Evaluation of Distributional Semantic Models of Ancient Greek: Preliminary Results and a Road Map for Future Work

Silvia Stopponi
CLCG, University of Groningen, The Netherlands
s.stopponi@rug.nl

Nilo Pedrazzini
The Alan Turing Institute / University of Oxford
United Kindgom
npedrazzini@turing.ac.uk

Saskia Peels
CLCG, University of Groningen, The Netherlands
s.peels@rug.nl

Barbara McGillivray
King's College London
United Kindgom
barbara.mcgillivray@kcl.ac.uk

Malvina Nissim
CLCG, University of Groningen
The Netherlands
m.nissim@rug.nl

Abstract

We evaluate four count-based and predictive distributional semantic models of Ancient Greek against AGREE, a composite benchmark of human judgements, to assess their ability to retrieve semantic relatedness. On the basis of the observations deriving from the analysis of the results, we design a procedure for a larger-scale intrinsic evaluation of count-based and predictive language models, including syntactic embeddings. We also propose possible ways of exploiting the different layers of the whole AGREE benchmark (including both human- and machine-generated data) and different evaluation metrics.

1 Introduction

The application of Natural Language Processing to the study of Ancient Greek semantics is an emerging research area which has proven to be a fruitful avenue for our understanding of the Ancient Greek language and culture. Previous work has focused on the training of Distributional Semantic Models (DSMs) on Ancient Greek corpora (Boschetti, 2009; Rodda et al., 2017, 2019; McGillivray et al., 2019; Perrone et al., 2021a), a task enabled by the relatively large quantity of extant texts available for this language. DSM evaluation is a necessary step to properly assess the usefulness of applying these models to large-scale studies of Ancient Greek, but is made particularly challenging by the lack of native speakers and, compared to modern languages, a limited number of experts available.

This paper offers an evaluation of DSMs for Ancient Greek against the newly created AGREE benchmark (Stopponi et al., 2024b) and a road map for further, wider evaluation. We exploit the layered nature of AGREE to assess at different levels four DSMs, and discuss results not only in terms of model comparison, but mostly in terms of best evaluation strategies, suggesting various precision- and recall-based options. On that basis, in Section 6 we propose a road map for a more comprehensive evaluation campaign, which would involve training a wider range of models, including dependency-based embeddings (see, among others, Padó and Lapata 2007; Levy and Goldberg 2014; Lapesa and Evert 2017; Lenci et al. 2022), already preliminarily tested in Stopponi et al. (2024a), and studying their behaviour with respect to a number of metrics. Specifically, we propose to assess the difference in performance between syntactic embeddings trained on manually tagged and on automatically tagged treebanks. We plan to evaluate the DSMs, trained with different parameters, against the full version of AGREE, including both human- and machine-generated judgements. We also suggest alternative ways to use the data collected for AGREE and possible evaluation metrics.

2 Previous work

Few resources exist as gold standards for the evaluation of DSMs on Ancient Greek. Vatri and Lähteenoja (2019) contains the manual annotation of the senses of the lemmas μῦς, ἁρμονία, and κόσμος (Vatri and McGillivray, 2018) and was used in Perrone et al. (2021a) and Perrone et al. (2021b) to evaluate models for semantic change detection.

Rodda et al. (2019) evaluated count-based DSMs for Ancient Greek against benchmarks obtained

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 49–58,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_006

from an ancient lexicon, a modern dictionary of synonyms, and the computational lexicon *Ancient Greek WordNet* (Boschetti et al., 2016). The data they released represent the first benchmark for the evaluation of Ancient Greek DSMs.[1] Reusing preexisting resources, as they did, allows incorporating in the evaluation the semantic knowledge of real speakers of Ancient Greek (as in the case of the ancient lexicon) and to leverage the semantic knowledge of highly specialized experts, from resources that can be the product of years of work. This data collection seems less biased by the aims of the research, however it also has downsides. Lexical resources, compiled by humans, can suffer from idiosyncrasies, for example being biased by the interests and language taste of their author, and if the author is not alive anymore, it is not possible to get explanations about specific choices. Moreover, ancient resources can reflect ideas of semantic relationships between words (e.g. word similarity) that are different from the contemporary conceptualization, as also noticed by Rodda et al. (2019, 6–8) and discussed in Stopponi et al. (2024b).

3 Training Data for DSMs of Ancient Greek

The largest corpus of Ancient Greek, the Thesaurus Linguae Graecae (Pantelia, 2022), containing more than 110 million tokens,[2] is only accessible through the web interface. However, scholars can use a number of open-access machine-readable Ancient Greek corpora, containing different ranges of text types.[3] Some corpora are annotated, for example with lemma, POS, and syntactic information. The Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018), a portion of which was used as training data for the study presented in this paper, contains 10,206,421 automatically lemmatized and POS-tagged tokens. But many corpora with syntactic annotation also exist: an overview of the most often used treebanks for Ancient Greek is in Table 1.

As the case of GLAUx shows (see Table 1), automatic parsing allows for the creation of larger treebanks, even if the syntactic annotation is expected to be less accurate. We thus plan to train syntactic embeddings on two corpora, GLAUx and the largest possible manually-annotated treebank, created from a collation of the available corpora.

4 The AGREE Benchmark

The AGREE benchmark contains pairs of lemmas semantically related to 36 selected 'seed' lemmas (12 nouns, 12 adjectives, and 12 verbs), for a total of 638 lemma pairs.[4] The judgements were collected via questionnaires distributed to a large number (> 50) of academic scholars of Ancient Greek. The final benchmark, AGREE, incorporates a mix of expert-elicited pairs and expert-assessed, machine-generated pairs. The machine-generated items are pairs of [seed lemma - nearest neighbour], with nearest neighbours extracted from Word2Vec models (Mikolov et al., 2013) that underwent expert judgement and were assessed as highly related. For the experiments reported in this paper, we only use the human-elicited portion of the benchmark: *AGREE-task1*. This portion can be further divided into the subset of pairs that were proposed by one expert only, and the subset of pairs that were proposed by more than one annotator, under the assumptions that the latter might be cases of a stronger relatedness, and/or higher frequency.

5 Evaluation of DSMs of Ancient Greek

5.1 Procedure

For this study we evaluated two count-based and two predictive DSMs trained on a portion of the Diorisis corpus (Vatri and McGillivray, 2018), merging text from the Archaic, Classical and Hellenistic periods, since the AGREE benchmark (and especially the pairs proposed by experts) is particularly suited to the evaluation of models trained on texts from those periods (Stopponi et al., 2024b). The lemmatized version of Diorisis was used, to reduce the impact of word sparsity. Stop word filtering was performed, according to the list also used in Rodda et al. (2019)[5]. Stop word filtering reduced the size of the corpus from 5,768,916 to 2,960,459 tokens. The four models were evaluated against AGREE-task1, by comparing the top 5, 10, 15 (k) nearest neighbours of each of the 36 seed lemmas in the benchmark with the lemmas related to the same seed in AGREE-task1. The nearest neighbours extracted from the models were compared

[1] https://zenodo.org/record/3552763#.YfAItOrMKWA

[2] https://wiki.digitalclassicist.org/Thesaurus_Linguae_Graecae

[3] A review of most available open-access corpora for Ancient Greek is in Keersmaekers (2021, 40).

[4] https://zenodo.org/record/8027490.

[5] https://figshare.com/articles/dataset/Ancient_Greek_stop_words/9724613, by A. Vatri.

Treebank	N. tokens	Manual annotation	Texts
Ancient Greek Dependency Treebank (Perseus, Bamman and Crane, 2011)	ca. 550K*	yes	Literary, full list at http://perseusdl.github.io/treebank_data/
PROIEL Treebank (Haug and Jøhndal, 2008)	ca. 250.5K	yes	*The Greek New Testament*, *Histories* (Herodotus), *Chronicles* (Sphrantzes)
Gorman Trees (Gorman, 2020)	ca. 240K*	yes	Literary prose, full list at https://perseids-publications.github.io/gorman-trees/
Pedalion Trees (Keersmaekers et al., 2019)	ca. 300K	yes	Literary, full list at https://perseids-publications.github.io/pedalion-trees/
Harrington Treebanks (Harrington, 2018)	ca. 18K*	yes	*Nicene Creed*; *Book of Susanna* (Septuaginta), *Verae historiae* (Lucian of Samosata), *Vita Aesopi*
PapyGreek (Vierros and Henriksson, 2021)	ca. 44K	syntactic layer only	Papyri
Aphthonius (Yordanova, 2018)	ca. 7K*	yes	*Progymnasmata* (Aphtonius)
GLAUx corpus (Keersmaekers, 2021)	ca. 11,860K	no	Literary, papyrological, epigraphical. A sample was released at https://perseids-publications.github.io/glaux-trees/

Table 1: Some available treebanks for Ancient Greek. If the size of the treebank is followed by a *, it is taken from Keersmaekers et al. (2019, 110). The size of the PapyGreek treebanks has been calculated by summing up all the 'word' elements in the XML files.

to: all the lemmas in AGREE-task1, the lemmas in AGREE-task1 proposed by more than one expert, and the lemmas in AGREE-task1 proposed by only one expert. Precision and recall were adopted as evaluation metrics and defined as follows:

$$\text{Precision@K} = \frac{\text{overlap model's near. neighb. and benchmark}}{k}$$

$$\text{Recall@K} = \frac{\text{near. neighb. model also in benchmark}}{\text{n. related lemmas benchmark}}$$

5.2 Models

The models selected for evaluation are two Word2Vec models, one SGNS and one CBOW, and two count-based models. The matrices of the count-based models were weighted with PPMI and one one of the two dimensionality reduction was performed with Singular Value Decomposition (SVD). The two count-based models were built by using the software provided by the LSCDetection repository (Schlechtweg et al., 2019) with $window = 5$ and the following other parameters: $k = 1$ and $alpha = 0.75$ for PPMI, 300 dimensions and $gamma = 0.0$ for SVD. The two Word2Vec models were trained with the Gensim library (Řehůřek and Sojka, 2010) and the following parameters: $size = 30$, $window = 5$, $min\_count = 5$, $negative = 20$.

5.3 Results

The average precision and recall are reported in Table 2. We immediately see that recall is generally low. This can be explained by the fact that there are on average 14 neighbours per lemma[6] in AGREE-task1, so that the denominator in recall@k is generally larger than the numerator when $k = 5$ or $k = 10$. The recall consequently increases (on average) if k also increases, while the opposite happens for precision, which increases if k decreases. Taking into account recall for $k < 15$ makes thus little sense, since it is never possible to achieve full recall when the lemmas related to a certain seed in the benchmark are more than the extracted k-nearest neighbours. Conversely, it is theoretically possible to achieve 100% precision if all the extracted k-nearest neighbours are also in the benchmark. The higher precision with smaller values of k seems to confirm that the closest neighbours in the semantic space are actually more strictly related to the seed lemma, while the strength of the seed-

[6] Min. = 6, max. = 24, standard deviation = 4.43.

k	Precision	Recall
5	**0.20**	0.06
10	0.16	0.09
15	0.13	**0.11**

Table 2: Average precision and recall calculated against the whole AGREE-task1 benchmark and divided by k.

Model	Precision	Recall
SGNS	0.11	0.06
CBOW	0.15	0.08
SVD	0.16	0.09
PPMI	**0.22**	**0.12**

Table 3: Average precision and recall calculated against the whole AGREE-task1 benchmark, divided by model.

neighbour relationship declines for neighbours that are further away from the seed.

Model architecture also has an impact, with count-based performing better than predictive models. This is in line with what is observed by Lenci et al. (2022). Moreover, the model without dimensionality reduction performs better than the one to which SVD was applied, as shown in Table 3. Further, Word2Vec CBOW seems to perform better than Word2Vec SGNS. However, parameter optimization was not performed for this preliminary study, and a limited number of model architectures was tested. In future, larger evaluation will probably give a better picture of the differences between count-based and predictive models.

For example, for the seed lemma εἰρήνη, 'peace', there are 9 related lemmas in AGREE-task1: πόλεμος, 'war', σπονδή, 'drink-offering/treaty', ἥσυχος, 'quiet' (adj.), ἡσυχία, 'quiet, silence' (noun), σπένδω, 'make a drink-offering', μάχη, 'battle', γαληνός, 'calm', πολιτεία, 'citizenship', συγγραφή, 'writing', ὁμολογέω, 'agree', νίκη, 'victory', ὄλβος, 'happiness', γαλήνη, 'stillness', and φιλία, 'friendship'. Both the CBOW and the PPMI model have precision 0.2 with $k = 5$, i.e. among the first 5 nearest neighbours returned there is one that is also in AGREE-task1. The recall is 0.07 (1/14). The overlapping lemma is σπονδή, 'drink-offering/treaty' for the CBOW model, (which also returns as the other four nearest neighbours διάλυσις, 'separating/ending', συμμαχία, 'alliance', Λακεδαιμόνιος, 'Spartan', and πολεμέω, 'fight') and it is πόλεμος, 'war' for the PPMI, which also returns συμμαχία, 'alliance', Φίλιππος, 'Philip', πολεμέω, 'fight', and πρεσβεία, 'embassy'. We notice that both models return συμμαχία, 'alliance' among their first 5 neighbours. This word was not proposed by the experts in the first phase of data collection for the AGREE benchmark, but is however semantically related to εἰρήνη, 'peace'. More in general, we deem all the top 5 nearest neighbours returned by both models as acceptable results, since they all are related to εἰρήνη, 'peace'; the two models just differ in results from one other, as well as from the benchmark. Of course, there are also cases in which the overlapping lemma(s) are the same between models. One example is μέγας, 'big', for which there are 15 related lemmas in AGREE-task1.[7] Both the CBOW and the PPMI model have precision 0.2 (1/5) and recall 0.07 (1/14) with k = 5, and the lemma overlapping with the AGREE-task1 benchmark is the same for both models, μέγεθος, 'greatness'. Again, the extracted nearest neighbours that are not in the benchmark are not necessarily unrelated to the seed μέγας, 'big'. The CBOW model also returns τηλικοῦτος, 'of such an age/so large', ἄξιος, 'weighing as much/worthy', ῥοπή, 'weight', and ὑπερβάλλω, 'surpass/exceed', while the PPMI model also returns ἐλάσσων, 'smaller', ἴσος, 'equal', ἄρος, 'use/profit', and πολύς, 'many'. Except from ἄρος, 'use/profit', they all relate to μέγας, even if, intuitively,with a different strength and with different types semantic relations.

The internal layering of the benchmark AGREE-task1, which accounts for the number of experts who proposed a specific lemma, allows for other observations (Table 4). On average, the lemmas returned by only one expert (*AGREE-task1-only1* in 4) are more (13.02 per seed lemma) than those returned by several experts (*AGREE-task1-more1*, 4.69 per seed). We could hypothesize that the relatedness among the latter may be stronger or more evident, since more than one expert independently had proposed the same lemmas as related to the relevant seed word. When we evaluate against lemma pairs proposed by more than one expert higher pre-

[7] They are μικρός, 'small', ὅρκος, 'oath', βασιλεύς, 'king', θαῦμα, 'wonder', θεός, 'god', μακρός, 'long', ὀλίγος, 'little', βραχύς, 'short', μέγεθος, 'greatness', αὐξάνω, 'increase', μεγαλοψυχία, 'greatness of soul', ἥρως, 'hero', γίγας, 'giant', καλός, 'beautiful', and μεγαλοφροσύνη, 'greatness of mind'.

Benchmark subset	Prec	Rec
AGREE-task1	**0.16**	**0.09**
AGREE-task1-more1	0.09	0.05
AGREE-task1-only1	0.07	0.04

Table 4: Average precision and recall calculated against different subsets of the AGREE-task1 benchmark. The results with the three values of k were averaged.

cision and recall scores are observed, possibly suggesting that pairs proposed by more experts are more closely related to their seed lemma, and possibly more frequent. This is particularly true for the PPMI model, which achieves an average of 0.22, 0.14, and 0.09 precision, and an average of 0.12, 0.07, and 0.05 recall against, respectively, the whole AGREE-task1, the pairs proposed by more than one expert, and the pairs proposed by only one expert (the results are averaged across the three values of k). This is observed when averaging the results of all models, but it does not necessarily hold for each model. The CBOW model, for example, achieves a higher precision against the set of pairs proposed by only one expert than against those proposed by more experts. Both Word2Vec models instead achieve the same precision and recall on both subsets of AGREE-task1. The results discussed until now are summarised in Table 5.

Another dimension of the benchmark is the part-of-speech (POS) of the seed lemmas. In Table 6 we see that evaluating against pairs including an adjective seed lemma the highest precision is achieved, followed by noun seeds and verb seeds. The recall is higher when evaluated against pairs including adjective or noun seeds. However, the differences in precision and recall are very small.

Finally, dividing the results by lemma reveals a great variety in precision and recall among the different lemmas. For example, with $k = 5$ the highest precision is achieved. The average precision per lemma calculated against the whole AGREE-task1 is 0.20, with standard deviation 0.16. There is indeed a large variability between the average precision against the "best" and the "worst-performing" lemmas. Those yielding the highest precision are some nouns and adjectives: ἅρμα, 'chariot', average precision 0.6; ψευδής, 'false', 0.55; ἐλεύθερος, 'free', 0.45; πατήρ, 'father', 0.45; and ἄγριος, 'wild', 0.45. However, they are immediately followed by verbs, ἔρχομαι, 'go' and ὁράω, 'see', both with average precision 0.4. The lowest precision, 0, is achieved with the seed lemmas ἀκτή, 'headland', κλυτός, 'renowned', ναίω, 'dwell', ῥῆσις, 'speech', σῆμα, 'sign/mark', and τεύχω, 'make/build', all with average precision 0. Nevertheless, as we already observed, a low precision does not necessarily correspond to bad results (i.e. unrelated lemmas), even if it is true that some of the nearest neighbours returned by the models to these are unrelated or intuitively less strictly related to the seed lemmas. Moreover, a higher precision seems to correspond to higher-frequency words, while the lemmas yielding the lowest precision also have a low frequency in the corpus.[8] In Table 7 the average precision and recall for each lemma are reported, calculated against the whole AGREE-task1 and with $k = 15$. Note that changing the value of k the order of the seed lemmas, ranked by precision, also changes.

6 Road Map for Future Work

We plan a larger evaluation including more model architectures, different parameters and different evaluation metrics, with the aim of understanding the differences between model types, rather than finding the 'best' model (see also Lenci et al., 2022), and evaluation adequacy. More investigation is needed to understand whether the difference between count-based and predictive models trained on Ancient Greek lies in the quality of results (i.e., if some architectures actually return less relevant nearest neighbours), or only in the kind of relationships they capture. Further experiments will also concern dependency-based embeddings.

Moreover, this extended study will exploit the full dataset produced for the AGREE benchmark, including the second part of the dataset, not used for the current evaluation. Since in the second phase of the data collection the experts assigned relatedness scores to human- and machine- generated lemma pairs, these items items allows ranking the lemma pairs according to their degree of relatedness, and thus for a more nuanced evaluation.

[8] The frequency in the subcorpus of the mentioned "best performing" lemmas is: ἅρμα: 541, ψευδής: 1048, ἐλεύθερος: 940, πατήρ: 5685, ἄγριος: 348, ἔρχομαι: 5251, ὁράω: 4987, while the frequency of the mentioned "worst-performing lemmas is: ἀκτή: 177, κλυτός: 142, ναίω: 283, ῥῆσις: 48, σῆμα: 213, τεύχω: 255.

		Precision				Recall						
Bench. subset	k	PPMI	SVD	CBOW	SGNS	PPMI	SVD	CBOW	SGNS	Tot. prec.	Tot. rec.	Tot. pairs
AGREE-task1	all k	0.22	0.16	0.15	0.11	0.12	0.09	0.08	0.06	0.16	0.09	638
	k = 5	0.28	0.19	0.19	0.14	0.08	0.06	0.05	0.04	0.20	0.06	
	k = 10	0.22	0.16	0.14	0.11	0.12	0.09	0.08	0.07	0.16	0.09	
	k = 15	0.17	0.13	0.11	0.09	0.15	0.11	0.10	0.08	0.13	0.11	
AGREE-task1-more1	all k	0.14	0.09	0.07	0.06	0.07	0.05	0.04	0.03	0.09	0.05	169
	k = 5	0.19	0.11	0.08	0.07	0.06	0.03	0.02	0.02	0.11	0.03	
	k = 10	0.13	0.10	0.06	0.05	0.08	0.06	0.04	0.03	0.09	0.05	
	k = 15	0.10	0.08	0.06	0.04	0.09	0.07	0.05	0.04	0.07	0.06	
AGREE-task1-only1	all k	0.09	0.07	0.08	0.06	0.05	0.04	0.04	0.03	0.07	0.04	469
	k = 5	0.09	0.08	0.11	0.07	0.03	0.02	0.03	0.02	0.09	0.02	
	k = 10	0.09	0.07	0.08	0.06	0.05	0.04	0.04	0.03	0.07	0.04	
	k = 15	0.07	0.06	0.06	0.04	0.06	0.05	0.05	0.04	0.06	0.05	

Table 5: Average precision and recall calculated against different subsets of the AGREE-task1 benchmark, divided by model type and by k. The recall for values of k lower than 15 has been reported for completeness, but it has limited usefulness (see above). The column 'Tot. pairs' contains the total number of pairs in the relevant subsets.

POS	Precision	Recall
A	0.18	0.09
N	0.15	0.09
V	0.15	0.08

Table 6: Average precision and recall calculated against the whole AGREE-task1 benchmark and divided by POS of the seed lemmas.

6.1 Models

We will test a selection of popular DSMs belonging to the first two generations defined by Lenci et al. (2022), i.e. count-based models (PPMI and GloVe) and predictive models (Word2Vec and FastText). In particular, we will test:

1. two count-based models trained by using Positive Pointwise Mutual Information (PPMI) as association measure,[9] with and without dimensionality reduction with the Singular Value Decomposition (SVD);
2. GloVe (Pennington et al., 2014));
3. FastText (Bojanowski et al., 2017);
4. the two architectures of word2vec (Mikolov et al., 2013), the Skip-gram with Negative Sampling (SGNS) and the Continuous-Bag-of-Words (CBOW);
5. two 'syntax-filtered' models (Padó and Lapata, 2007; Lapesa and Evert, 2017; Lenci et al., 2022), a SGNS one but using direct dependency between tokens to extract co-occurrences rather than mere token windows and one trained using the SuperGraph approach described in Al-Ghezi and Kurimo (2020). The latter method consists in using dependency relations between tokens to generate graph structures for every sentence in a treebank, before merging all graphs into one SuperGraph. The SuperGraph then serves as input to Node2Vec (Grover and Leskovec, 2016), a modification of the SGNS architecture which enables the training of word representations starting from nodes in a graph.

Contextual models will not be included, instead. Even if some work exists on the training of contextual models of Ancient Greek (Singh et al., 2021; Keersmaekers and Mercelis, 2021; Yamshchikov et al., 2022; Riemenschneider and Frank, 2023) (despite the fact that contextual models require huge quantities of training data (Lenci et al., 2022, 1274)), the only existing evaluation datasets for semantic models of Ancient Greek (Rodda et al., 2019 and Stopponi et al., 2024b) were created

[9] About association measures, see Evert et al. (2008).

Lemma	Precision	Recall	Lemma	Precision	Recall
ἅρμα, 'chariot'	0.32	0.22	εἰρήνη, 'peace'	0.10	0.11
ὁράω	0.30	0.24	Ἀθηναῖος, 'Athenian'	0.08	0.08
ναῦς, 'ship'	0.27	0.25	νόστος, 'return'	0.08	0.07
χρυσός, 'gold'	0.27	0.27	παλαιός, 'old'	0.08	0.07
ἄγριος, 'wild'	0.23	0.17	ζεύγνυμι, 'yoke'	0.08	0.09
ἐλεύθερος, 'free'	0.23	0.17	μέγας, 'big'	0.08	0.08
ἔρχομαι, 'go'	0.23	0.19	μῦθος, 'word/story'	0.07	0.07
πατήρ, 'father'	0.22	0.30	ἀκτή, 'headland'	0.07	0.07
ψευδής, 'false'	0.20	0.18	μοχθέω, 'labour'	0.07	0.07
κακός, 'bad'	0.17	0.12	Σάμος, 'Samos'	0.07	0.06
οἰκέω, 'inhabit'	0.17	0.11	ἄλκιμος, 'brave'	0.05	0.04
αὐξάνω, 'increase'	0.17	0.14	ῥῆσις, 'speech'	0.05	0.04
ὀρφανός, 'orphan'	0.17	0.14	τέμνω, 'cut'	0.03	0.03
πόντος, 'sea'	0.15	0.12	κλυτός, 'renowned'	0.02	0.01
φιλέω, 'love'	0.15	0.15	λείπω, 'leave/quit'	0.02	0.01
αἴθω, 'light up'	0.13	0.10	τεύχω, 'make/build'	0.02	0.01
πρέσβυς, 'old man, elder'	0.13	0.11	ναίω, 'dwell'	0.00	0.00
ἑνδέκατος, 'eleventh'	0.13	0.11	σῆμα, 'sign/mark'	0.00	0.00

Table 7: Average precision and recall calculated against the whole AGREE-task1 benchmark and with $k = 15$, divided by seed lemma. The lemmas are ranked by average precision.

for the evaluation of static (type-based) embeddings. Although type-based embeddings can be obtained from contextualized token embeddings, e.g. by averaging the model representations of each word (see the discussion in Lenci et al., 2022, 1290–1291), their superiority over type embeddings obtained from static DSMs has been questioned (Lenci et al., 2022, 1289–1293). This evaluation will thus be limited to the evaluation of static embeddings, leaving the training and evaluation of contextual embeddings for future work.[10] All the models will be trained with two different context windows, e.g. 5 and 10. According to the large-scale evaluation of Lenci et al. (2022), model architecture and context window size are the two parameters that significantly affect model performance (especially model architecture). We thus concentrate on testing of these two.

[10]It should be noted that the training corpus of Lenci et al. (2022, 1279) are English texts from the Web. Their conclusions could thus not entirely apply to Ancient Greek, a language with a different syntax and morphology.

6.2 Dependency-based embeddings

Ancient Greek syntactic embeddings obtained with the SuperGraph method have already been compared with window-based models by Stopponi et al. (2024a), clearly suggesting that the former capture functional rather than topical similarity, as had already been shown at least since Levy and Goldberg (2014) on the basis of English models. Given this ontological difference between the two, an open question, then, is whether syntactic embeddings should be evaluated on a par with traditional count-based and word2vec models, namely whether there are arguments for using the same benchmark to judge the quality of models regardless of whether syntactic information is integrated in their training or not. Previous large-scale comparisons of dependency-based and window-based DSMs suggested that the latter, when fine-tuned, generally outperform the former in most downstream tasks (Kiela and Clark, 2014; Lapesa and Evert, 2017). Given the generally greater computational costs associated with dependency parsing and the extraction of syntactic collocates (i.e. tokens with a direct dependency

relation), it has been questioned whether the training of dependency-based embeddings is justifiable after all. However, there is evidence, at least as far as high-resource languages such as English are concerned, that dependency-based embeddings outperform window-based models in a limited but coherent number of tasks. This has been shown to be consistently the case, for instance, of categorization tasks, namely grouping lexical items into semantically coherent categories (Rothenhäusler and Schütze, 2009; Lapesa and Evert, 2017; Lenci et al., 2022), as well as thematic fit estimation, namely evaluating the typicality of the argument of a verb given a thematic role (e.g., agent or patient) (Baroni and Lenci, 2010; Chersoni et al., 2017). Different tasks such as categorization and synonymity tests present, in many ways, the same ontological differences occurring between dependency- and window-based models as a whole. This alone would seem to warrant the training of different models (and, as a result, the development of different evaluation methods) depending on the task at hand. Classic distributional semantic models (i.e. window-based) are generally fined-tuned to capture attributional similarity (Turney, 2006), namely the number of attributes, or properties, shared by the referents of two given words. As pointed out by Baroni and Lenci (2010), words that share many collocates will show a high attributional similarity since common collocates can be seen as a proxy for some of the attributes that the two words denote. Pairs such as *dog-puppy* will then have a high attributional similarity but not necessarily a high relational similarity (Turney, 2006), which in turns refers to sharing similar semantic relations to their nearest neighbours. In Baroni and Lenci's 2010 example, the pair *dog-tail* will be more similar to *car-wheel* than it is to *dog-animal*, even though attributionally that is clearly not the case.

Building on the preliminary observation made in Stopponi et al. (2024a) about the relational, rather than attributional, similarity captured by Ancient Greek dependency-based models, we thus plan to test different Ancient Greek models on different tasks depending on the kind of similarity the model is trained to capture. Categorization and thematic fit task, for example, can be set up with the help of the richly annotated resources for the language (e.g. the verbal semantic annotation in the PROIEL treebank) for dependency-based models, in addition to similarity judgement tasks, which may be instead better suited to evaluate window-based DSMs.

6.3 Evaluation Metrics

We observed above how precision and recall only provide an absolute evaluation against the benchmark, capturing whether the words in the benchmark are returned by the models or not, but they do not allow us to take into account the strength of the semantic relationship between lemmas. Moreover, only the first k neighbours returned by the model are evaluated, while there is no information about how close to the seed lemma in a semantic space the related lemmas in the benchmark are which are not among the first k neighbours. Furthermore, the use of recall in this kind of evaluation can be problematic when the number of k is lower than the number of pairs in the benchmark.

To overcome these limitations, we plan to include additional evaluation strategies. One option is to use the evaluation items that were rated on a 0-100 relatedness scale (AGREE-task2), to calculate for each seed lemma the correlation between: (i) the scores assigned to pairs including that lemma in the benchmark; (ii) the cosine distances between the same word pairs in a semantic space. The scores can also be used to *rank* the items, and a correlation can be calculated between ranks and cosine distances. Taking into account degrees of relatedness may be a more adequate way to evaluate models on a phenomenon such as semantic relatedness.

Another possibility is to exploit the information about the number of raters who proposed the words collected in the first phase (AGREE-task1), for example by giving greater weight to pairs suggested by multiple raters. However, this will first require a deeper investigation on the nature of the pairs proposed by one versus several experts, and the impact this might have on evaluation. Relatedly, frequency should also be considered to verify the ways and extent to which precision and recall are impacted by high-frequency items (both the human-elicited ones and those returned by the models).

7 Conclusion

We presented and discussed the results of an evaluation of four Distributional Semantic Models of Ancient Greek, two count-based and two predictive models. The gold standard was a subset of the AGREE benchmark, AGREE-task1, including pairs of related lemmas proposed by experts of Ancient Greek. The evaluation showed that count-

based models achieved higher precision and recall on AGREE-task1, and higher precision and recall were also achieved on average when evaluating against pairs of related lemmas proposed by more than one expert. Another important finding was the great difference in performance between different lemmas. We also presented a plan for a more extended evaluation, including more model architectures, parameters, and evaluation metrics. This evaluation will take into account different degrees of relatedness between lemmas and allow for a better understanding of the differences between DSMs of Ancient Greek and of the possible impact of such differences on computational studies in Ancient Greek lexical semantics.

Acknowledgements

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

References

Ragheb Al-Ghezi and Mikko Kurimo. 2020. Graph-based syntactic word embeddings. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*, pages 79–98. Springer.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

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

Federico Boschetti. 2009. *A Corpus-based Approach to Philological Issues*. University of Trento.

Federico Boschetti, Riccardo Del Gratta, and Harry Diakoff. 2016. Open Ancient Greek WordNet 0.5. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics 'A. Zampolli', National Research Council, in Pisa. http://hdl.handle.net/20.500.11752/ILC-56 (accessed 4 July 2022).

Emmanuele Chersoni, Enrico Santus, Philippe Blache, and Alessandro Lenci. 2017. Is structure necessary for modeling argument expectations in distributional semantics? In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Stefan Evert et al. 2008. Corpora and collocations. *Corpus linguistics. An international handbook*, 2:1212–1248.

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

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *CoRR*, abs/1607.00653.

Matthew Harrington. 2018. Perseids project. treebanked commentaries at Tufts University.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

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

Alek Keersmaekers and Wouter Mercelis. 2021. Improving morphological analysis of Greek with transformer-based approaches: First results with ELECTRA.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, France. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30, Gothenburg, Sweden. Association for Computational Linguistics.

Gabriella Lapesa and Stefan Evert. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 394–400, Valencia, Spain. Association for Computational Linguistics.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, 56(4):1269–1313.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Maria C Pantelia. 2022. *Thesaurus Linguae Graecae Digital Library*. University of California, Irvine.

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

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. 2021a. Lexical semantic change for Ancient Greek and Latin. In N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, and S. Hengchen, editors, *Computational approaches to semantic change*, pages 287–310. Language Science Press.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021b. Lexical semantic change for Ancient Greek and Latin: Computational approaches to semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational approaches to semantic change*, volume 6, chapter 9, pages 287–310. Language Science Press.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. *arXiv preprint arXiv:2305.13698*.

Martina A Rodda, Philomen Probert, and Barbara McGillivray. 2019. Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique Des Langues*, 60(3):63–87.

Martina A Rodda, 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):11–24.

Klaus Rothenhäusler and Hinrich Schütze. 2009. Unsupervised classification with dependency based word spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 17–24, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021*, pages 128–137. Association for Computational Linguistics.

Silvia Stopponi, Nilo Pedrazzini, Saskia Peels-Matthey, Barbara McGillivray, and Malvina Nissim. 2024a. Natural language processing for Ancient Greek: Design, advantages, and challenges of language models. *Diachronica*.

Silvia Stopponi, Saskia Peels-Matthey, and Malvina Nissim. 2024b. AGREE: A new benchmark for the evaluation of distributional semantic models of Ancient Greek. *Digital Scholarship in the Humanities*.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

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

Alessandro Vatri and Viivi Lähteenoja. 2019. Ancient Greek semantic annotation datasets.

Marja Vierros and Erik Henriksson. 2021. PapyGreek Treebanks: A dataset of linguistically annotated Greek documentary papyri. *Journal of open humanities data*.

Ivan Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Polina Yordanova. 2018. Treebank of Aphtonius, Progymnasmata.