτρεμέω "keep still"
BOW (Verbs)	ὁράω "see"	φεύγω "flee"	ὀσφραίνομαι "smell", βδελύσσομαι "be loathsome", περιβλέπω "look around", προσπλέω "sail toward", αἱμάσσω "make bloody", ἐνεργάζομαι "produce in", ἱππότης "horseman", γλαυκός "gleaming"
DepMinimal (Nouns)	σιωπή "silence"	δῆμος "populace"	καταδικάζω "convict", καταψηφίζομαι "vote against", εὐταξία "good order", καρτερέω "be steadfast", νεανίας "young man", κλέω "celebrate", στένω "groan", θαῦμα "wonder"
DepMinimal (Verbs)	ὁράω "see"	κάθημαι "sit"	ἐπιποθέω "desire", πτήσσω "scare", ἀσχημονέω "disgrace oneself", ὀλιγάκις "seldom", ἀποδειλιάω "be fearful", προσελαύνω "drive to", κρεμάννυμι "hang"
DepHead-Child (Nouns)	σιωπή "silence"	κίνδυνος "danger"	ἀσφαλής/head "safe", ὑποσημαίνω/head "indicate", συνωθέω/head "force together", ἐπιρριπτέω/head "throw oneself", καρτερέω/head "be steadfast", ὑποπτεύω/head "suspect", γοῦν/child "at any rate", πνίγω/head "choke"
DepHead-Child (Verbs)	ὁράω "see"	ἵστημι "stand"	πόρρωθεν/child "from far", μακρόθεν/child "from far", πρόσφημι/head "speak to", ἄντα/child "over against", ἐγγύθεν/child "from far", διαταράσσω/head "confuse", κάθημαι/child "sit", ὀρχέομαι/child "dance"
DepSyntRel (Nouns)	σιωπή "silence"	χρόνος "time"	ἐξίστημι/head/adverbial "change", καιρός/head/coordinate "time", κατέχω/head/adverbial "hold fast", ἀγανακτέω/head/adverbial "be irritated", παραδίδωμι/head/adverbial "hand over", ὑβρίζω/head/adverbial "maltreat", ἔξεστι/head/adverbial "be possible", δουλεύω/head/adverbial "serve"
DepSyntRel (Verbs)	ὁράω "see"	φημί "say"	ἀμελέω/child/object "neglect", γελάω/child/object "laugh", ἐπαίρω/child/object "raise", ταράσσω/child/object "disturb", ἡσσάομαι/child/object "be inferior", ὁρμάω/child/object "start", κλαίω/child/object "weep", διαλέγομαι/child/object "converse"
DepMorph (Nouns)	σιωπή "silence"	βία "violence"	παρέρχομαι/head/dative "pass by", καταψηφίζομαι/head/accusative "vote against", ὄχλος/child/genitive "crowd", κατέχω/head/dative "hold fast", παρίημι/head/dative "let go", ἀποδέχομαι/head/genitive "accept", ὑπείκω/head/dative "withdraw", συλλαμβάνω/head/dative "collect"
DepMorph (Verbs)	ὁράω "see"	εὑρίσκω "find"	κάθημαι/child/participle_accusative "sit", ἀναβαίνω/child/participle_accusative "go up", χαλεπός/head/infinitive "difficult", ἵστημι/child/participle_accusative "stand", διάκειμαι/child/participle_accusative "be", ῥίπτω/child/participle_accusative "throw", ἔρχομαι/child/participle_accusative "go", προσέχω/child/participle_accusative "offer"

These data show that using a simple bag-of-words context model can lead to a large number of spurious associations. The association between δικαστής "judge" and μητρυιά "step-mother", for instance, is based on the frequent use of the two words in a rhetorical speech without there being any direct link between the words (e.g. *ἄχθομαι μὲν οὖν , ὦ ἄνδρες* ***δικασταί****, ἐπὶ τῇ* ***μητρυιᾷ*** *χαλεπῶς ἐχούσῃ* "I am in pain, **men of the jury**, because my **stepmother** is doing badly"). Similarly, the association between

γλαυκός "gleaming" and φεύγω "flee" is based on contexts in which the object of flight is described as γλαυκός, e.g. ***γλαυκοῖο φυγὼν*** *Τρίτωνος ἀπειλὰς* "**fleeing** the threats of the **gleaming** Trito". It is exactly these kinds of associations that the dependency-based models filter out.[7]

Examining the differences between the *DepMinimal* and *DepHeadChild* model, we can observe that in many cases it is quite obvious what the direction of the arc should be without knowing it in advance. For instance, a verb such as καταδικάζω "convict" would typically be the head of a noun such as σιωπή "silence" and δῆμος "people" and not its child, and an adverb such as ὀλιγάκις "seldom" would typically be the child of a verb such as ὁράω "see" and κάθημαι "sit" rather than its head, so adding the direction of the arc would be superfluous. In some cases adding the direction of the arc might even be detrimental. To give an example, nouns will typically be the head of relative clauses or attributive participles, while in a main clause they would be considered a child of the respective verb. Both ὁράω and θεάομαι "see", for instance, have a feature κάλλος/head "beauty" with a high PPMI value from sentences such as ***κάλλος*** *οἷον οὔπω πρότερον* ***ἐτεθέατο*** "such a **beauty** as he **had never seen** before", in which ἐτεθέατο (from θεάομαι) is considered to be the child of κάλλος, even though it also functions as the object of the relative clause. As a result, in such cases grouping these instances under a single feature "κάλλος" would be more effective.

Even in cases in which there is a clear hierarchical relationship, it is not obvious if this hierarchy is always relevant: in cases with adverbial clauses or participle groups, for instance, such as **ἀναβλέψας** τοῖς ὀφθαλμοῖς **εἶδεν** αὐτὸν τὸν τόπον "**looking up** with his eyes he **saw** this place" it is clear that the fact that the participle ἀναβλέψας (of ἀναβλέπω, "look up") is in a dependency relationship with εἶδεν (of ὁράω, "see") is relevant for the meaning of ὁράω, but it is less obvious that the fact that ἀναβλέψας is a child of εἶδεν is equally meaningful (a sentence such as ἀνέβλεψε τοῖς ὀφθαλμοῖς καὶ εἶδεν αὐτὸν τὸν τόπον "he **looked up** with his eyes and **saw** this place" would roughly convey the same meaning). This is not to say that the fact that ἀναβλέπω is in a subordinate relationship is entirely meaningless (otherwise the writer would obviously not have chosen to encode such a subordinate relationship explicitly by the use of the participle), but this might not be an aspect of meaning that is particularly useful to detect word similarity.

However, the direction of the arc is certainly not irrelevant in all cases. For instance, in the list of words that have a high PPMI value with both σιωπή "silence" and δῆμος "people" in the *DepMinimal* model, we can find nouns such as ὄχλος "crowd", for which ὄχλος is usually the head (or in a coordinate relationship) in the case of δῆμος (e.g. **ὄχλοι** παντοίων **δήμων**: "crowds of all sorts of people"), but in

[7] Of course such less direct dependency links might sometimes be informative as well: in a sentence such as "fleeing the dangerous men", for instance, the word "dangerous" does provide useful information about the meaning of "flee". One possible way to include such contexts is to include indirect paths as well (such as *flee* > *man* > *dangerous*) and weigh the paths according to their length (as well as the type of syntactic relation), see Padó and Lapata (2007). Meanwhile, words which have no dependency path at all between them, such as δικαστής and μητρυιά in the example above, would still be excluded.

the case of σιωπή it usually is a child (e.g. τῶν **ὄχλων** ἡ **σιωπή**: "the silence of the crowds") – "a crowd of silence" would be atypical. As there is little difference in performance between the two models, the advantages to explicitly code the dependency link on the feature seem to be as important as the drawbacks. Therefore a model that combines the strengths of both models would be preferable, i.e. only encode head/child information when it helps to make relevant semantic distinctions and not when it is e.g. simply related to specific conventions of the dependency-based format.

One way to further refine the dependency-based models is to add further syntactic and morphological labels to it, such as in the *DepSyntRel* and *DepMorph* models. However, a negative effect of such an approach would possibly be data sparsity, seeing that it further subdivides a given feature in several new features which each would be less frequently attested than the feature without label, and we are dealing with a relatively small corpus to start with. This would not be a problem if there was no connection between several syntactic uses of a word, if e.g. the "adverbial" use of word X would be entirely different in meaning from its "object" use: in such a case making this sub-distinction would only help to model meaning distinctions. However, this is clearly not always the case: looking at e.g. the top 5% of features with the highest PPMI values for both σιωπή and σιγή (both "silence"), we see several reoccurring features with a different syntactic label such as κατέχω/adverbial and κατέχω/subject, ἀκούω/adverbial and ἀκούω/object, and so on. One issue is that a specific semantic role can be encoded in different syntactic constructions, such as the patient, which would be encoded as the subject of an active verb but the object of a passive verb. Another issue is that the boundaries between labels such as "object" and "adverbial" are often rather fluid, which becomes increasingly problematic when dealing with an automatic parsing system. While this latter problem is not relevant for constructions that use morphology instead of syntactic relations, the problem of using different syntactic constructions to encode the same semantic role still arises there.

Finally, we can also see an important difference in the type of semantic information that is encoded in the *DepSyntRel* and *DepMorph* models as opposed to the other syntactic models. There does seem to be a greater emphasis on constructions that show a similar syntactic behavior: the nearest neighbors of ὁράω show a large number of verbs that are more broadly situated in the evidential domain rather than especially connected with acts of seeing such as φημί "claim", οἶδα "know", μανθάνω "learn", νομίζω "think" and so on. Looking at the shared features with high PPMI values, almost all of them are verbal objects, denoting some kind of information that is manipulated, e.g. **ἰδοῦσα** δὲ τὰς αἶγας **τεταραγμένας** "**seeing** that the goats **had been disturbed**" and **τεταράχθαι** μὲν αὐτὴν […] **ἔφη** μοι ἡ Θεοπάτρα "Theopatra **said** to me that she **had been disturbed**". Using morphology instead of syntactic labels further emphasizes the high co-occurrence of ὁράω with participial complementation, which is considered to be more objective than infinitival complementation: therefore verbs such as νομίζω "think" are pushed down from the 6th position in the list of nearest neighbors (with *DepSyntRel*) to the 41st (with *DepMorph*), while verbs such as εὑρίσκω "find" appear in the top 10, from constructions such as **εὑρὼν**

παῖδα τὸν ἐμὸν **καθήμενον** "**finding** my child **sitting down**" which are quite comparable to something like τὸν Κροῖσον αὐτὸν **ὁρᾷς** ἤδη ἐπὶ κλίνης χρυσῆς **καθήμενον** "you already **see** Croesus himself **sitting** down on a golden throne". In such constructions the meaning of ὁράω is in fact quite similar to εὑρίσκω, but the use of such syntactic and morphological features might overemphasize this specific aspect of the meaning of these verbs as opposed to other usages. Similarly, most features of σιωπή in *DepSyntRel* are related to its usage as an adverbial (specifically of manner). Since the label "adverbial" is used as a catch-all term for all sorts of adverbial relations, this can explain the high cosine similarity with χρόνος, which is similarly often used with an adverbial function, even though it is a quite different adverbial relation (of duration rather than manner). Using the morphological rather than the syntactic label further narrows it to usages with the dative case, which is common for manner adverbials (duration is typically expressed in the accusative), but the dative case is still quite broad and can be used to express all sorts of other semantic roles such as instrument (which would be the typical semantic role for βία "violence"). In other words, it is clear that the use of syntactic and morphological features does reveal aspects of meaning that are not present in other models, but it is less obvious that this information is also appropriate for tasks such as word similarity detection.

5.5 Performance with specific words

Next, we took a closer look at how well the models performed overall with specific words. Table 10 summarizes the average performance of some select noun classes across all five word models (the standard deviations per category are between brackets), see 'Supplementary material' for the full results. Starting with nouns, one category of nouns that performs particularly well are words in the natural domain: καρδία "heart", ὀδούς "tooth", ὗς "pig", ἶρις "iris flower", ἠιών "shore", δελφίς "dolphin", σκοπιά "hill-top", στρύχνον "winter cherry" and ὀποβάλσαμον "balsam" return many synonyms or related words in their nearest neighbors, although this is the less the case with κῦμα "wave", οὐλή "scar" and ψάμμος "sand". As a general category, however, these words are clearly easier to model than other nouns, as can be seen in Table 10: the ratio related vs. unrelated words is clearly considerably higher than average (while they return less synonyms, this is probably because most of these words are so specific that they do not have a large number of synonyms to start with). Another group of nouns that seems to be modelled well are nouns referring to people, i.e. ἀδελφός "brother", νεανίσκος "young man", λοχαγός "commander" and πολεμιστής "soldier". However, one of these words (πολεμιστής) performs somewhat worse than average, this category does not contain many words to start with, and the words in this category do have a higher token frequency than average. Concrete objects/structures also perform a little better than average (ἄγαλμα "statue", πλοῖον "ship", ταμιεῖον "treasury", ἱππόδρομος "chariot-road", θήκη "case", πέδιλον "sandal" and ἰμάσθλη "whip"), while qualities or emotions (ἀλήθεια "truth", ἡδονή "pleasure", ἔρις "strife", ἄχος "distress", οἶστρος "frenzy", καλοκἀγαθία "nobleness") perform about average. Finally, the words that are clearly the most difficult to model refer to events or processes: μάχη "fight", συμφορά "accident", σιωπή "silence", ἀνάμνησις

"remembrance", προσευχή "prayer", παραφυλακή "guard", ἑστίασις "feasting", πρόβασις "increase", ἔλασις "driving away", εὔπλοια "fair voyage" and εἰδωλατρία "idolatry". This is slightly skewed by the outlier παραφυλακή (see also below), which returns on average 7.4 unrelated words, but most of them also have a lower than average ratio of related vs. unrelated words.

Table 10: Mean classification of 10 nearest neighbors per word class, with standard deviations between brackets.

	Synonym	Related	Distantly-related	Same domain	Unrelated
AVERAGE	1.0 (1.2)	3.0 (2.0)	3.3 (2.0)	1.8 (1.5)	0.8 (1.4)
Natural domain (N=12)	0.4 (0.6)	4.1 (2.4)	4.1 (2.3)	1.1 (1.3)	0.3 (0.5)
People (N=4)	0.7 (0.7)	4.2 (1.7)	3.2 (1.1)	1.7 (0.9)	0.3 (0.4)
Concrete objects (N=7)	2.0 (1.8)	2.3 (1.5)	3.6 (1.5)	1.7 (1.2)	0.4 (0.7)
Qualities/emotions (N=6)	0.9 (1.1)	4.5 (2.1)	3.0 (3.0)	0.8 (1.0)	0.8 (1.1)
Events/processes (N=10)	0.9 (1.0)	2.4 (1.4)	2.9 (1.4)	2.2 (1.4)	1.6 (2.1)

As for verbs, it is more difficult to exactly pinpoint a number of semantic classes that perform well, since the results seem more random there. There are some tendencies, however: many verbs that are easy to model refer to some concrete physical action such as οἴχομαι "go away", ἀπελαύνω "drive away", σκεπάζω "cover", κρούω "knock" and λῃΐζομαι "plunder". Verbs that belong to the mental domain also perform well (although they are all very frequent) such as δοκέω "seem", μανθάνω "learn" and κρίνω "judge". Other than that, there are no clear tendencies, although some bad-performing verbs are semantically quite vague or abstract, or have wide-ranging meanings, such as συμβαίνω (for which the LSJ dictionary lists meanings ranging from "stand with the feet together" to "come to an agreement", "correspond with", "to be an attribute of", "happen" and so on), προαπαντάω ("go forth to meet", "take steps in advance" or "to be interposed") and ἀνασκευάζω ("pack up the baggage", "remove", "ravage", "to be bankrupt", "reverse a decision", "build again").

For verbs, these differences are probably best explained by their general semantic properties: it is not surprising that verbs that are semantically quite specific and concrete, e.g. physical contact verbs such as σκεπάζω "cover", would have more useful context information than very ambiguous verbs such as συμβαίνω (see above), of which its meanings might be too disparate to model with a single vector. Animacy might also be a factor: verbs that have human objects might typically use pronouns or proper names to refer to these human referents, while these physical contact verbs typically have concrete non-animate objects, which might provide these models with more useful context information. This could also explain why verbs with typically verbal complements such as cognitive verbs are modelled well, since these complements are directly expressed as well. This is simply a hypothesis, however, that should be further explored in future research.

As for nouns, the same principles generally hold: nouns that are referentially more abstract such as nominalized processes tend to be modelled quite badly, while very concrete nouns perform well.

However, especially for nouns the influence of genre also seems to be an important factor. The most prominent example are nouns that typically belong to the scientific or natural domain, which were the easiest to model, as discussed above. We can give several reasons for this: first of all, there are many scientific texts in the Greek corpus. The works of four authors, i.e. Galen (medicine), Hippocrates (medicine), Aristotle (philosophy, including biology and physics) and Theophrastus (botany), together consist of 4.6 million tokens, or 1/8 of the total corpus. Secondly, such nouns tend to be well-demarcated, which makes them easier to model than more abstract concepts. Finally, these texts tend to be "definitional", i.e. they precisely try to describe the concept under question, and as a result many useful context features are provided. See, for instance, some occurrences of the word ἶρις "iris" in Theophrastus's *Enquiry into Plants*:

(1) ἀνθεῖ δὲ καὶ ἡ **ἶρις** τοῦ θέρους καὶ τὸ στρούθιον καλούμενον· (…) ὁ μὲν ἀσφόδελος μακρὸν καὶ στενότερον καὶ ὑπόγλισχρον ἔχει τὸ φύλλον, (…), ἡ δὲ **ἶρις** καλαμωδέστερον· (…) ἔνια δὲ ἔχει, καθάπερ ἡ σκίλλα καὶ ὁ βολβὸς καὶ ἡ **ἶρις** καὶ τὸ ξίφιον· (Theophrastus, *Enquiry into Plants* 6.8.3)
The **iris** also blooms in summer, and the plant called soap-wort; (…) Asphodel has a long leaf, which is somewhat narrow and tough, (…), and **iris** one more like a reed. (…) some however have a stem, as squill purse-tassels **iris** and corn-flag (translation A. Hort).

The context features we find in those sentences are clearly suited to demarcate the meaning of ἶρις, e.g. ἀνθεῖ "blooms", καλαμωδέστερον "more like reed", and other flowery plants ἶρις coordinates with such as σκίλλα "squill", βολβός "purse-tassels" and ξίφιον "corn-flag".

Having more data for a given lemma obviously helps to model its meaning. However, this needs to be nuanced in two ways. First of all, there are situations in which having more data can be more detrimental, if the type of data is not really suited to model the meaning of the target word. This is, for instance, the case for παραφυλακή "guard", which occurs in the majority of its usages in the papyri (124/149 times) in contexts such as the following:

(2) παρὰ Αὐ]ρηλίου Παπνουθίου Πκυλίου μητρὸς [. . . . .]ιας ἀπὸ ἐπ[οι]κείου Σεντοποιὼ ὕπο [τὴν **παρα]φυλ̣ακ̣ὴν** τ̣[ῶ]ν̣ ἀπὸ κώμης Πτι[μενκυρκ]εω[ς] Ι̣ρ̣ιμ̣έν̣[ων] τοῦ Ἑρμουπολίτο[υ νομοῦ] (BGU 6 1430)
"Of Aurelius son of Papnuthius son of Pkylius, his mother […], from the hamlet Sentapouo under the **guard** of the Shepherds from the village Temencyrcis from the Hermopolites nome"

(3) ἐν περιχώματι Τραισε ὑπὸ τὴν **παραφυλακὴν** τῶν ἀπὸ κώμης Ἄρεως τοῦ Ἑρμουπολίτου νομοῦ (SB 14 11373)
"(…) in the Traise dyke under the **guard** of the people from the Areos village of the Hermopolites nome"

(4) συσταθεὶ]ς ὑφ’ ὑμῶν εἰς **παραφυλακ(ὴν)** [τῆς μητρο]πόλεως (P. Ryl. 2 88)
"(…) being assigned by you for the guard of the metropolis"

While there are some context elements that may be useful to model the meaning of παραφυλακή, i.e. κώμης "village" and μητροπόλεως "metropolis", in general these texts are quite formulaic, which has as a result that the same construction might be repeated several times, as in (2) and (3), and that these contexts might be quite generic (especially in texts such as contracts), e.g. "this person has done so and so in this place at this time", as opposed to contexts such as (1). In other words, it is not only the quantity of the data that matters, but the quality as well: some types of data are clearly more suited to model lexical semantics than others.

Finally, even if we have a large amount of data with useful context features, the vectors we calculate might not always encode the desired semantic information. For instance, looking at the nearest neighbors of words such as πρόβασις "increase" and ἐπισυντίθημι "add successively", we can see that most words are in the mathematical domain: e.g. διάμετρος "diameter", ἀριθμός "number" and περίοδος "period, circumference" for πρόβασις and πολλαπλασιάζω "multiply", διπλόω "double" and μερίζω "divide" for ἐπισυντίθημι. This is probably caused by the fact that the Greek corpus contains a large amount of mathematical material, with a specialized vocabulary (therefore these context features will receive high PPMI values), which pulls the vector toward the mathematical meaning of the word. However, these words have non-technical meanings as well, which might be subdued due to this factor – also note that in our evaluation we considered a word to be "synonymous" or "related" if this was true for at least one meaning, so the fact that some vectors might be "skewed" towards a particular meaning is not measured by the metrics we used above. There are multiple ways to resolve this issue: either by selecting or weighting parts of the corpus so that these non-technical meanings would also be represented, or by abandoning the use of one single vector to represent all meanings and either constructing vectors for specific genres or working with token-based models (see De Pascale 2019 for an application of both strategies in the context of dialectology). At any rate, it is necessary to take a closer look at the question of how the heterogeneity of the Greek corpus impacts the composition of our vector representation in the future.

5.6 Comparison with other benchmarks

As noted in Section 5.1, various other benchmarks for the evaluation of distributional semantic models for Ancient Greek exist, including Ancient Greek WordNet (Bizzoni et al. 2014), an automatically created WordNet for Ancient Greek based on bilingual English-Greek dictionaries, Justus Pollux's Onomasticon, an ancient work from the second century AD describing semantically related words, Schmidt's (1876-1886) *Synonymik der griechischen sprache* containing lists of Ancient Greek synonyms, and the AGREE benchmark (Stopponi et al. 2023), containing measures of word relatedness

scored by various independent researchers.[8] All of these benchmarks consist of lists of related words, while the AGREE benchmark also contains a score from 0 to 100 how related these words were considered on average by the scholars.

To evaluate, for each pair that was considered semantically related in the benchmarks, we calculated how high each word of the pair appeared in the list of semantically related words (descending by cosine) of the other word. After that, we calculated the median of all these rankings as a metric of how well the models are able to detect closely semantically related words (we used median instead of mean since in many cases the ranking was very low, which would have a large effect on the mean).[9] Additionally, for the AGREE benchmark, we calculated the correlation between the ratings of experts and the cosine similarity of the distributional models, using Spearman correlation. The results are presented in Tables 11-12 (since the benchmarks based on Schmidt and Pollux did not contain verbs, only results for nouns are presented there).

Table 11: Median rank of semantically similar word pairs according to the benchmarks among each other's neighbors returned by each word model (not SVD-scaled).

		BOW	DepMinimal	DepHC	DepSyntRel	DepMorph
WordNet	**Nouns, N=11631**	764	**694**	701	747	748
	Verbs, N=33015	1228	1188	1182	**1177**	1185
Pollux (Nouns, N=2631)		527.5	**463**	490	547.5	585.5
Schmidt (Nouns, N=2793)		238	**209**	**209**	248.5	278.5
AGREE	**Nouns, N=129**	23	**22**	**22**	25	33
	Verbs, N=67	31	**18.5**	23.5	27	22.5

Table 12: Spearman correlation of model ratings and expert ratings in the AGREE benchmark.

	BOW	DepMinimal	DepHC	DepSyntRel	DepMorph
Nouns, N=226	**0.538**	**0.538**	0.529	0.505	0.511
Verbs, N=234	0.370	0.414	0.420	**0.441**	**0.441**

[8] The three first resources were used by Rodda et al. (2019), as noted in Section 2.2, and a digital (greatly abridged) version of the Onomasticon and Schmidt's lexicon were compiled by them (*https://github.com/alan-turing-institute/ancient-greek-semantic-space*). Ancient Greek WordNet can be found at *http://hdl.handle.net/20.500.11752/ILC-56*. The AGREE benchmark is found at *https://zenodo.org/record/7681749*.

[9] Both Rodda et al. (2019) and Stopponi et al. (2023) evaluate similarity in terms of precision and recall, comparing the word pairs in the benchmarks to the k (5, 10, 15) nearest neighbors of these target words in distributional models. Recall represents how many of the related words in the benchmarks were included in the list of nearest neighbors, while precision represents how many of the nearest neighbors were included in the benchmarks. However, this seemed problematic to us as 1) the benchmarks are not generally exhaustive, complicating the calculation of precision, since the nearest neighbors might contain related words that are not included in the benchmarks, 2) k is an arbitrary choice, and some words might have much more related words than others and 3) some words might have less than k related words in the benchmarks, complicating the calculation of recall (as also noted by Stopponi et al. 2023).

Firstly, regarding the different benchmarks, the AGREE benchmark was the only resource explicitly created to evaluate distributional models, and it is clear that it is much more suited for this than the other benchmarks: the median word ranked much higher in the list of nearest neighbors than for any other benchmark we evaluated. The only benchmark that was somewhat close was Schmidt's *Synonymik*, and the median word included there was still very low in the list of nearest neighbors of its supposed semantically related words (at place 237 on average across all models) when compared to AGREE (at place 25 on average across all models for nouns). The median rank was in particular very low with Ancient Greek WordNet: Rodda et al. (2019) also noted that this resource did not match the results of their distributional models very well, which is likely an artifact of the substantial level of noise introduced by the automatic creation of this resource.

Regarding model performance, unlike the results discussed in the previous sections, in general there is not a large difference between bag-of-words models and syntactic models when evaluated against these benchmarks, with the bag-of-words model in several cases even performing best. This is true for both the comparison of the ranks of semantically related words as well as the correlation between experts' and models' ratings (although there seems to be a difference between nouns and verbs). Inspecting the data more closely, this is likely because several of these benchmarks contain words that are only related in a very topical way. For example, focusing on the differences between the *BOW* and *DepMinimal* model with the AGREE benchmark (the best performing benchmark), some word pairs that occur much lower on average in each other's list of nearest neighbors in *DepMinimal* vs. *BOW* are νόστος ('return home') vs. θάλασσα ('sea'), νόστος ('return home') vs. ὁδός ('way'), πατήρ ('father') vs. σέβας ('respect'), as well as words that are clearly very closely semantically related but will have a very different syntactic behavior, such as πόντος ('sea') vs. ἁλιεύς ('fisherman'), ῥῆσις ('speech') vs. ἀγορά ('marketplace', 'assembly'), and πρέσβυς ('old man') vs. ἡλικία ('age').[10]

6 Discussion

The aim of this study was to test the validity of distributional semantic models for Ancient Greek – and presumably, the results can be expanded to other highly inflectional and historical languages as well – in particular by focusing on the type of context features that are suited best to model lexical semantics. These context features involved an increasing level of analysis, ranging from (1) a simple 4 words window bag-of-words model, to all words that are in a dependency relationship, both excluding (2) and including (3) the direction of the dependency arc and the dependency relationship with a syntactic (4) and morphological (5) label (see Table 2).

[10] Additionally, there were some words that are morphologically nouns but semantically adjectives that are typically combined with the other noun in the pair, such as ναῦς ('ship') vs. *κορωνίς* ('curved') and πόντος ('sea') vs. οἶνοψ ('wine-colored').

To evaluate the results of these different distributional models, we investigated how useful the (raw, PPMI-weighted) vectors are to detect word similarity, and what types of similarity they detected, by a (subjective) labeling of the nearest neighbors retrieved by each vector model. We found that dependency-based vectors are much better suited to return synonymous and/or taxonomically related words than a simple bag-of-words context model. This is especially striking since we used automatically parsed data, which still had a considerable error rate. The importance of using syntactic dependencies is likely caused by the free word order of Greek, since the relevant contextual information might not always be present in a small context window of preceding or following words.

Among the different dependency-based models, on the other hand, the differences are less pronounced. There are several reasons for this: (a) some technicalities of the dependency format (e.g. how coordination structures are encoded) create differences that are linguistically meaningless; (b) the direction of the arc might not always correspond to a meaningful relationship, at least not for the purpose of detecting word similarity (e.g. participles modifying other verbs); (c) some syntactic contrasts might in some cases be rather arbitrary (e.g. "adverbial" vs. "object"); (d) differences in syntactic structure do not always have a one-to-one correspondence to meaning differences (e.g. the object of an active construction and the subject of a passive construction both correspond to the patient or theme of the same verb); and (e) using syntactic and morphological features could introduce some high-level information about the syntactic usage of a word (e.g. the complementation patterns in which it typically takes part) which might not in all cases be optimal to detect word similarity. As a result, adding a too large amount of linguistic analysis could lead to data sparsity by dividing features in several sub-features of which the contrasts between them are not that significant. This is not to say that using a higher level of linguistic analysis is entirely detrimental: as there are no big quantitative differences between the different dependency models, it is rather the case that the benefits and the drawbacks of an increasing level of analysis outweigh each other. Therefore in the future it would be worthwhile to take a closer look at the different levels of granularity of specific labels and decide in which cases it would be beneficial for the detection of semantic similarity to make more fine-grained distinctions and in which cases it would not. Another, more automated way to reduce such "artificial" differences is to use a dimension reduction technique such as SVD, by including labels of various levels of granularity together in the PPMI matrix and letting the dimension reduction detect the most relevant distinctions.

Evaluating our results against independent benchmarks, we found that the difference between bag-of-words and syntactic models was less pronounced there, likely because these benchmarks contain several topically related words for which the syntactic models would reduce the strength of the association.

There are several ways to expand on this current work. First of all, we have shown that a wide mix of context features, i.e. bag-of-words context features, dependencies, syntactic relations and inflectional morphological features, all encode useful information for distributional semantic modelling. We could also add derivational morphological features to this list, which has already been noticed by Boschetti

(2010), but which we did not consider here due to a lack of derivational morphological annotation in the corpora we used. While we created a separate model for each of these categories of features, it would be useful to integrate the strengths of each of them in a single model, as detailed above.

Secondly, while this paper was specifically concerned with type-level distributional models, it would be useful to apply these insights to token-level models as well. Detecting word similarity on the type level ignores the fact that some words may be highly similar with respect to one meaning but highly dissimilar with respect to another meaning. Additionally, this study exclusively made use of a context-count architecture, which has been shown to perform inferiorly in comparison with context-predict architectures: therefore it will be useful to compare results with the latter models as well, both on the type level (e.g. *word2vec*, see also Stopponi et al. 2023) and on the token level (e.g. *RoBERTa*, see also Riemenschneider and Frank 2023).

Finally, we have shown that the lack of homogeneity of the Greek corpus with regard to genre is an important open problem – probably even more important than diachrony, seeing that many late literary writers wrote in a style similar to Classical Attic Greek. For many words the meaning is highly dependent on and/or predictable by the type of text in which they are used, and therefore their vectors can be skewed toward the meaning in some genres that are overrepresented in the corpus. In other words, this problem is highly related to the polysemy problem, and token-based models may therefore also be used to identify such genre-specific meanings. What is more, some text types provide more useful context features than others, e.g. highly descriptive scientific texts vs. formulaic texts such as contracts. As a result, even using more in-domain data might be detrimental if these data are less useful from a practical point of view (e.g. repetitive contexts). While this paper involved a very general task, in the future it will be necessary to take a closer look at the genre composition of the corpus from which the vectors are created, and filter out texts that are less suited for the task on hand or reduce their influence in some other way (e.g. by weighting them).

Acknowledgements

This paper is a thoroughly revised and updated version of a part of the PhD research of the first author (*A Computational Approach to the Greek Papyri*), supervised by the second author. We want to thank Toon Van Hal for his constructive feedback and his extensive work in annotating the dataset used in this study. We are also very grateful to the first reviewer for their constructive feedback, which has greatly improved the quality of this paper. Finally, we would like to thank the team of Martina Rodda, Philomen Probert and Barbara McGillivray, as well as Silvia Stopponi, for the highly valuable resources they created to evaluate distributional models which were also used in this study.

Supplementary material

All the material produced by this research can be found on https://github.com/alekkeersmaekers/greek-count-vectors

References

Baroni, M., Dinu, G., Kruszewski, G. (2014). Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247. Baltimore.

Bizzoni, Y., Boschetti, F., Diakoff H., Del Gratta, R., Monachini, M., Crane, G. R. (2014). The Making of Ancient Greek WordNet. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1140–1147. Reykjavik.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Boschetti, F. (2010). *A Corpus-Based Approach to Philological Issues*. University of Trento. (PhD thesis)

Botha, J., Blunsom, P. (2014). Compositional Morphology for Word Representations and Language Modelling. In: *International Conference on Machine Learning*, pp. 1899–1907. Beijing.

Bullinaria, J. A., Levy, J. P. (2007). Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39(3), pp. 510–526.

Bullinaria, J. A., Levy, J. P. (2012). Extracting Semantic Representations from Word Co-Occurrence Statistics: Stop-Lists, Stemming, and SVD. *Behavior Research Methods*, 44 (3), pp. 890–907.

Burns, P. J. (2019). Building a Text Analysis Pipeline for Classical Languages. In: Berti, M. (Ed.). *Digital Classical Philology*, pp. 159–176. Berlin: De Gruyter Saur.

Croft, W. (2013). Radical Construction Grammar. In: Hoffmann, T., Trousdale, G. (Eds.). *The Oxford Handbook of Construction Grammar*, pp. 211–232. Oxford: Oxford University Press.

De Pascale, S. (2019). *Token-Based Vector Space Models as Semantic Control in Lexical Lectometry.* University of Leuven. (PhD thesis)

Dik, H. (1995). *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus.* Amsterdam: Gieben.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10), pp. 635–653.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. In: *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Blackwell.

Heylen, K., Peirsman, Y., Geeraerts, D., Speelman, D. (2008). Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pp. 3243–3249. Marrakech.

Jones, H. S., Liddell, H.G., MacKenzie, R., Scott, R., Thompson, A. A. (1996). *A Greek-English Lexicon*. Oxford: Clarendon.

Keersmaekers, A. (2020). The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In: *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pp. 39–50. Online.

Kolb, P. (2009). Experiments on the Difference between Semantic Similarity and Relatedness. In: *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pp. 81–88. Odense.

Lapesa, G., Evert, S. (2014). A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics,* 2, pp. 531–546.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics,* 4, pp. 151–171.

Levy, O., Goldberg, Y. (2014). Dependency-Based Word Embeddings. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* pp. 302–308. Baltimore.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2,* pp. 768–774. Montreal.

List, N. (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), pp. 137–152.

Luong, T., Socher, R. Manning, C. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113. Sofia.

McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., Vatri, A. (2019). A Computational Approach to Lexical Polysemy in Ancient Greek. *Digital Scholarship in the Humanities,* 34(4): pp. 893–907.

Mercelis, W., Van Hal, T., Keersmaekers, A. (Forthcoming). Tongue, Language or Noise? Word Sense Disambiguation in Ancient Greek with Corpus-Based Methods. In: *International Colloquium of Ancient Greek Linguistics*. Madrid.

Padó, S., Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2): pp. 161–199.

Peirsman, Y., Heylen, K., Geeraerts, D. (2008). Size Matters: Tight and Loose Context Definitions in English Word Space Models. In: *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pp. 34–41. Hamburg.

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., McGillivray, B. (2021). Lexical semantic change for Ancient Greek and Latin. In: Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., Hengchen, S. (Eds.). *Computational approaches to semantic change*, pp. 287–310. Berlin: Language Science Press.

Riemenschneider, F., Frank, A. (2023). Exploring Large Language Models for Classical Philology. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15181–15199. Toronto.

Rodda, M. A., Senaldi, M. S. G., Lenci, A. (2017). Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1): pp. 11–24.

Rodda, M. A., Probert, P., McGillivray, B. (2019). Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique des Langues*, 60p.(3), pp. 63-87.

Schmidt, J. H. H. (1876–1886). *Synonymik Der Griechischen Sprache*. Leipzig: Teubner.

Singh, P., Rutten, G., Lefever, E. (2021). A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek. In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 128–137. Punta Cana (online).

Spanopoulos, A. I. (2022). *Language Models for Ancient Greek.* National and Kapodistrian University of Athens. (BA thesis)

Stopponi, S., Pedrazzini, N. Peels, S., McGillivray, B., Nissim, M. (2023). Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work. In: *Proceedings of the 1st International Workshop on Ancient Language Processing (ALP) at RANLP*. Varna.

Turney, P. D., Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37: pp. 141–188.

Vatri, A., McGillivray, B. (2018). The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 3(1): pp. 55–65.

Yamshchikov, I., Tikhonov, A. Pantis, Y. Schubert, C., Jost, J. (2022). BERT in Plutarch's Shadows. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6071–6080. Abu Dhabi.