Computational Linguistics in the Netherlands Journal 14 (2025) 421-450 Submitted 12/2024; Published 07/2025

Lexical semantic change detection for Ancient Greek: dataset creation and evaluation of a word-embedding-based technique

Silvia Stopponi* S.STOPPONI@RUG.NL
Malvina Nissim* M.NISSIM@RUG.NL
Saskia Peels-Matthey* S.PEELS@RUG.NL

*Center for Language and Cognition, University of Groningen
Postbus 716, 9700 AS Groningen, The Netherlands

Abstract

We create a benchmark for the evaluation of lexical semantic change detection in Ancient Greek and use it to assess the validity of two metrics of lexical semantic change on diachronic embeddings models. Stopponi et al. (2024b) assessed the viability of lexical semantic change detection for Ancient Greek with word2vec models, using two existing measures. However, only a manual evaluation was conducted since a benchmark for the evaluation of this task for Ancient Greek was still missing. We create such a benchmark by extracting cases of semantic change from close-reading studies in Ancient Greek lexical semantics. We also create a parallel benchmark of semantically stable items and assess the effectiveness of the most relevant of the two metrics in distinguishing semantically changed from semantically stable items. Finally, we qualitatively evaluate the candidates for semantic change detected by filtering words by low vector coherence value and high frequency. The results show that the method is effective at retrieving cases of semantic change, especially when coupled with frequency information, but also reinforce the idea that performing lexical semantic change detection on an ancient language and building a robust evaluation benchmark are particularly challenging tasks. In conclusion, we propose a constructive way to leverage this method as a research companion, by integrating it with the close-reading method.

1. Introduction

Lexical semantic change is a relevant phenomenon from both a linguistic and an cultural point of view, given that the changing usage of words could signal a cultural change. Computational methods for semantic change detection have been mainly applied to modern languages. An overview is in Tahmasebi et al. (2021a), leveraging both large corpora and native speakers to train and evaluate large models. This is not the case for ancient languages, such as Ancient Greek, on which we focus. Ancient languages lack native speakers, and knowledge about these languages typically relies on a written corpus, limited in size, which cannot be substantially increased.

Most existing studies on lexical semantic change in Ancient Greek were carried out with the close-reading (or 'philological') method, allowing for a detailed analysis of word occurrences in context. However, due to the time-consuming nature of the method, no philological study in lexical semantics can include the whole corpus of Ancient Greek. Few studies tried to address lexical semantic change with computational methods: Boschetti (2009), Boschetti (2018), Rodda et al. (2017), and Stopponi et al. (2024b). The last one, in particular, assessed the viability of lexical semantic change (LSC) detection for Ancient Greek with word embedding models.[1] Two measures of semantic change were tested: vector coherence (VC) and J, following the method in Cassani et al. (2021), who modelled lexical semantic change in a diachronic corpus of English. VC quantifies the coherence of word embeddings of the same word extracted from different corpus slices (and thus different time

1. A summary of the existing work on LSC for Ancient Greek is in Stopponi et al. (2024b, 48).

periods), while J measures the change in the nearest neighbours of the same word across different corpus slices. Stopponi et al. (2024b) concluded that both measures are effective at retrieving usage stability in Ancient Greek, but that VC is more useful than J to detect usage change, a signal of potential semantic change.[2] J would indeed tend to "equate diachronic coherence to closed semantic domains" (Cassani et al. 2021, 17), i.e. to be higher for words which keep the same nearest neighbours through time, as it is obvious for a measure based on the overlap between nearest neighbours through time. An example of this is the high J obtained by numerals and ordinals, as noticed by Cassani et al. (2021, 17) for English and by Stopponi et al. (2024b, 53) for Ancient Greek. Indeed, numerals and ordinals exhibit not only usage stability but also consistency in their nearest neighbours (other numerals and ordinals) over time.

No evaluation of the measures against an external benchmark was performed by Cassani et al. (2021), but only a manual inspection of the words with the lowest and highest values. Stopponi et al. (2024b) did the same, by manually inspecting the 50 lemmas[3] with the highest and lowest VC and J, without evaluating the adopted change metrics against an external benchmark, since a resource for the evaluation of LSC detection for Ancient Greek did not exist yet. One aim of the present research is to address that need, by releasing a benchmark of Ancient Greek lemmas which underwent attested semantic change, as reported in the scholarship. The resource is subsequently used to assess the effectiveness of the method (word embeddings $+VC$ measure) at detecting those changes. Given that the reliability of the J measure had previously been questioned, we only briefly discuss the results of a preliminary inspection with this metric (reported in Section 5 and Figure 1), without any extended analysis.

2. Related Research

The development and progress of automatic semantic change detection follows closely the development and progress of meaning representation and processing in language technology (see Tahmasebi et al. (2021b) for an overview of the early approaches). At first, count-based (co-occurrence based) methods were tested, such as Latent Semantic Analysis (Sagi et al. 2009, Sagi et al. 2011), Positive Pointwise Mutual Information weighting of the matrices (as in Hamilton et al. (2016b) and Rodda et al. (2017)), or the Temporal Random Indexing (Basile et al. 2014, Caputo et al. 2015). Then, most work on LSC detection moved to using word embeddings, e.g., (Kulkarni et al. 2015, Hamilton et al. 2016b, Hamilton et al. 2016a). To allow for comparison between vectors from different spaces, various strategies have been developed, for example, the method by Kim et al. (2014a) and Haagsma and Nissim (2017), the Temporal Referencing by Dubossarsky et al. (2019), TWEC by Di Carlo et al. (2019), and CADE by Bianchi et al. (2020)). Most recently, contextual embeddings, such as ELMo's (Peters et al. 2018) or BERT's (Devlin et al. 2019), have been used, including many of the systems participating in the SemEval 2020 Task 1 (Schlechtweg et al. (2020, 19–22); Kutuzov and Giulianelli (2020)), and also large language models (GPT-4 in Wang and Choi (2023)), but the usability of such models for ancient languages is hampered by their low-resource nature (Spanopoulos 2022).

More in general, ancient languages, including Ancient Greek, remained at the margins of the scholarship in LSC detection. In the specific case of Ancient Greek, most existing studies about lexical semantic change are carried out with the so-called close-reading method. A scholar adopting this method carefully reads a selection of texts, typically searching for specific elements, for example thematic, linguistic, or stylistic features. This method does not exclude integration with a quantitative approach, since the research does not necessarily need to be confined to a limited number of passages (selected with the aid of an online database, notably the Thesaurus Linguae

2. Usage change does not necessarily imply semantic change. However, the opposite holds when adopting a distributional view of meaning: if meaning is represented through usage, then a change in meaning can only be detected through usage change. In this paper, when we talk about 'usage change', we refer to possible candidates for semantic change.
3. They use the lemmatised version of the training corpus.

Graecae–TLG). The close-reading method sometimes entails the systematic reading of whole works, collections of works, or even the whole production of an author, and the recording of all occurrences of the target phenomenon. One may subsequently conduct a quantitative investigation on such a dataset. An advantage of the close-reading method is that the scholar is exposed to the text and can thus discern better than a computational system which passages are actually relevant towards the research aims, while discarding noise (e.g. passages of features which just resemble the target, but are not relevant). During exposure to the text, the scholar can discover unnoticed or unknown factors influencing the analysis, or recalibrate the research questions and dataset. However, the close-reading method also brings disadvantages with it, compared to computational linguistic methods. The time needed for reading is the most important constraint, making it impossible for a single scholar to investigate a large amount of data. For example, a complete close-reading analysis is impossible even on a relatively small corpus such as the Diorisis Corpus. Time constraints also limit the range of phenomena that can be investigated at one time. In addition, the close-reading method is strictly tied to the personal knowledge, preferences, perhaps biases of the scholar involved, and possibly to varying levels of consistency throughout a prolonged study.

Some computational work on lexical semantic change for ancient languages does exist. For example, on Latin (Bamman and Crane 2011, Eger and Mehler 2016, Perrone et al. 2021) and Ancient Greek (Boschetti 2009, Rodda et al. 2017, Perrone et al. 2019, McGillivray et al. 2019, Perrone et al. 2021, Stopponi et al. 2024b).

3. Data and Method

3.1 The case of Ancient Greek

Ancient Greek was spoken for more than two thousands years in an area around the Mediterranean. Thousands of speakers used it in all activities of daily life, and a considerable amount of documents of different kinds were written in this language. Yet only a scarce amount of them has come down to us, surviving centuries of historical events. The extant Ancient Greek corpus includes well-known literary works written by more than two thousands authors, covering a range of genres, such as epic poetry (e.g. Homer), drama (e.g. Sophocles and Aristophanes), and philosophy (e.g. Plato and Aristotle). These are just the tip of the iceberg of a vast literary production which went mostly lost.[4] Ancient Greek is a low-resource language since its corpus is limited in size and, contrary to the situation for some modern low-resource languages, its corpus cannot be substantially increased due to the lack of native speakers. Moreover, the extant Ancient Greek corpus covers a long timespan (e.g. the corpus adopted in this study, the Diorisis Ancient Greek Corpus by Vatri and McGillivray (2018), spans through more than a thousand years, see Table 1), and it is not balanced for genre. This means that, for example, the Archaic portion of the literary corpus (ca. 800–500 BCE) mainly consists of epic poetry, while in Classical times (ca. 500–323 BCE) we find for example drama, philosophy, (forensic) oratory, and historiography. Later, in Hellenistic and Roman times, we find a continuation of the same genres (epic, drama, philosophy, etc.), but also the *Septuagint* (the Greek Old Testament), and the *New Testament*. The imbalance in genres through time entails an imbalance in topics and style, which affects semantic change detection. Many words are not present in all time slices, making it sometimes impossible to track the evolution of their meaning throughout the whole corpus. Another difference between high- and low-resource languages is that frequency filtering cannot be too aggressive, due to the high amount of low-frequency words. For example, Gonen et al. (2020) discarded the 20% least frequent words in each of the corpora of English, French and Hebrew

4. For this analysis, we do not use other existing large corpora of Ancient Greek texts, namely epigraphic data (inscriptions, preserved on stone and other durable material) and papyri, but only literary texts.

they used, and all words occurring less than 200 times. In this study, we only discarded from each corpus slice lemmas[5] occurring less than 5 times and did not perform any stop-word filtering.

The Diorisis corpus only contains literary data, texts which typically have been copied for centuries from manuscript to manuscript. For this kind of texts there is a larger availability of annotated corpora, processed for NLP purposes. However, literature is not the only type of written evidence for Ancient Greek. Inscriptions and papyri are the other two relevant kinds of written documents. Typically they are original products from Ancient Greece, not copies of lost 'originals'. Inscriptions are often found on stone, but they can also be inscribed on pottery, for example (pieces of) vases, or other objects and durable materials. Papyri are mostly non-literary texts, generally written on papyrus. The so-called 'documentary papyri' typically contain contracts, letters, petitions, and similar, and they mostly come from Egypt, where they were preserved by the specific climatic conditions of the desert. Digitised corpora of inscriptions are often partially annotated or not annotated at all, as in the case of the corpus provided by the Packard Humanities Institute,[6] and different collections follow different annotation standards, for example for lemmatisation, so that it is difficult to merge them into a larger corpus with a coherent annotation scheme. One reason for discrepancies in lemmatisation between collections, especially when they include early inscriptions, are the different varieties of Greek used, according to the temporal and geographic provenance of the texts. It can thus happen that the same words are assigned different lemmas in different collections, depending on the specific dialect at hand. This does not happen in the annotation of the Diorisis corpus, which was lemmatised all together, with the same automatic tool. The language of papyri, which mostly come from Egypt, is more uniform and these texts are more homogeneous in subject matter. Even if there is at least one corpus of papyri annotated for NLP purposes (released by Keersmaekers (2020)), the current work only focuses on literary texts, and does not include inscriptions or papyri.

3.2 Training data

The diachronic corpus of Ancient Greek adopted in this work is the Diorisis Ancient Greek Corpus (Vatri and McGillivray 2018). The corpus contains 820 texts spanning from the first extant Greek literature to the 6th century CE, for a total of 10,206,421 tokens, which were automatically lemmatised and POS-tagged. Due to the small size of the corpus, we used lemmas instead of wordforms to reduce data sparsity. The sparsity issue is particularly relevant because to train diachronic word embedding models the corpus must be divided into slices, five in our case, making the problem even bigger. The slices roughly reflect the traditional divisions into periods of the Greek literature, but not completely, to limit unbalance in size. For example, our 'Hellenistic' slice ends with the year 0, later than in the traditional division. All words occurring less than 5 times in a slice were filtered out from it. An overview of the size of the slices after frequency filtering is in Table 1. A certain imbalance in slice size is inevitable when also aiming at a balance in time span since the extant Ancient Greek texts are not evenly distributed though time. The Diorisis Corpus is thus poorer of texts belonging to the earliest and latest phases of the Ancient Greek literature, and this is reflected by the smaller size of the Archaic and Late Roman slice.

To give an example of the difference in available data and covered timespan between ancient and modern languages, the Corpus of Historical American English by Davies (2012) (COHA) used by Cassani et al. (2021) only covers the timespan between 1810 and 2009, yet includes more than 400 million words (making it 40 times larger than our corpus). It is balanced by genre per decade. Its six slices are thus larger and more homogeneous than our five slices of the Diorisis corpus, which is not only smaller, but also imbalanced by genre over time.

5. The version of the Diorisis corpus e use is the lemmatised one, as it commonly happens in computational studies of Ancient Greek.
6. `https://inscriptions.packhum.org/`

Time slice	N. tokens	Vocab. size	Timespan
Archaic	229,999	3,829	beginning–500 BCE
Classical	2,628,193	14,526	499–324 BCE
Hellenistic	2,164,057	12,698	323–0 BCE
Early Roman	4,276,672	19,652	1–250 CE
Late Roman	753,907	8,578	251–500 CE

Table 1: Number of tokens, vocabulary size, and timespan per slice.

3.3 Word Embedding Models

As in Stopponi et al. (2024b) and Cassani et al. (2021), the framework adopted to train word embeddings is CADE (Bianchi et al. 2020), a technique to train word2vec models on different corpus slices without the need for space alignment since all models are initialised in the same way. We trained our models with a context window of five, i.e. five words to the left and five to the right of each target word were taken into account during training. We specified the following parameters: (vector) size = 30, siter = 5, diter = 5, workers = 4, sg = 0, ns = 20.[7]

According to previous research, such as Antoniak and Mimno (2018) and Wendlandt et al. (2018), word embeddings suffer from instability, i.e. training the same model in different conditions–for example with different initialisation, corpus order, and corpus composition, as in Antoniak and Mimno (2018)–produces different vectors and thus different semantic relationships between words. Wendlandt et al. (2018) only assess nearest neighbours stability, while Antoniak and Mimno (2018) explore stability as variation across different trainings in: a. cosine similarity between pairs of words; b. overlap between nearest neighbours lists (Jaccard similarity). In this study, however, we do not rely on overlap in nearest neighbours to quantitatively evaluate the proposed method; instead, we only use them in qualitative evaluation to get cues about the direction of semantic change of specific words. We are aware that the lists of nearest neighbours of the same word can significantly vary among different trainings of the same model, but we have empirical evidence that different lists of neighbours just point to the same meanings in different ways, and that no 'wrong', but just different, associations are obtained when repeating model training with different initialisations.[8] Moreover, being aware of the low stability of the ranking of neighbours, we never take into account the order in which the nearest neighbours appear, nor do we advise drawing any conclusions from it.

3.4 Measures of change

Regarding the measures of change, Cassani et al. (2021) used VC, J, and LNC,[9] while we only extensively test the statistical significance of the results obtained with VC, and just briefly present the results of a preliminary investigation with J. The latter measure proved to be less reliable at retrieving semantic change (Stopponi et al. 2024b), and we thus discarded it from this study. Cassani et al. (2021) do not report any evaluation of the metrics' performance against an external resource, but they did a manual evaluation of some detected words with high and low diachronic coherence. They also assessed a strong positive correlation between the three measures, "suggesting that they largely capture similar patterns" (Cassani et al. 2021, 11), and observed that words with low VC

7. Explanation and more information about the parameters of CADE is in its code: `https://github.com/vinid/cade/blob/master/cade/cade.py`.
8. This can also be observed in Table 5 and Table 6 in Antoniak and Mimno (2018), where different neighbour lists do not contain 'wrong' associations with the target word; they simply point to different aspects of its meaning. Note that those lists are obtained in the 'bootstrap' setting, i.e. the different trainings were not executed on exactly the same corpus.
9. Local neighbourhood coherence (LNC) is a third measure adopted by Cassani et al. (2021) but not by us, for the reasons explained in Stopponi et al. (2024b).

also obtained low J and LNC. While we also observe a positive correlation between VC and J in this work, as we discuss in Section 5, this correlation does not translate into equal reliability for both metrics when used for the actual evaluation of lexical semantic change.

The measures adopted by Cassani et al. (2021) are adapted from previous studies. In particular, VC is an adaptation of an often used metric to quantify usage change, for example in Gulordava and Baroni (2011), Jatowt and Duh (2014), Kim et al. (2014b), Hamilton et al. (2016c), and Hamilton et al. (2016a), who call it 'global measure'. This 'global measure' consists in measuring the cosine distance between vectors of the same word between consecutive corpus slices. This idea was reused by Cassani et al. (2021) for VC, with the addition of an extra step, namely calculating the sum of all cosine similarities per word, computed between all combinations of its slice-specific vectors, to obtain a single value per item. Quantitative evaluation only exists for the original metric, and Cassani et al. (2021) do not provide any quantitative evaluation of the accuracy of VC. A quantitative evaluation of the original metric (the simple cosine similarity without sum) is in Gulordava and Baroni (2011), who report correlation between their similarity-based metric and human judgements they collected.[10] But word2vec did not exist yet at the time of publication of their study, which is why they evaluate the cosine similarity measure in combination with count-based vectors, not with predictive word embeddings. Qualitative evaluation of the metric, still coupled with count-based word vectors, is in Jatowt and Duh (2014), who selected case-studies from the *Oxford Dictionary of Word Origins* (Cresswell 2010) and the *Online Etymological Dictionary* (Harper n.d.).

The lack of an evaluation of the metrics against an external benchmark is probably due to the entirely different aim of Cassani et al. (2021), compared to ours. Their purpose is not to evaluate the metrics, but to investigate the relationship between age of acquisition of words and semantic change, i.e. whether words that are acquired later in life are more susceptible to shifts in meaning. Another difference between our study and Cassani et al. (2021) is the fact that they do not report the values of the metrics assigned to the whole lexicon, whereas we provide the average VC for all lexicon items, benchmark and non-benchmark.

VC and J are calculated by following two different approaches: VC leverages word vectors to calculate cosine similarities between vectors of the same word in different slices, while J leverages vectors to extract the nearest neighbours of each word in different slices and calculate the overlap between neighbour lists. The VC of a word is a sum of cosine similarities. More specifically, it is the sum of the cosine similarities between a word's embeddings in different corpus slices. The maximum VC value depends on the number of slice combinations; as we have ten possible combinations between five time slices and we use all of them, the VC thus varies between -10 and 10. A word would receive $VC = 10$ in the extreme case in which the cosine similarity was 1 in all ten slice combinations.

J is a sum of Jaccard coefficients, instead. The Jaccard coefficient of a word, calculated between two time slices, is the intersection of the two slice-specific lists of top k nearest neighbours of the word–$k = 10$ in this study–divided by the union of the two lists (thus without duplicates). It ranges between 0 and 1. The J of a word is the sum of Jaccard coefficients calculated between combinations of slices, in our case all ten possible combinations between the five corpus slices. This 'global' J score ranges in our case between 0 and 10, depending the maximum value on the number of slices. A word from our corpus would thus get $J = 10$ if there is a perfect overlap between the two lists of nearest neighbours for all combinations of slices.

In this study, the VC was initially only calculated for the words shared among all corpus slices, i.e. if a word was not present in one or more slices it was excluded from the analysis. But since the small size of the first and last slice yields a strong limitation on the size of the shared vocabulary, we then decided to compute the measure in two different settings: by taking all the five slices of the corpus into account ('five slices' setting) and by only taking into account the three central ones, larger and more homogeneous in genres ('three slices' setting). Restricting the corpus to three slices has the disadvantage of limiting the time span on which lexical semantic change detection is performed,

10. Differently from the other cited studies, Gulordava and Baroni (2011) divided their corpus into two slices only, so that they could compare for each word one cosine similarity value and one human judgement.

but the advantage of increasing the lexicon shared among all slices from 2,030 to 8,367 words. Since the aim of this paper is to investigate longer-term semantic change, we did not calculate VC or J of words between two consecutive slices only.

4. Benchmark creation

As a first step towards building a resource to evaluate LSC detection, we identified semantic fields corresponding to historical themes or developments which were potential focal points of change. For example, words related to the sociopolitical organisation are likely to have changed their meaning during the rise of the Hellenistic monarchies (after the death of Alexander the Great in 323 BCE) and after the Roman conquest of Greece (2nd–1st cent. BCE). Another potential trigger of semantic change is the transition from the prevalence of polytheism to the prevalence of Christianity. However, we had to deal with the scarcity of philological (non-computational) scholarship treating specific cases of semantic change in Ancient Greek. In particular, studies including clear indications of the meanings involved in the change are rare (e.g. word x changed from having meaning y to having meaning z, or added the new meaning z to the older meaning y). Some relevant studies include Gingrich (1954), Buck (1949), Finkelberg (1998), Horky (2019), and Luraghi (2022). Contrary to Hamilton et al. (2016c), who could use the *Oxford English Dictionary* to retrieve attested cases of semantic change in English,[11] for example by searching for the 'obsolete' tags, or from Jatowt and Duh (2014), who selected cases of changed words from the *Oxford Dictionary of Word Origins* (Cresswell 2010) and the *Online Etymological Dictionary* (Harper n.d.), we could not use dictionaries of Ancient Greek in this way, since semantic change is not tagged there, but only indirectly indicated, and not even systematically.

The items extracted from the literature were first double-checked with four dictionaries: Liddell et al. (1940) (Ancient Greek - English), Rocci (1939) (Ancient Greek - Italian), and Sluiter et al. (2024) (Ancient Greek - Dutch). If the meanings involved in a change were not recorded in at least one of the dictionaries, the word was not included in the benchmark. All items were manually checked to ensure they represented clear-cut changes, which were likely to involve a sensible modification in the contexts of occurrence of the word, and thus to be detected with a word-embedding-based method. Some examples of the decisions made are in Table 2. More examples of included and excluded items are in Appendix A.

The final selection includes 44 items, mostly nouns and verbs (only two are adjectives). Most changes are originated in the *Septuagint* (the Greek Old Testament), in the New Testament, or, more in general, in Hellenistic times.[12] We would like to point out that 42 out of 44 items were extracted from Gingrich (1954), *The Greek New Testament as a Landmark in the Course of Semantic Change*, sometimes reinforced by Buck (1949). It is thus not surprising that 15 items (34%)[13] are changes triggered by the rise of Christianity. The high proportion of items related to this topic is, however, inevitable, due to the mentioned difficulty in finding examples in the scholarship. The released benchmark file[14] includes for each lemma information about the reference(s) in the scholarship, the meanings involved in the change, the time of the change, and whether or not it is related to Christianity. The full list of benchmark items is also provided in Appendix B.

11. The gold data were used by the authors to evaluate the performance of different word embedding algorithms at detecting known shifts.
12. The Hellenistic period of Greek history conventionally goes from the death of Alexander the Great (323 BCE) to the Roman conquest of the Ptolemaic kingdom (30 BCE). Note that our 'Hellenistic' slice includes texts up to the year 0, due to the need to balance the slices in size.
13. This percentage refers to the composition of the whole benchmark. However, not all benchmark items could be used in this study, since some of them were absent from the shared lexicon among all corpus slices. The percentage of Christianity-driven changed items which could be used for evaluation is 35% in the 'three slices' setting and 25% in the 'five slices' setting.
14. `https://zenodo.org/records/13364555`

Word	Meaning change	In benchmark	Reason
δαιμόνιον, *daimonion*	'divinity' > 'evil spirit', 'demon'	yes	clear-cut change
κοινωνία, *koinōnia*	'communion', 'association' > 'contribution'	yes	clear-cut change
ἐντυγχάνω, *entynchanō*	'meet with' > 'read'	yes	clear-cut change
λαμπάς, *lampas*	'torch' > 'oil lamp'	no	too subtle
πτωχός, *ptōchos*	'beggar' > 'poor man'	no	too subtle
ὑπάγω, *hypagō*	'retire' > 'go', 'go away'	no	too polysemous

Table 2: Examples of included and excluded items. By 'clear-cut changes' we mean that the change at hand was double-checked with the dictionaries and deemed sufficiently evident to be detected with word embedding models and VC. We excluded cases such as λαμπάς, *lampas* and πτωχός, *ptōchos* because the change in meaning was very subtle and probably did not imply a detectable change in the contexts of occurrence of the word. By 'too polysemous' we mean that the word ὑπάγω, *hypagō* has an extremely wide range of meanings–16 in the dictionary by Liddell et al. (1940). This could make semantic change detection harder because the meanings involved in the change risk to end up representing just a small portion of the usages, mixed with many other irrelevant usages.

5. Evaluation

Contrary to Stopponi et al. (2024b), in this study we focus solely on VC, due to the low reliability of the J measure. The latter has the tendency to flatten most values towards zero, due to the fact that overlap between nearest neighbours to the same word in different time slices occurs quite rarely. This tendency can be seen in Figure 1. It should be noticed that this characteristic of J does not exclude correlation with VC. We assessed a positive correlation between VC and J, both in the 'three slices' and in the 'five slices' setting.[15] Such a correlation is also reported by Cassani et al. (2021). These results show that a positive correlation between metrics does not ensure reliability.

The first step of our evaluation consisted in assessing the average VC assigned to the benchmark (changed) words and whether the difference in average VC between benchmark and non-benchmark items was statistically significant. We expected the VC of benchmark items to be significantly lower than that of other items. In a subsequent step, we assessed whether word frequency had a significant effect on VC. Not all benchmark items could be used for evaluation in this study, since we only evaluated against items which were present in the shared lexicon between all corpus slices of a specific setting (either three or five slices). In the 'five slices' setting only 11 benchmark items were present in the shared lexicon and could thus be used for evaluation, while in the 'three slices' setting, with a larger shared lexicon, 37 benchmark items could be used.

Benchmark items obtain on average a lower (normalised)[16] VC (0.72) than non-benchmark items (average $VC = 0.74$). There is also a difference between all items in the 'three slices' (average $VC = 0.80$) and in the 'five slices' setting (average $VC = 0.68$). Both differences were expected, as it is logical that words changed their meaning more on average over a longer time span. Additionally,

15. In particular, in the 'three slices' setting we found a moderate positive correlation between VC and J ($rho = 0.32$), while in the 'five slices' setting the positive correlation is weaker ($rho = 0.19$). Both correlations are highly significant ($p < 0.001$).

16. The VC depends on the number of slices, being a sum of cosine similarities. It thus ranges between -10 and 10 in the 5 slices setting (where ten combinations of slices are possible, and ten cosine similarity values are summed up) and between -3 and 3 in the three slices setting. To make the average VC values comparable among the two settings, the values of VC were normalised, by dividing them by the number of slice combinations, 10 in the 'five slices' setting and 3 in the 'three slices' setting.

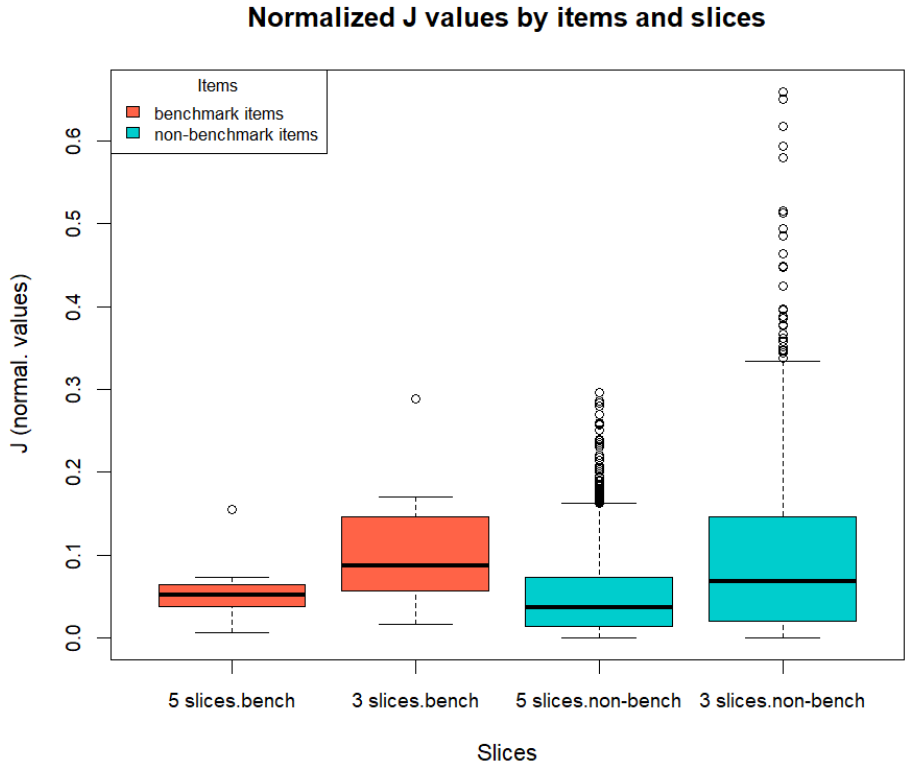


Figure 1: Boxplot representing the normalised J values divided by item type and number of slices, made with *R*.

we expect benchmark items to change on average more than non-benchmark items, which are a mix of words with varying degrees of usage stability. Finally, an important element to keep in mind when interpreting the statistical significance of the differences in VC is the imbalance in size between the two sets of benchmark and non-benchmark items, summarised in Table 3.

Setting	Type of items	N. lemmas
5 slices	Benchmark	11
	Non-benchmark	2019
	All lexicon	2030
3 slices	Benchmark	37
	Non-benchmark	8330
	All lexicon	8367

Table 3: Number of lemmas belonging to each set: benchmark, non-benchmark, and all lexicon.

5.1 Benchmark vs. non-benchmark items

The average VC values per item type and setting are summarised in Table 4. The values between parentheses, where there are any, are those valid for this analysis, calculated on a smaller shared

lexicon, as explained below, while the values outside parentheses are valid for the analysis in the following Section 5.2. A boxplot of the VC values divided by setting and type of items is in Figure 2.

Setting	**Type of items**		
	Benchmark	**Non-benchmark**	**Tot.**
3 slices	0.74	0.77 (0.80)	0.77 (0.80)
5 slices	0.64 (0.80)	0.68	0.68
Tot.	0.71 (0.72)	0.75 (0.74)	

Table 4: Average normalised VC per setting (three or five slices) and type of items (benchmark or non-benchmark). Lower VC means more potential change, higher VC means more usage stability, and thus more potential meaning stability. The values between parentheses belong to the first analysis, to test the significance of slices and lexicon. In that analysis only the smaller shared lexicon between all five slices could be used. When two values are available, the values outside parentheses belong to the second analysis, where the shared lexicon among three slices could be used because the dataset was split into two independent subsets.

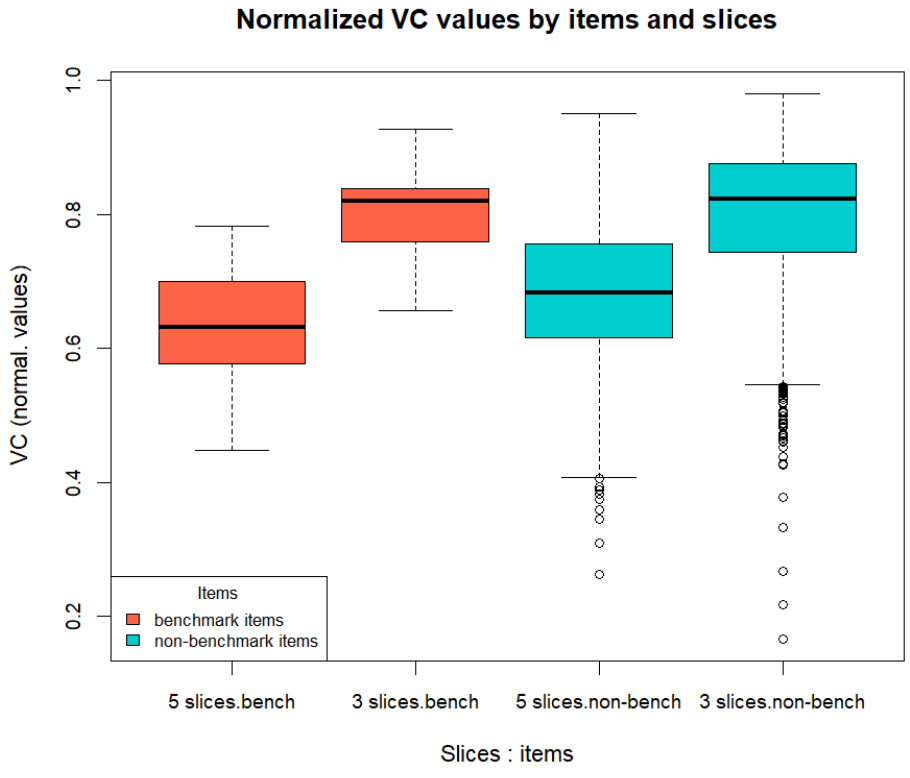


Figure 2: Boxplot representing the normalised VC values divided by item type and number of slices, made with *R*.

To assess the relationship between VC, item type (benchmark or non-benchmark), and setting (either three or five slices), we performed a linear mixed effects analysis.[17] Item type and number of slices were the fixed effects, while we set the lemma as random effect since two VC values were available for each lemma, one per setting. We indeed only included in this model the lemmas for which two VC values were available, i.e. those which were present in the shared lexicon of both settings. In practice, the subset of analysed lemmas coincides with the lexicon shared among all five slices, which only includes 11 benchmark items. This was done to avoid imbalance in the data set, with some benchmark items having two VC values and others (those not present in the shared lexicon of the five slices) only one. P-values were obtained by likelihood ratio tests of the model with a specific fixed effect to assess (item type or setting) against a model without that effect. The result was that item type does not have a significant effect on VC ($\chi^2(1) = 0.77, p = 0.38$), while the factor 'number of slices' does have en effect on it ($\chi^2(1) = 2019.5, p < 0.001$). In particular, the level 'slices = three' increases VC (normalised) by about 0.12 ($\pm$ 0.002 standard errors). This means that the difference in VC observed between the 'three slices' and the 'five slices' setting is highly significant, while the difference in VC between benchmark and non-benchmark items (with benchmark items obtaining on average a lower VC) could just be due to chance. However, the nonsignificance of 'items' could be explained by the the small set of 11 benchmark items available for this test and by the large imbalance in size between these and the 2019 non-benchmark items. Indeed, the difficulty of finding examples of changed words in the scholarship, combined with the manual selection of items to include in the final resource, strongly impacted the benchmark size in relation to the whole lexicon (see Table 3). Even if benchmark items are ascertained cases of semantic change, the size of the resource may be too limited for the difference in VC between benchmark and non-benchmark items to be significant. We thus decided to run an additional test by splitting the dataset into two parts, one per setting. In this way, for the 'three slices' setting we could use more than 11 evaluation items, since 37 benchmark lemmas are in the shared lexicon across the three slices.

5.2 Frequency and benchmark vs. non-benchmark items

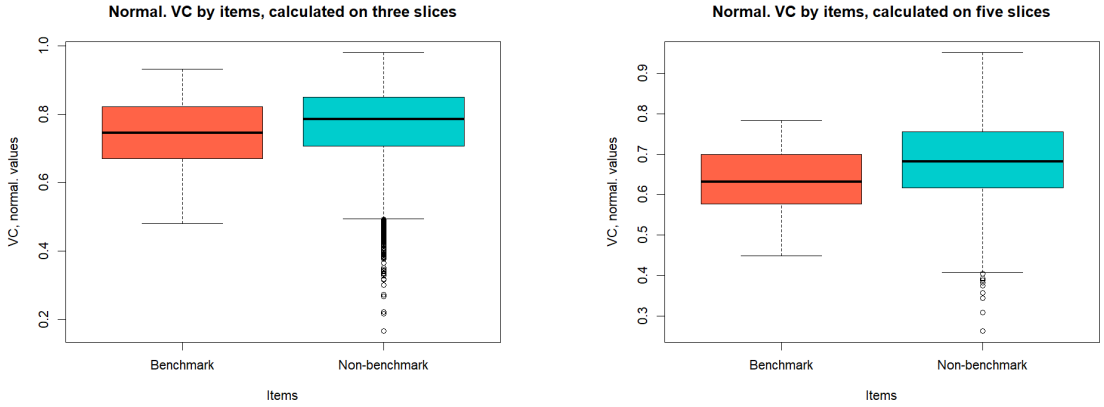
To explore the data further, we ran another test, which also included word frequency as a fixed effect, since we considered it as a relevant factor that could have impacted the pace of semantic change as assessed by the model. The relationship between frequency and language change has already been investigated, for example by Winter et al. (2014), Vejdemo and Hörberg (2016), and Hamilton et al. (2016c).[18] These scholars consider word frequency as a relevant factor affecting semantic change. In particular, higher frequency of occurrence would correlate with slower meaning change.[19] Moreover, we consider frequency as a factor which can influence not only a word's degree of change, but also its stability, since word embeddings for lower-frequency words are based on fewer contexts. Those contexts may be not be completely representative of the prevalent meaning of the word, and could bias the vector representation or be too noisy for the representation to be meaningful, conditioning the quality of semantic change detection. The impact of frequency on vector stability as already been observed by Wendlandt et al. (2018), who assessed that frequency has a non-negligible impact on the stability of word2vec embeddings (though it behaves as a minor factor). The decision of including frequency as a factor was backed up by some heteroscedasticity of residuals observed in the previous analysis, a potential signal of missing factors. We represent word frequency in our dataset as absolute word frequency, i.e. we use for each word the sum of its frequencies in all slices (either three or five, according to the setting). As an alternative to absolute frequency, to possibly better grasp the degree of frequency variability over time, we also experimented with representing word frequency as *the standard deviation from its average frequency* over three or five slices. Due to the collinearity of the two ways of representing frequency, as obviously absolute frequency and frequency variability

17. Using R (R Core Team 2021) and lme4 (Bates et al. 2015).
18. The limitations of some existing studies are discussed by Cassani et al. (2021, 3).
19. See the discussion in Cassani et al. (2021, 22–23).

correlate, they cannot be both included in the same model. We report here results obtained using absolute frequency, since we found neither an advantage nor any actual difference in the explanatory power of the models when using frequency variability instead.

Another factor which could have affected the degree of semantic change associated to each word is polysemy. Indeed, semantic change can involve an increase or decrease in the number of meanings, with a new meaning that is just added to the previous one(s), without the old one(s) being lost, or with a meaning getting lost. The change results in these cases in an increased or reduced polysemy.[20] However, we did not find any straightforward strategy to incorporate polysemy in the analysis, since there is no obvious way of distinguishing a discrete amount of senses for a word.

Since we already assessed that 'slices' had a significant effect on VC, we decided to separately test the significance of 'frequency' for 'slices = three' and 'slices = five'. The factor 'items' was kept, to assess whether it had any significance in the two settings separately, especially when more words were taken into account in the 'three slices' setting. An advantage of splitting the dataset was indeed that more benchmark and non-benchmark items could be used in this setting, with a shared lexicon of 8,367 words, including 37 benchmark items, while the size of the lexicon stayed the same for the 'five slices' setting (2,030 words, including 11 benchmark items). During the previous test the number of usable benchmark and non-benchmark items had been indeed limited by the need to have two observations per word. As evident from Table 4, benchmark items obtain in both settings an average lower VC than non-benchmark items. In particular, in the 'three slices' setting the average normalised VC of benchmark items is 0.74 against 0.77 for non-benchmark items. In the 'five slices' setting the average normalised VC is 0.64 for benchmark and 0.68 for non-benchmark items.



(a) normalised VC values by items, calculated on three slices. The benchmark items are 37.

(b) normalised VC values by items, calculated on five slices. The benchmark items are 11.

Figure 3: Boxplots representing the normalised VC values by items, calculated over three or five slices.

Figure 3 shows these differences by setting. Whether they are significant (represented by the factor 'items') and whether 'frequency' has a significant effect on VC was tested in both settings with a multiple linear regression model. The outcome was that the effect of 'items' on VC is nonsignificant again, in both settings, while the positive effect of 'frequency' is always significant. Moreover, the two setting-specific models, even if significant overall, only explain a very small amount of variance in VC. More in detail:

20. Semantic broadening and narrowing have been captured with computational methods (contextualised BERT embeddings and three different metrics of change) by Giulianelli et al. (2020).

- 'three slices' setting: the model is significant as a whole ($p < 0.001$). However, the two factors only explain 0.39% variance in VC (adjusted R-squared = 0.0039). The effect of frequency is highly significant ($p < 0.001$), but it is a very small positive effect. The effect of the factor 'items' is nonsignificant, instead ($p = 0.0552$).

- 'five slices' setting: the whole model is significant ($p < 0.01$), but also in this case it only explains a very small amount of variance in VC, 0.39% (adjusted R-squared = 0.0039). The factor 'frequency' has a significant effect on VC ($p < 0.01$), but it is a very small positive effect. Also in this setting, 'items' has no significant effect on VC, instead ($p = 0.12$).

The tested method (word2vec $+VC$) returned valid results at a manual inspection, regarding words with both the highest and the lowest usage stability (Stopponi et al. 2024b), and the lower average VC calculated on a longer timespan (five slices vs. three slices) is also coherent with a higher potential amount of semantic change.[21] However, the statistical nonsignificance of the 'items' factor and the small amount of variance explained by the models above point to some issues with our benchmark and evaluation:

a) the benchmark is probably not a homogeneous set of items. Philological work, on which we based the selection of benchmark items, can identify cases of semantic change and highlight different usages of it as proofs of the change, but cannot quantify how strongly a word changed, nor compare the strength of semantic change of different words. The benchmark thus seems to include words with a different degree of semantic change, and with a different degree of polysemy resulting from change. [22] This would explain the variety in VC values assigned to benchmark items. Indeed, we suggests that computational methods and philological methods can be complementary in this respect, since computational methods can give a quantitative evaluation of the degree of change. While philological methods can assess and verify cases of semantic change, building on the knowledge and experience of scholars, and with outcomes validated by human judgements, computational methods can bring in a different dimension into the analysis, the degree of change. During the manual filtering of the candidates to go into the benchmark, we discarded words such as λαμπάς, *lampas* (see Table 2), considering their semantic change as too subtle to entail sensible modifications of its co-occurrences with other words. However, we did not discard other cases such as αὐγή, 'light' > 'dawn' and πηγή, 'stream of water', 'spring' > 'well', even if their degree of semantic change is not the highest because we considered them as still valid examples of change. These two words obtained the highest VC and J among benchmark items.

b) the method (word2vec models coupled with the vector coherence measure) is perhaps not powerful enough to detect all kinds and degrees of semantic change, including those present in the benchmark. If the context of usage of a word does not change enough through time, a low VC is indeed impossible to obtain. Since what cosine similarity captures is semantic relatedness (with high relatedness = high cosine similarity), if a new sense of a word is highly related to the pre-existing ones, the word is less likely to obtain a low VC because high cosine similarities are measured between its vectors in different slices.

c) non-benchmark items may include many words which also changed their meaning through time. Therefore, in our evaluation the changed items in the benchmark were compared with a set which also includes changed items, and the two sets (benchmark and non-benchmark) do

21. But note that the lower usage stability detected among the whole corpus, compared to restricting the analysis to three slices, could also be influenced by the greater variety in genres across a longer timespan.

22. An example of the different polysemy which can result from semantic change are the words ἐπίσκοπος, *episkopos* and κόσμος, *kosmos*. While the first undergoes a more radical change, passing from the meaning 'overseer', 'official' (in Jewish and other non-Christian societies) to 'Christian official', the second undergoes an increase in polysemy, by adding the meaning 'world' to the pre-existing meanings.

not represent two categories with a completely different behaviour, even if benchmark items obtain in all settings and analyses a lower average VC than non-benchmark items. We expect that many (probably most) words changed their meaning at least to some extent during such a long period of time–more than a thousand years.[23]

5.3 Changed vs. stable items

To address the last problem, we ran a third evaluation, in which we compare 11 words from the benchmark of changed items with a specular set of 11 stable, unchanged words. This second set was independently selected by a highly proficient student of Ancient Greek, not involved in this study. The only guidelines were about semantic stability and frequency: the words needed to be semantically stable from Archaic until Late Roman times, not to have changed under the pressure of strong well-known cultural, societal, or political changes (for example the advent of Christianity), and they could not be rare. Some examples of semantic areas where stable words are likely to be found[24] were also provided: family relations, food, natural elements, and animals. The student double-checked his candidates against the LSJ dictionary, to exclude to have overlooked relevant changes, for example in case the dictionary listed a specialistic meaning or a meaning tied to Christianity or to the Roman culture. To keep the task feasible and the quality of the chosen words high, the student was only asked to provide 11 words, the same number as the benchmark items tested in the 'five slices' setting. In this way we could compare two sets (changed and unchanged words) of equal size, without making ourselves an arbitrary selection among benchmark items. [25] All 11 stable words turned out to be present in all corpus slices with an acceptable frequency (≥ 50), except from three cases.[26] The full list of unchanged words is in Appendix C.

As expected, in both settings the benchmark of 11 changed items achieves an average lower VC than the 11 stable items. In the 'five slices' setting the average VC for changed items is 0.64, while it is 0.72 for stable items; in the 'three slices' setting it is 0.80 for changed items and 0.87 for stable items. These differences in VC are represented in Figure 4. To test their significance, we constructed two multiple linear regression models of normalised VC, both as a function of item type (changed vs. stable items) and of absolute frequency.[27]

The first model, concerning the 'three slices' setting, was significant ($F(2, 19) = 7.56, p = 0.0039$) and explained 38% of variance in the data (adjusted R-squared = 0.38), a much stronger effect than when comparing benchmark and non-benchmark items. The effect of both factors 'item type' ($p = 0.049$) and 'absolute frequency' ($p = 0.016$) was significant. Items belonging to the benchmark of 11 stable lemmas showed a positive effect on VC, with a VC on average 0.05 higher, compared to the 11 changed items. Absolute frequency also showed a small positive effect on VC. This result shows that, when the two sets of items are of comparable size and we have insight and control on the nature of the included items, our method is able to discriminate between changed and unchanged

23. According to what observed by Stopponi et al. (2024b), confirming what already found by Cassani et al. (2021) for English, stop-words and words denoting natural elements, animals, family relationships, and numbers seem to be the most resistant to change.
24. See the discussion in Stopponi et al. (2024b).
25. The 11 changed items used in this analysis are five nouns, three adjectives, and three verbs: ἐπίσκοπος, ἀφίημι, δαίμων, φυλακή, ὡραῖος, κηρύσσω, κόσμος, πηγή, λείπω, αὐγή, σῖτος. See Appendix B for their transliteration and translation.
26. The three cases all concern the Archaic slice, where πέμπτος *pemptos* has frequency 12, καθεύδω *katheudō* 7, and πλέω *pleō* 32 (still an acceptable frequency to obtain a meaningful vector representation).
27. We did consider using log-transformed values in place of absolute frequencies. We eventually opted to keep the absolute values on the ground of the experimental setup and hypothesis, which are concerned with testing the impact of frequency on vector stability. The difference between a frequency of 3 and a frequency of 6 is hypothesized to be less impactful for model stability than the difference, e.g., between 30 and 60. Very high numbers associated with most common words are also meaningful in absolute terms. See also the discussion in Section 5.4. We would be ready to include and discuss log-transformed results in the Appendix, or even the main body, should reviewers deem this useful or more appropriate.

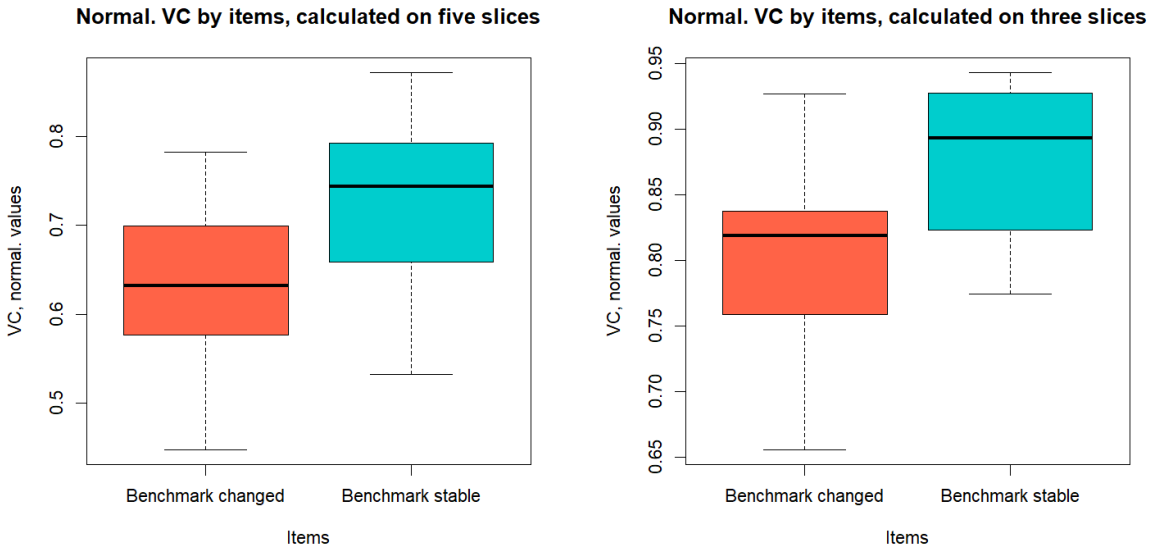


Figure 4: Boxplots showing the difference in VC between the two benchmarks of stable and changed items, divided by setting (five or three slices).

words. In this case, the large size of the effect (0.38) compensates for the small size of the sample (22 items in total).

The model for the 'five slices' setting was not significant, instead. The factors 'item type' and 'absolute frequency' had no effect on VC and the difference in average VC between the two groups of items could thus just be due to chance. The nonsignificance of this model, compared to the previous one, may be explained with a combination of small sample size (22 items) and small effect size. The adjusted R-squared is indeed 0.087, i.e. the model explains 8.7% of variance. The longer timespan in the 'five slices' setting, compared to the 'three slices', probably also plays a role towards the nonsignificance of the model, since a higher degree of change affects all the lexicon (benchmark *and* non-benchmark) when the whole corpus is used, especially because of the presence of the Archaic slice. This subcorpus, which mostly contains epic and didactic poetry and hymns, differs in genre from the following ones. Another relevant factor seems to be the presence of two smaller slices (Archaic and Late Roman), which likely introduce more randomness into the vector representations.

5.4 Qualitative evaluation of the detection of high-frequency-low-VC words

The significant relationship assessed between absolute frequency and VC in Section 5.2 (and, for the 'three slices' setting, in Section 5.3) reinforces the idea, already present in the scholarship, that frequency is an important factor to consider when dealing with semantic change. In this section we explore the relationship between VC and frequency further. We are particularly interested in evaluating cases of words with low VC and high frequency because embeddings for high-frequency words are calculated on a high number of contexts and are thus supposed to be more stable (Wendlandt et al. 2018). Their detection as candidates for semantic change is less likely to be due to chance since their vectors were not built from a few, possibly non-representative contexts, as can happen for very low-frequency words. Our hypothesis is that low-VC-high-frequency words could be more reliable candidates for semantic change.

In Table 5 we report for different types of items the average absolute frequency–defined as the sum of a word's frequency across all three or five slices–and the average frequency variability–defined

as the standard deviation from a word's average frequency. We observe that benchmark items have in both settings a lower average absolute frequency. One cause could be the absence in the benchmark of grammatical ('stop') words, which typically occur very frequently (and are often semantically stable). A higher frequency variability can be observed in the 'five slices' setting (697, calculated for all items) compared to the 'three slices' setting (167). This was to be expected, due not only to the longer timespan of the corpus and to its greater variation in genres and topics, but also to the inclusion of the first and last slices, Archaic and Late Roman, smaller in size. Many words are particularly rare there, especially in the Archaic slice, and this provokes a stronger imbalance in frequency across subcorpora, compared to the 'three slices' setting.

Setting	Type of items	N. items	Avg. absolute frequency	Frequency variability (Std Dev)
5 slices	Benchmark	11	1516	273
	Non-benchmark	2019	4071	699
	All lexicon	2030	4057	697
3 slices	Benchmark	37	1023	224
	Non-benchmark	8330	1126	167
	All lexicon	8367	1125	167

Table 5: Average absolute frequency and frequency variability per setting and item type: benchmark, non-benchmark, and all lexicon.

Even if there seems to be a relationship between high frequency and high VC, some lemmas with medium-high frequency also obtain lower VC values. An example is the word ὁ, *ho*, used as definite article and as relative or demonstrative pronoun. This lemma has the highest absolute frequency in the shared lexicon of the 'five slices' setting, 1,407,579, but has at the same time a below-average VC, 0.61 (the average VC calculated on all shared lexicon is 0.69 in this setting). The relatively low VC obtained by ὁ in this setting is caused by the very low cosine similarity (0.18) between its vectors in the Archaic and Classical slice, compared to the higher similarities ($\geq$ 0.85) between the other consecutive slices. This is likely to be due to the different usage of ὁ in the Homeric poems, contained in the Archaic slice, where ὁ is more commonly used with the older demonstrative function, instead of as a definite article, the most frequent usage later.

To identify the most representative cases of low-VC-high-frequency lemmas, we selected the 50 lemmas with the lowest VC in both settings and ranked them by absolute frequency. We show here some examples of lemmas which are at the top of the lists for the two settings, chosen after having excluded that the differences in vector representations between slices were caused by persistent lemmatisation errors. We propose an interpretation of the detected changes, based on the manual inspection of the words' nearest neighbours (NNs) and of the cosine similarities between their vectors in different time slices. As additional validation, we compare this information with what is found in two dictionaries of Ancient Greek. [28] Note that the cosine similarities we report, calculated between different slice combinations, may not be directly comparable with each other, for example because the total amount of change varies across different combinations of slices and because the amount of randomness in vector representations may be higher for smaller slices, where vectors are built from a lower number of occurrences. We thus advise to put the specific cosine similarity values in relation with the average cosine similarity for that combination of slices (see Table 6).

The examples below are neither exhaustive explorations of all meanings of the target words nor complete descriptions of all their changes through time, but rather a summary of initial observations

28. The LSJ dictionary (Liddell et al. 1940), a general dictionary Ancient Greek - English, and the *Homeric Dictionary* (Autenrieth 1887), focusing on word meaning in the Homeric poems only.

Slice combination	N. lemmas shared vocab.	Avg. cos. similarity
Archaic - Classical	2603	0.58
Classical - Hellenistic	8682	0.75
Hellenistic - Early Roman	10768	0.79
Early Roman - Late Roman	7021	0.70

Table 6: Average cosine similarity for combinations of consecutive slices. For each word in the shared lexicon between two slices, cosine similarity was calculated between its two slice-specific vectors. Such cosine similarities were subsequently averaged per pair of slices. The number of shared lemmas varies across slice combinations, resulting in different sample sizes for the averages.

and hypotheses about possible paths of change. We report these examples to show how our method may be used as a starting point for further research. Needless to say, only its integration with a close-reading analysis of the contexts of occurrence can lead to firm conclusions about whether a word underwent meaning change and about the specific meanings involved.

In the 'five slices' setting, among the words at the top of the list of low-VC-high-frequency lemmas, we find those listed below. The NNs are followed by their cosine similarity value with the target word. Not all meanings present in the LSJ dictionary have been reported, but only those relevant to this analysis:[29]

- δείκνυμι *deiknymi*, 'show, display, proof', VC 0.47, absol. freq. 4429: this lemma seems to acquire in the Classical subcorpus a specific usage related to oratory and philosophy, to which the nearest neighbours seem to point: συλλογισμός *syllogismos*, 'reasoning, syllogism, inference', 0.82; συμπέρασμα *symperasma*, 'conclusion (of a syllogism)', 0.81; καθόλου *katholou*, 'in general', 0.77;[30] πρότασις *protasis*, 'proposition, enunciation', 0.76; ἀντιστρέφω *antistrefō*, 'turn to the opposite side (also said of an argument)', 0.76; ἀπόδειξις *apodeixis*, 'making known, proof', 0.75. The nearest neighbours in the Hellenistic and Early Roman period, instead, do not relate specifically to argumentation. The first in the Hellenistic slice are: εἰμί *eimi*, 'to be, exist', 0.66; λέγω *legō*, 'say', 0.65; ἐρῶ *erō*, 'say, speak (future tense)', 0.63; and λογίζομαι *logizomai*, 'count, calculate', 0.59. In the Early Roman period: λέγω *legō*, 'say', 0.67; δηλόω *dēloō*, 'show', 0.66; λαλέω *laleō*, 'talk', 0.66; and ἐξευρίσκω *exeuriskō*, 'find out', 0.66. They seem to refer to a more general usage of δείκνυμι. A close-reading analysis could clarify whether the same holds for the Archaic slice (where the first nearest neighbours are λεύσσω *leussō*, 'see', 0.84; ἐπιτρέπω *epitrepō*, 'turn, commit, pass', 0.83; and κικλήσκω *kiklēskō*, 'call, summon', 0.83) and how widespread the more specialistic usage of the word in Classical times really is. The cosine similarity between the vectors of δείκνυμι in consecutive slices is 0.14 between Archaic and Classical slice, 0.57 between Classical and Hellenistic, 0.65 between Hellenistic and Early Roman, and 0.79 between Early and Late Roman.

- ὕπατος *hypatos*, 'highest, uppermost, consul', VC 0.34, absol. freq. 1817: as already discussed in Stopponi et al. (2024b), this term acquired the meaning 'consul' from Roman times (already from our 'Hellenistic' slice, for texts written under the Roman rule in Greece). This strong change is evident not only from the analysis of the nearest neighbours, but also from the cosine

29. The following words are all examples of so-called 'lexical words' but at the top of the low-VC-high-frequency list we also find grammatical words, such as ἔπειτα *epeita*, 'then/afterwards/therefore' (VC 0.48, absol. freq. 4714); αὖ *au*, 'again' (VC 0.39, absol. freq. 4078); and σχεδόν *schedon*, 'near/similar to/about' (VC 0.41, absol. freq. 2096). The direction of change in usage of these words is less easy to determine from the analysis of NNs, cosine similarities, and dictionary entries only, and we did not undertake a systematic exploration of their contexts of occurrence.

30. According to the Thesaurus Linguae Graecae, Aristotle is the author who uses this word the most.

similarity values between the vectors of ὕπατος in consecutive slices. It is indeed 0.51 between Archaic and Classical slice, -0.09 between Classical and Hellenistic, 0.94 between Hellenistic and Early Roman, and 0.45 between Early and Late Roman. ὕπατος also appears among the low-VC-high-frequency words in the 'three slices' setting.

- σημαίνω *sēmainō*, 'indicate, give a sign', VC 0.48, absol. freq. 1790: this lemma also seems to be mainly used in the Classical subcorpus in contexts tied to oratory and philosophy, while the usage in the Hellenistic and Early Roman slices points to a military usage of the word. Examples of nearest neighbours in the Classical slice are γνωστός *gnōstos*, 'known, knowable', 0.69; ἀντίθεσις *antithesis*, 'opposition, antithesis', 0.69; and νοέω *noeō*, 'perceive, think', 0.69. In the Hellenistic slice the first neighbours are: σάλπιγξ *salpinx*, 'war trumpet', 0.85; σαλπιγκτής *salpinktēs*, 'trumpeter', 0.81; ἀλαλάζω *alalazō*,'raise the war-cry', 0.76; σαλπίζω *salpizō*, 'sound the trumpet', 0.76. In the Early Roman slice: σάλπιγξ *salpinx*, 'war trumpet', 0.73; σαλπίζω *salpizō*,'sound the trumpet', 0.67; and ἐπισημαίνω *episēmainō*,'mark, indicate, give sign', 0.64. The cosine similarities between vectors of σημαίνω in consecutive slices are: 0.27 between Archaic and Classical slice, 0.49 between Classical and Hellenistic, 0.77 between Hellenistic and Early Roman, and 0.59 between Early and Late Roman.

- λαμπρός *lampros*, 'bright, limpid/well-known, illustrious/ splendid, brilliant', VC 0.48, absol. freq. 1768: its nearest neighbours in the Archaic and Classical slice refer to a 'physical' brightness. The closest in the Archaic space are: αὐγή *augē*, 'light (of the sun)', 0.88; κυάνεος *kyaneos*, 'dark, black', 0.86; ἀστήρ *astēr*, 'star', 0.85; λαμπετάω *lampetaō*, 'shine', 0.85; and νεφέλη *nephelē*, 'cloud', 0.84. In the Classical space: λάμπω *lampō*, 'shine', 0.72; αὐγή *augē*, 'light (of the sun)', 0.72; σέλας *selas*, 'light, brightness', 0.71; and ἐκλάμπω *eklampō*, 'shine', 0.69. Conversely, the nearest neighbours from the Hellenistic slice refer to a figurative brightness, the moral or sporting excellence of a person: ἡρωικός *hēroikos*, 'of the heroes, heroic', 0.82; ἀγών *agōn*, 'assembly, struggle', 0.78; γενναῖος *gennaios*, 'noble, excellent', 0.76; ἔκθυμος *ekthymos*, 'spirited, ardent', 0.73; ἀθλητής *athlētēs*, 'athlete, champion', 0.73; and γυμνικός *gymnikos*, 'of/for gymnastic exercises', 0.70. The cosine similarities between vectors of λαμπρός in consecutive slices are: 0.57 between Archaic and Classical slice, 0.36 between Classical and Hellenistic, 0.76 between Hellenistic and Early Roman, and 0.75 between Early and Late Roman. They confirm that a strong change could have taken place between our Classical and Hellenistic slice.

Examples of good candidates for semantic change at the top of the list of low-VC-high-frequency words for the 'three slices' setting are the following:

- προφήτης *profētēs*, 'interpreter of the divine will, prophet', VC 0.42, absol. freq. 893: this word, first used to indicate an interpreter of the will of some god in the Greek polytheistic religion, is used in Hellenistic times in the *Septuagint* to indicate a prophet of the Jewish religion, and later reused in Christian contexts. This is evident from the nearest neighbours analysis. The nearest neighbours in the Classical slice, clearly pointing to the Greek polytheistic religion, are: λίσσομαι *lissomai*, 'pray', 0.91; Κρονίδης *Kronidēs*, 'son of Cronos (epithet for the god Zeus)', 0.91; Θέμις *Themis*, 'goddess of justice and prophecy', 0.90; and μάντευμα *manteuma*, 'oracle', 0.90. Those in the Hellenistic slice, related to the Jewish religion, are: προφητεύω *prophēteuō*, 'to be an interpreter of the gods/a prophet', 0.88;[31] ἄγγελος *angelos*, 'messenger/angel', 0.81; ἰδού *idou*, 'behold!', 0.80; Ἰησοῦς *Iesous*, Greek translation of different Jewish names in the *Septuagint*, 0.79; ῥῆμα *rhēma*, 'word, saying', 0.76. Finally, the neighbours in the Early Roman period point to a Christian usage of the term: Ἰησοῦς *Iesous*, Greek translation of different

31. προφητεύω *prophēteuō* only occurs ten times before the *Septuagint* in the TLG corpus, and occurs most frequently later, in Christian authors. This word (and the concept it conveys) seems thus to acquire special prominence in Jewish and Christian texts.

Jewish names, including the Jesus of the *New Testament*, 0.87; καθώς *kathōs*, 'how/when', 0.85 (this adverb is typical of post-Classical Greek and used the most often in the *Septuagint* and the *New Testament* in the Diorisis corpus); ἀπόστολος *apostolos*, 'messenger, apostle, 0.84; ἄγγελος *angelos*, 'messenger/angel', 0.84; προφητεύω *prophēteuō*, 'to be an interpreter of the gods/a prophet', 0.79. That the strongest change happens between the Classical and Hellenistic slice is confirmed by the low cosine similarity between these two slices 0.24, compared to 0.86 between Hellenistic and Early Roman.

- παραβάλλω *paraballō*, *VC* 0.27, absol. freq. 682: this word is highly polysemic. According to the LSJ dictionary, the main meanings are: 'throw beside/by', 'expose', 'set beside or parallel with, compare', 'throw, turn, bend sideways', 'come near, approach'. The cosine similarity between the slice-specific vectors of παραβάλλω is low between both slice combinations, being 0.34 between Classical and Hellenistic slice and 0.20 between Hellenistic and Early Roman. From the analysis of the nearest neighbours it is not completely clear which meanings are involved in the difference in usage between the three slices, especially because of the word's polysemy. It is realistic to hypothesise that, if not all, at least several of the word's meanings coexist in each slice, informing the related vector. However, we do detect a difference in associations between the Classical and the Early Roman slice, with the Hellenistic which seems to present an intermediate situation. The nearest neighbours from the Classical slice point indeed to meanings such as 'throw (beside), turn, bend sideways'. The first are: ἀνατείνω *anateinō*, 'lift up/spread out/persevere', 0.85; νεύω *neuō*, 'incline/nod/decline', 0.84; πλάγιος *plagios*, 'pointed sideways, oblique', 0.83; ἐξάπτω *exaptō*, 'fasten to, hang by, cling to', 0.82; and ὀχάομαι *ochaomai*, 'leap', 0.81. The neighbours from the Early Roman slice seem to be associated to the meaning 'expose' of παραβάλλω, instead: μιμέομαι *mimeomai*, 'imitate/represent', 0.73; ἐπιδείκνυμι *epideiknymi*, 'exhibit, show off', 0.68; ἔπαινός *epainos*, 'approval, praise', 0.66; ἀπεικάζω *apeikazō*, 'form from a model, represent', 0.65; and ἀποφαίνω *apophainō*, 'show forth/display/represent', 0.63. The nearest neighbours in the Hellenistic slice could point to an in-between situation: παράκειμαι *parakeimai*, 'lie beside or before/to be mentioned', 0.66; ἔκκειμαι *ekkeimai*, 'to be cast out, exposed', 0.64; ἐλλείπω *elleipō*, 'leave in/ leave out/fail', 0.64; πλάτος *platos*, 'bredth/width', 0.63; ἀναγράφω *anagraphō*, 'engrave, register', 0.59. Only a close-reading analysis could confirm whether this hypothesis about different prevailing meanings of the word in different time periods is correct and whether a change in meaning happened.

- ἐπιβολή *epibolē*, *VC* 0.22, absol. freq. 564. This is another highly polysemous word, meaning 'throwing, laying on' with both a concrete and a psychological sense: it can mean 'application of the mind/conception/notion', 'setting upon a thing, design, enterprise', 'hostile attempt, assault', but also 'a thing put over for shelter or protection' and 'fine/penalty/requisition/impost'. The last meaning is very present in the Classical slice, where it appears for example in the *Wasps* by Aristophanes and in orations by Demosthenes, Lysias, and Andocides. The nearest neighbours in this slice recall a judicial and institutional setting: δήμαρχος *dēmarchos*, 'demarch, chief official of a deme in Attica', 0.85; πωλητής *pōlētēs*, 'seller/official who farmed out taxes', 0.84; λογιστής *logistēs*, 'auditor, esp. at Athens, a board which audited the accounts of magistrates going out of office', 0.83; ἀποστολεύς *apostoleus*, 'one who dispatches/at Athens, magistrate who had to fit out a squadron for service', 0.83; ἡλιαία *ēliaia*, 'public place or hall/supreme court at Athens', 0.82. The nearest neighbours in the Hellenistic period recall some of the other meanings of the word, instead, both concrete and psychological, including the meaning 'assault', suggesting usage in military context: πρᾶγμα *pragma*, 'deed, act/affair', 0.83; ἐπίνοια *epinoia*, 'thought/invention', 0.83; πόλεμος *polemos*, 'war', 0.82; ἐλπίς *elpis*, 'hope/expectation', 0.78; ὁρμή *ormē*, 'assault, attack', 0.77. Finally, the nearest neighbours in the Early Roman slice seem to point to a psychological usage of ἐπιβολή, such as 'application of the mind/conception/notion': σχέσις *schesis*, 'state, condition, habit', 0.82; φαντασία *phantasia*, 'appearance/imagination', 0.81; νόημα *noēma*, 'percep-

tion/thought/purpose/understanding', 0.81; τόνος *tonos*, 'that by which a thing is stretched, or that which can itself be stretched/mental or physical exertion/tension', 0.81; πρόληψις *prolēpsis*, 'preconception', 0.80. The cosine similarity between the vectors of ἐπιβολή suggests a stronger difference in usage between Classical and Hellenistic slice (cosine similarity 0.10) than between Hellenistic and Early Roman (cosine similarity 0.43). Again, only a close-reading analysis can confirm which meanings are predominant in the three subcorpora and whether actual meaning change is at hand.

This analysis shows how low VC does not necessarily entail semantic change, but it quantifies variability in usage in the first place. Variability can originate from actual semantic change, but also from other factors, such as the different genre composition of the slices, or a high degree of polysemy of the word. The analysis also shows the potential of the method as a starting point for further research. An example of a research area which could be expanded, taking as a point of departure one of our automatic detections, are Ancient Greek words which changed their meaning under the influence of the Romans. A result belonging to this group is the word ὕπατος, *hypatos*, which added to its earlier meanings 'highest, uppermost' the meaning 'consul', from Roman times. ὕπατος could be the starting point for a wider investigation about Roman-driven semantic change in the Ancient Greek lexicon, an apparently underexplored area.[32] Just looking at the ten nearest neighbours of ὕπατος, we find at least two other potential cases of words which changed their meaning for reasons tied to the Roman expansion: δήμαρχος, *dēmarchos*, originally 'demarch' (at Athens), later 'tribune of the plebs' (a Roman office); and συνάρχω, *synarchō*, 'to be a colleague in office', used in Roman times to refer to the consuls, ruling together. Expanding on this line of research, by leveraging both computational and close-reading methods, could answer questions about the impact of the Roman arrival on the Ancient Greek lexicon, for example about its extent, permanence, and diffusion across different genres.

6. Conclusions

We described the creation of a benchmark for the evaluation of methods for lexical semantic change detection for Ancient Greek and the results of the evaluation of a specific metric, the vector coherence, coupled with word2vec embeddings. Our study exposed the challenges of performing this kind of evaluation on an ancient language, namely Ancient Greek. First of all, creating a reliable and representative benchmark of semantically changed words is a hard task, mainly due to the unavailability of native speakers and, in the case of Ancient Greek, to the scarcity of scholarly work clearly identifying cases of semantic change, including the specific meanings involved in the change and the timespan. A manual evaluation showed that the proposed method is able to detect valid cases of semantic change, and we consider detections with high frequency of occurrence as particularly reliable candidates for semantic change. By using our benchmark, we assessed that the method assigns a higher degree of (potential) change (lower average VC) to benchmark items–known cases of semantic change–than to non-benchmark items, which receive a higher average VC. Consistently, it also assigned a higher degree of change to benchmark items, compared to a set of semantically stable items. Nevertheless, the differences in VC were statistically significant in only one case, when changed and stable items were compared in the 'three slices' setting. The small size of the benchmark seems to be a major factor conditioning the effectiveness of the evaluation, together with the size and composition of the training subcorpora. The small size of the training corpus we adopted, the Diorisis Ancient Greek Corpus, limited indeed the possibilities of LSC detection. Using a larger cor-

32. Though there is plenty of studies about the reciprocal influence of Roman and Greek culture, and there is scholarly work about Latin loans in Ancient Greek, such as Dickey (2023), work investigating specific, preexisting Ancient Greek words which changed their meaning after the arrival of the Romans seems to be lacking.

pus of Ancient Greek, such as the recently released GLAUx[33] and Opera Graeca Adnotata,[34] would improve embedding stability, and thus the reliability and effectiveness of detection. Moreover, the benchmark includes different kinds and degrees of semantic change, and we do not know whether the method is equally effective at detecting all of them. We suspect it could be more effective at capturing cases of 'sharp' semantic change, those with a particularly strong change in context of usage through time, while subtler changes may be assigned a higher VC, or even go unnoticed by the method. More research is needed to clarify this, ideally by testing the method against larger benchmarks on larger corpora of different languages. Finally, this study leaves open various possibilities for the usage of VC. For example, while in this work we followed the recommendation by Cassani et al. (2021) to sum over all slices, one could experiment with different ways of calculating the metric, such as summing up only the cosine similarities calculated between consecutive slices.

At this stage, we see this method as a support to philological work, not as a predictive tool. It can work as a research companion, complementing and enhancing the scholar's knowledge and intuition, and help navigating semantic change beyond personal biases and knowledge limitation. We consider it as a tool which can suggest new ideas and research paths, as exemplified in Section 5.4 and suggested by Jatowt and Duh (2014), or help recalling relevant knowledge. But–as it should always be the case in NLP research–the human user retains the ultimate judgement on the meaningfulness of the tool's output.

We are aware that obstacles exist to the use of this method in the daily practice of many scholars of Ancient Greek linguistics who are not familiar with computational methods: the inaccessibility of most computational techniques to scholars without programming skills is a well-known issue. A user-friendly interface to leverage this or other computational methods for LSC detection for Ancient Greek does not exist (yet), but a first step into this direction is our AGALMA interface (Stopponi et al. 2024a),[35] an open-access tool making word vector representations accessible to anyone with knowledge of the Ancient Greek language. AGALMA (Ancient Greek Accessible Language Models for linguistic Analysis) allows to extract the nearest neighbours to a target word in the five time slices of the Diorisis corpus, to compute cosine similarity values between word vectors, and to generate 3D visualisations of the semantic spaces. AGALMA comes with a list of Frequent Questions and Answers intended to make users aware of the functioning, the potential, and the limitations of the tool. It is thus a first, concrete opportunity for 'non-computational' classicists to extract information from language models trained on Ancient Greek and to use them for their own research.

7. Acknowledgements

This work was partially supported by the Young Academy Groningen through the PhD scholarship of Silvia Stopponi.

We also acknowledge the financial support of Anchoring Innovation. Anchoring Innovation is the Gravitation Grant research agenda of the Dutch National Research School in Classical Studies, OIKOS. It is financially supported by the Dutch ministry of Education, Culture and Science (NWO project number 024.003.012). For more information about the research programme and its results, see the website `www.anchoringinnovation.nl`.

We thank Sven Smeman for selecting the set of stable words. We thank Greta Zella and many participants in the workshop *Days of Computational Approaches to Ancient Greek and Latin* (Leuven, November 2024) for helping us improve this paper through fruitful discussion.

33. GLAUx, available at `https://www.glaux.be/` and `https://github.com/alekkeersmaekers/glaux`, with its 20 millions tokens is double the size of the Diorisis corpus.
34. Opera Graeca Adnotata is the largest machine-actionable corpus for Ancient Greek, counting more than 40 million tokens. It is available at `https://github.com/OperaGraecaAdnotata/OGA`.
35. Available at `https://huggingface.co/spaces/GroNLP/agalma`.

8. Authors' contributions[36]

Conceptualisation: Malvina Nissim, Saskia Peels-Matthey, Silvia Stopponi
Methodology: Malvina Nissim, Saskia Peels-Matthey, Silvia Stopponi
Software: Silvia Stopponi
Formal analysis: Silvia Stopponi
Investigation: Saskia Peels-Matthey, Silvia Stopponi
Data Curation: Saskia Peels-Matthey, Silvia Stopponi
Writing - Original Draft: Malvina Nissim, Saskia Peels-Matthey, Silvia Stopponi
Writing - Review and Editing: Malvina Nissim, Saskia Peels-Matthey, Silvia Stopponi
Supervision: Malvina Nissim, Saskia Peels-Matthey
Funding acquisition: Malvina Nissim, Saskia Peels-Matthey

9. Appendix A: Selection of the evaluation items

To give a more complete picture of the process of selection of the items included in the benchmark, we offer here more examples of included and excluded items.

Among candidates extracted from the scholarly work, there were items which were not separate lemmas in the Diorisis corpus, and which were unlikely to be separately lemmatised in any other corpus. An example is πρεσβύτερος, *presbyteros*, the comparative of the adjective πρέσβυς, *presbys* 'old'. The comparative literally means 'older', but it was also used to refer to officials in Jewish and other non-Christian societies, and later changed its meaning to refer to Christian officials. Since πρεσβύτερος is a comparative, all its occurrences are lemmatised under πρέσβυς in the Diorisis corpus. Therefore, a change in the meaning of just πρεσβύτερος would be difficult to detect, because the vector includes all the other attestations of the lemma πρέσβυς as well.

Similarly, the word ἔθνη, *ethnē*, which changed its meaning from 'nations' to the Christian concept of 'Gentiles', is the plural of ἔθνος, *ethnos* 'nation/ people/tribe'. For the same reasons as above, this word was not included in the benchmark.

Other remarkable cases are the words ἐπίσκοπος, διάκονος, ἀπόστολος, and μάρτυς, adopted in the Christian society and culture to denote specific kinds of people. All of them were included in the benchmark since they underwent a strong change in meaning. ἐπίσκοπος, *episkopos*, originally meaning 'overseer', was used for officials in Jewish and other non-Christian societies, and later came to designate a Christian official. διάκονος, *diakonos* is a parallel case. From the original meaning 'servant' it was subsequently used for officials in Jewish and other non-Christian societies, and later for Christian officials. ἀπόστολος, *apostolos* changed its meaning from 'messenger' into 'apostle', while μάρτυς, *martys* changed from 'witness' to 'martyr'.

10. Appendix B: Benchmark of changed items

Lemma	**Reference**	**Which Change**
ἀγάπη, *agapē*	Gingrich (1954, 190)	Different emotional association to the Christian usage of the word. It becomes 'Christian love'.
διαθήκη, *diathēkē*	Gingrich (1954, 191)	'last will', 'testament' > 'God's covenant' ('new covenant').

36. Authors in alphabetical order.

κληρονομία, *klēronomia*	Gingrich (1954, 191)	'heritage' > 'property'. Following the Septuagint (LXX), the New Testament (NT) uses it "in a figurative sense, but also enlarges it to include the whole Christian heritage of salvation".
μετανοέω, *metanoeō*	Gingrich (1954, 191)	'change of mind' before the LXX > Hebrew idea of repentance > 'conversion', 'requirement for entrance into the kingdom' in the NT.
διάβολος, *diabolos*	Gingrich (1954, 191)	'slanderer' > 'evil par excellence'.
δαίμων, *daimōn*	Gingrich (1954, 191)	'divinity' > 'evil spirit', 'demon'.
δαιμόνιον, *daimonion*	Gingrich (1954, 191)	'divinity' > 'evil spirit', 'demon'.
κηρύσσω, *kērryssō*	Gingrich (1954, 191)	'proclaim as a herald' > 'preach'.
ἐπίσκοπος, *episkopos*	Gingrich (1954, 191)	'overseer', official in Jewish and other non-Christian societies > 'Christian official'.
διάκονος, *diakonos*	Gingrich (1954, 191)	'servant', official in Jewish and other non-Christian societies > 'Christian official'.
ἀπόστολος, *apostolos*	Gingrich (1954, 192)	'messenger' > 'apostle'.
εὐαγγέλιον, *euangelion*	Gingrich (1954, 192)	'reward for good news' in Hom. > 'good news' in later Greek and in the LXX > 'good news about Christ', 'gospel' > 'gospel' as a literary genre.
παρασκευή, *paraskeuē*	Gingrich (1954, 192)	'preparation' > 'day of preparation before the Sabbath'.
μάρτυς, *martys*	Gingrich (1954, 193)	'witness' > 'martyr'.
λαλέω, *laleō*	Gingrich (1954, 193), Buck 1254	'gossip', 'chatter' in Classical Greek > 'say' in Hellenistic Greek.
βρέχω, *brechō*	Gingrich (1954, 193)	'wet' (i.e. a transitive verb, of persons walking through water, and wetting their knees or feet) > 'rain' (in addition to 'wet').
ὑπάρχω, *yparchō*	Gingrich (1954, 193)	'begin' > 'equivalent to εἰμί'.
ἀφίημι, *aphiēmi*	Gingrich (1954, 193)	'send' > acquires the meanings 'let go' and 'leave'.
λείπω, *leipō*	Gingrich (1954, 195)	'leave' > acquires meaning 'be lacking'.

αἰσθάνομαι, *aisthanomai*	Gingrich (1954, 194)	'perceive' > 'understand' in Hellenistic Greek.
βάσανος, *basanos*	Gingrich (1954, 194), Buck 1115	'touchstone' > 'test' > "torture', 'pain', 'disease' in the NT.
δέρω, *derō*	Gingrich (1954, 194), Buck 553	'flay', 'skin' > 'beat' in the NT.
ἐπαγγελία, *epangelia*	Gingrich (1954, 194)	'announcement' > 'promise'.
ἐπακολουθέω, *epakoloutheō*	Gingrich (1954, 194)	'follow' > 'authenticate', 'confirm'.
κοινωνία, *koinōnia*	Gingrich (1954, 194)	'association' > 'contribution'.
περιπατέω, *peripateō*	Gingrich (1954, 194)	'walk around' > 'walk' and metaphorical 'live'.
προσευχή, *proseuchē*	Gingrich (1954, 194)	'prayer' > 'place of prayer'.
χρηματίζω, *chrēmatizō*	Gingrich (1954, 190)	'have dealings with' > 'give a divine revelation', 'be named' in the NT.
στέγω, *stegō*	Gingrich (1954, 194)	'cover' > 'endure', 'pass over in silence'.
ἄνοια, *anoia*	Gingrich (1954, 194)	'foolishness' > 'anger'.
ἁπλότης, *aplotēs*	Gingrich (1954, 194)	'simplicity' > 'generosity'.
ἄτοπος, *atopos*	Gingrich (1954, 194)	'out of place' > 'wicked', 'wrong'.
αὐγή, *augē*	Gingrich (1954, 194)	'light' > 'dawn'.
ἐντροπή, *entropē*	Gingrich (1954, 194)	'respect' > 'shame'.
ὑποπιάζω, *ypopiazō*	Gingrich (1954, 195)	'give someone a black eye' > 'annoy', 'trouble', 'mortify'.
ὄψον, *opson*	LSJ. This change was reported by Gingrich (1954, 195) for the word ὀψάριον, *opsarion*, but it is an extremely rare word, only occurring 14 times in the TLG. Consequently, automatic detection of semantic change over a long timespan would not have been possible. Since an analogous development is recorded in the LSJ for the more frequent ὄψον, we included this word in the benchmark instead.	'cooked food', 'relish' > 'cooked fish'.
πηγή, *pēgē*	Gingrich (1954, 195)	'stream of water', 'spring' > 'well'.
σῖτος, *sitos*	Gingrich (1954, 195), Buck 514	'grain' > 'wheat'.
τρώγω, *trōgō*	Gingrich (1954, 195)	'nibble', 'gnaw' > 'eat'.

φυλακή, *phylakē*	Gingrich (1954, 196)	'guarding' > 'prison'.
ὡραῖος, *ōraios*	Gingrich (1954, 196)	'seasonable', 'ripe' > 'beautiful'.
ἀρέσκω, *areskō*	Luraghi (2022)	'make amends' > 'appease', in archaic poetry > acquires the additional meaning 'like' in the Classical period.
ἐντυγχάνω, *entynchanō*	Gingrich (1954, 189–190)	'meet with' > 'read' in the NT.
κόσμος, *kosmos*	(Finkelberg 1998, 122); Horky (2019)	'order', 'ornamentation' > acquires the additional meaning of 'world, earth'.

11. Appendix C: Benchmark of stable items

The list of 11 stable items is the following. N, A, and V indicate respectively a noun, an adjective, and a verb, while m., f., and n. stand for 'masculine', 'feminine', and 'neuter':

1. ὀφθαλμός *ophthalmos*, 'eye' (N m.)
2. γόνυ *gony*, 'knee' (N n.)
3. ποταμός *potamos*, 'river' (N m.)
4. ὄρος *oros*, 'mountain' (N n.)
5. θυγάτηρ *thygatēr*, 'daughter' (N f.)
6. ταχύς *tachys*, 'quick' (A)
7. μέλας *melas*, 'black' (A)
8. πέμπτος *pemptos*, 'fifth' (A)
9. καθεύδω *katheudō*, 'sleep' (V)
10. πλέω *pleō*, 'sail' (V)
11. πέτομαι *petomai*, 'fly' (V)

References

Antoniak, Maria and David Mimno (2018), Evaluating the stability of embedding-based word similarities, *Transactions of the Association for Computational Linguistics* **6**, pp. 107–119, MIT Press, Cambridge, MA. https://aclanthology.org/Q18-1008.

Autenrieth, Georg (1887), *A Homeric Dictionary, for Schools and Colleges: Based Upon the German of Dr. Georg Autenrieth*, Harper & brothers.

Bamman, David and Gregory Crane (2011), Measuring historical word sense variation, *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 1–10.

Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro (2014), Analysing word meaning over time by exploiting temporal random indexing, *Analysing Word Meaning Over Time by Exploiting Temporal Random Indexing* pp. 38–42, Pisa University Press.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* **67** (1), pp. 1–48.

Bianchi, Federico, Valerio Di Carlo, Paolo Nicoli, and Matteo Palmonari (2020), Compass-aligned distributional embeddings for studying semantic differences across corpora, *arXiv preprint arXiv:2004.06519*.

Boschetti, Federico (2009), A corpus-based approach to philological issues.

Boschetti, Federico (2018), *Copisti digitali e filologi computazionali*, CNR Edizioni. `http://hdl.handle.net/20.500.11752/OPEN-89`.

Buck, Carl Darling (1949), *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A Contribution to the History of Ideas*, University of Chicago Press.

Caputo, Annalina, Pierpaolo Basile, and Giovanni Semeraro (2015), Temporal random indexing: A system for analysing word meaning over time, *IJCoL. Italian Journal of Computational Linguistics* **1** (1-1), pp. 61–74, Accademia University Press.

Cassani, Giovanni, Federico Bianchi, and Marco Marelli (2021), Words with consistent diachronic usage patterns are learned earlier: A computational analysis using temporally aligned word embeddings, *Cognitive science* **45** (4), pp. e12963, Wiley Online Library.

Cresswell, Julia (2010), *Oxford dictionary of word origins*, Oxford University Press, USA.

Davies, Mark (2012), The 400 million word corpus of historical american english (1810-2009) mark davies brigham young university the 400 million word corpus of historical american english (1810–2009), *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010*, Vol. 325, John Benjamins Publishing, p. 231.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), Bert: Pre-training of deep bidirectional transformers for language understanding, *in* Burstein, Jill, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. `https://aclanthology.org/N19-1423`.

Di Carlo, Valerio, Federico Bianchi, and Matteo Palmonari (2019), Training temporal word embeddings with a compass, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6326–6334.

Dickey, Eleanor (2023), *Latin loanwords in ancient Greek: a lexicon and analysis*, Cambridge University Press.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (2019), Time-out: Temporal referencing for robust modeling of lexical semantic change, *in* Korhonen, Anna, David Traum, and Luís Márquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 457–470. `https://aclanthology.org/P19-1044`.

Eger, Steffen and Alexander Mehler (2016), On the linearity of semantic change: Investigating meaning variation via dynamic graph models, *in* Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Berlin, Germany, pp. 52–58. https://aclanthology.org/P16-2009.

Finkelberg, Aryeh (1998), On the history of the greek koσmoσ, *Harvard Studies in Classical Philology* pp. 103–136, JSTOR.

Gingrich, F. Wilbur (1954), The greek new testament as a landmark in the course of semantic change, *Journal of Biblical Literature* pp. 189–196, JSTOR.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández (2020), Analysing lexical semantic change with contextualised word representations, *in* Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 3960–3973. https://aclanthology.org/2020.acl-main.365.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg (2020), Simple, interpretable and stable method for detecting words with usage change across corpora, *in* Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 538–555. https://aclanthology.org/2020.acl-main.51.

Gulordava, Kristina and Marco Baroni (2011), A distributional similarity approach to the detection of semantic change in the google books ngram corpus., *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pp. 67–71.

Haagsma, Hessel and Malvina Nissim (2017), A critical assessment of a method for detecting diachronic meaning shifts: Lessons learnt from experiments on dutch, *Computational Linguistics in the Netherlands Journal* **7**, pp. 65–78.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016a), Cultural shift or linguistic drift? comparing two computational measures of semantic change, *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, Vol. 2016, NIH Public Access, p. 2116.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016b), Diachronic word embeddings reveal statistical laws of semantic change, *in* Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, pp. 1489–1501. https://aclanthology.org/P16-1141.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016c), Diachronic word embeddings reveal statistical laws of semantic change, *arXiv preprint arXiv:1605.09096*.

Harper, Douglas (n.d.), Online etymology dictionary. http://www.etymonline.com/.

Horky, Philipp Sidney (2019), When did kosmos become the kosmos, *Cosmos in the Ancient World* pp. 22–41, Cambridge University Press.

Jatowt, Adam and Kevin Duh (2014), A framework for analyzing semantic change of words across time, *IEEE/ACM joint conference on digital libraries*, IEEE, pp. 229–238.

Keersmaekers, Alek (2020), A computational approach to the greek papyri: Developing a corpus to study variation and change in the post-classical greek complementation system, *PhD diss., KU Leuven*.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (2014a), Temporal analysis of language through neural language models, *in* Danescu-Niculescu-Mizil, Cristian, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Association for Computational Linguistics, Baltimore, MD, USA, pp. 61–65. https://aclanthology.org/W14-2517.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov (2014b), Temporal analysis of language through neural language models, *in* Danescu-Niculescu-Mizil, Cristian, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, Association for Computational Linguistics, Baltimore, MD, USA, pp. 61–65. https://aclanthology.org/W14-2517.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena (2015), Statistically significant detection of linguistic change, *Proceedings of the 24th international conference on world wide web*, pp. 625–635.

Kutuzov, Andrey and Mario Giulianelli (2020), Uio-uva at semeval-2020 task 1: Contextualised embeddings for lexical semantic change detection, *in* Herbelot, Aurelie, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, Barcelona (online), pp. 126–134. https://aclanthology.org/2020.semeval-1.14.

Liddell, Henry George, Robert Scott, H. S. Jones, and R. McKenzie (1940), *A Greek and English Lexicon*, Oxford: Clarendon Press.

Luraghi, Silvia (2022), The verb aréskein in ancient greek: Constructions and semantic change, *Acta Linguistica Petropolitana* **18** (1), pp. 226–245.

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), pp. 893–907, Oxford University Press.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray (2019), GASC: Genre-aware semantic change for Ancient Greek, *in* Tahmasebi, Nina, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Association for Computational Linguistics, Florence, Italy, pp. 56–66. https://aclanthology.org/W19-4707.

Perrone, Valerio, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray (2021), Lexical semantic change for Ancient Greek and Latin, *Computational approaches to semantic change* pp. 287–310.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018), Deep contextualized word representations, *in* Walker, Marilyn, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237. https://aclanthology.org/N18-1202.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rocci, Lorenzo (1939), *Vocabolario greco-italiano*, Società Editrice Dante Alighieri.

Rodda, Martina A, Marco SG Senaldi, and Alessandro Lenci (2017), Panta rei: Tracking semantic change with distributional semantics in ancient greek, *IJCoL. Italian Journal of Computational Linguistics* **3** (3-1), pp. 11–24, Accademia University Press.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2009), Semantic density analysis: Comparing word meaning across time and phonetic space, *in* Basili, Roberto and Marco Pennacchiotti, editors, *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, Association for Computational Linguistics, Athens, Greece, pp. 104–111. `https://aclanthology.org/W09-0214`.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2011), Tracing semantic change with latent semantic analysis, *Current methods in historical semantics* **73**, pp. 161–183, De Gruyter Mouton Berlin, Germany.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi (2020), Semeval-2020 task 1: Unsupervised lexical semantic change detection, *arXiv preprint arXiv:2007.11464*.

Sluiter, Ineke, Lucien van Beek, Ton Kessels, and Albert Rijksbaron (2024), *Woordenboek Grieks/Nederlands*, Amsterdam University Press.

Spanopoulos, Andreas I. (2022), Language Models for Ancient Greek.

Stopponi, Silvia, Mark den Ouden, Saskia Peels-Matthey, and Malvina Nissim (2024a), AGALMA, the Ancient Greek Accessible Language Models for linguistic Analysis. `https://huggingface.co/spaces/GroNLP/agalma`.

Stopponi, Silvia, Saskia Peels-Matthey, and Malvina Nissim (2024b), Viability of automatic lexical semantic change detection on a diachronic corpus of literary ancient greek, *The First Workshop on Data-driven Approaches to Ancient Languages (DAAL 2024)*, Ghent University, pp. 47–57. `https://www.dbbe2024.ugent.be/wp-content/uploads/2024/06/DAAL_proceedingsU1.pdf`.

Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen (2021a), *Computational approaches to semantic change*, BoD–Books on Demand.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2021b), Survey of computational approaches to lexical semantic change detection, *Computational approaches to semantic change*, Language Science Press Berlin.

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

Vejdemo, Susanne and Thomas Hörberg (2016), Semantic factors predict the rate of lexical replacement of content words, *PloS one* **11** (1), pp. e0147924, Public Library of Science San Francisco, CA USA.

Wang, Ruiyu and Matthew Choi (2023), Large language models on lexical semantic change detection: An evaluation, *arXiv preprint arXiv:2312.06002*.

Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea (2018), Factors influencing the surprising instability of word embeddings, *in* Walker, Marilyn, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 2092–2102. `https://aclanthology.org/N18-1190`.

Winter, Bodo, Graham Thompson, and Matthias Urban (2014), Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure, *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)*, World Scientific, pp. 353–360.