Vector space models of Ancient Greek word meaning, and a case study on Homer

Martina Astrid Rodda[*, **] — **Philomen Probert**[*] — **Barbara McGillivray**[**, ***]

[*] *University of Oxford*
(martinaastrid.rodda@jesus.ox.ac.uk; philomen.probert@wolfson.ox.ac.uk)
[**] *The Alan Turing Institute*
[***] *University of Cambridge (bm517@cam.ac.uk)*

ABSTRACT. *Our paper describes the creation and evaluation of a Distributional Semantics model of ancient Greek. We developed a vector space model where every word is represented by a vector which encodes information about its linguistic context(s). We validate different vector space models by testing their output against benchmarks obtained from scholarship from the ancient world, modern lexicography, and an NLP resource. Finally, to show how the model can be applied to a research task, we provide the example of a small-scale study of semantic variation in epic formulae, recurring units with limited linguistic flexibility.*

RÉSUMÉ. *Notre article démontre à la fois la création et l'évaluation d'un modèle de sémantique distributionnelle du grec ancien. Tout d'abord nous avons développé un modèle d'espace vectoriel où chaque mot est représenté par un vecteur qui codifie les informations qui concernent ses contextes linguistiques. Ensuite nous avons validé différents modèles d'espace vectoriel en testant leur output par rapport à des références obtenues à partir de trois sources: un savant de l'Antiquité, la lexicographie moderne et la ressource WordNet. Enfin, en vue de démontrer comment le modèle peut être appliqué à une activité de recherche, nous fournissons une étude de cas, à petite échelle, de la variation sémantique dans les formules épiques, à savoir les unités recurrentes qui ont une flexibilité linguistique limitée.*

KEYWORDS: *Distributional Semantic Models, Diorisis Ancient Greek corpus, vector-space models, evaluation of distributional resources, ancient Greek epic poetry, formulaic language, semantic variation.*

MOTS-CLÉS : *Modèles de sémantique distributionnelle, Diorisis Ancient Greek corpus, espaces vectoriels, évaluation des ressources distributionelles, poésie épique du grec ancien, formules linguistiques, variation sémantique.*

1. Introduction: the broader research question

This paper presents the creation and evaluation of a computational model that was developed in the context of a broader study of linguistic variation in ancient Greek epic formulae. As the requirements of this research question shape the entirety of the study presented here, we shall begin by outlining some details about the archaic Greek epic tradition and its language. Archaic Greek epic is the product of an oral tradition of which the Homeric poems are the main remnant. Poems about heroes and gods were composed orally and performed by skilled singers in front of an audience. A singer's abilities lay equally in their knowledge of the heroic myths and in their ability to use traditional language according to and beyond the expectations of their listeners.[1]

Throughout this tradition, devices were developed for easier composition and understanding; the most important of these is formulaic language. Formulae are recurring linguistic units which allow for limited flexibility in structure and meaning. They range from noun-epithet pairs (*swift-footed Achilles*, *golden Aphrodite*) to more complex phrases, e.g. speech introductions (*and to him/her spoke...*). The latter often include open slots that need to be filled with additional material: a speech introduction will almost inevitably contain a reference to the name of the speaker (*and to him/her spoke golden Aphrodite*) and/or to some attendant circumstances (*and to him/her Aphrodite spoke in reply*).

While these open slots do not come in a fully pre-defined shape, as opposed to the less flexible parts of the formula itself, restrictions on their flexibility still apply: for instance, in the formulaic structure discussed above, it is possible to address someone *in reply* or *in anger* or *with a smile*, but not *with winged words*, a famous phrase that can only be used in combination with other types of speech introduction formulae.[2]

Formulaic behaviour has been compared to that of idioms and other linguistic constructions characterised by limited syntactic flexibility (Kiparsky 1976; Bozzone 2014; Antović and Cánovas 2016): both formulae and multi-word expressions in everyday language are restricted in their patterns of variation and change. Meaning change and flexibility, in particular, play an important role in the evolution of language usage. Research has highlighted how semantic flexibility promotes the productivity of constructions with open slots in modern languages: the range of different meanings a construction can accommodate in its open slots has an effect on whether it survives and spreads through time (Barðdal, 2008; Perek, 2016). As the behaviour of formulae is similar to that of linguistic constructions, we can expect the semantic openness of a formula to also influence the vitality of its usage through time; however, this mechanism has never been studied in early Greek epic until now.

1. Cf. e.g. Foley (1999). For a discussion of formulaic language in a usage-based linguistic perspective, see Kahane (2018).

2. The metre of archaic Greek epic (hexameters) also plays an important role in these restrictions, a circumstance that this paper will mostly ignore due to its focus on semantics.

Our paper lays the groundwork for an approach to ancient Greek epic formulae that can take into account the role of semantic flexibility in their behaviour and evolution. To do so, we present the first computational model which uses Distributional Semantics to assess the scope of linguistic variation in formulaic language. As this paper's main focus is on the technical requirements and optimisation of the model, we will only show a simple example of its application, by sketching out an analysis of the semantic flexibility of a small group of transitive-verb formulae. Further work will explore the links between semantic flexibility and diachronic variation.

1.1. *Motivation and relationship to previous work*

Distributional Semantics offers a quantitative approach to the study of semantic flexibility and represents an especially promising tool in Historical Linguistics, where no native speaker input can be sought. The meaning of a word is defined in a distributional perspective as a function of its collocates in a corpus: words that share a linguistic context are also related in meaning (Harris, 1954; Fabre and Lenci, 2015). Distributional Semantics naturally presents itself as a valuable method to assess the range of meaning flexibility in formulaic language: it provides a way to combine and process a large quantity of information and to make judgements that do not rely on the intuition of the researcher, as well as being quantifiable to a very fine-grained level.[3] In applying Distributional Semantics to the study of ancient languages, we build on previous work on computational semantics in ancient Greek, which applied similar methods to questions such as semantic drift, the semantics of verbs of motion, and polysemy (Rodda *et al.*, 2017; Grewcock, 2018; McGillivray *et al.*, 2019). Our main interest lies in creating a computational method that will carry useful information for scholars in the humanities, beyond the purposes of the current project.

For these reasons, we focus at length on the validation of the model and the fine-tuning of parameters relative to the pre-processing of the input (see below, 2.2). We will also briefly address how the configuration of parameters relates to existing literature on the evaluation of Distributional Semantics Models (DSMs) (Bullinaria and Levy, 2007; Bullinaria and Levy, 2012; Tanguy *et al.*, 2015). However, our priority is not achieving high absolute rates of accuracy compared to the benchmark sets, but establishing which combination of parameters is most accurate and useful for the purposes of this Digital Humanities study. While the fine-tuning of these hyperparameters is not fully portable across corpora, our article also shows a step-by-step breakdown of the evaluation process, which can be applied to different datasets. Moreover, due to the size of the corpus used (the largest freely available lemmatised corpus of ancient Greek literature), the evaluation provided in this article is significant for research on ancient Greek in general.

3. On these requirements in general, see Jenset and McGillivray (2017). On the potential of digital humanities in addressing traditional philological tasks, see Boschetti (2018).

2. Method

2.1. *Corpus*

Data for this study was gathered from the Diorisis Ancient Greek corpus (Vatri and McGillivray, 2018) (`https://www.doi.org/10.6084/m9.figshare.6187256`). This corpus comprises 820 literary Greek texts, spanning chronologically from Homer (8th century BC?) to the 5th century CE, for a total of over 10 million words. The texts are mainly sourced from the Perseus Canonical Greek Literature repository (along with The Little Sailing and Bibliotheca Augustana digital libraries); each text is fully lemmatised using a custom-built dictionary, and PoS-tagged with TreeTagger (Schmid, 1994) trained on the Ancient Greek Dependency Treebank (Celano, 2014) and the Ancient Greek portion of the PROIEL treebank (Haug and Jøhndal, 2008).

The Diorisis corpus is fully lemmatised; all vector space models discussed here were built on the lemmas, not inflected words. We decided to include the entirety of the material available, without chronological filtering. This decision was made both to compensate for the relatively small size of the corpus itself (in comparison with modern language corpora), and on account of the nature of the "gold-standard" datasets that will be introduced in section 3.1.

2.2. *Corpus processing*

Data was extracted through a Python script written for this purpose, available at `https://zenodo.org/badge/latestdoi/174973156`. The script extracts the collocates for each word in the corpus, i.e. the words that occur together with the target word, within a window of co-occurrence defined by the user; for our study we used windows of 1, 5, and 10 words on both sides of the target word (not including the target itself). Context windows are sensitive to sentence boundaries: we only included words in the same sentence as the target word. Sentence boundaries are already encoded in the Diorisis corpus, and are defined according to ancient Greek punctuation rules: a full stop, semicolon (used as a question mark), or middle dot (equivalent to a colon or semicolon) all end a sentence.

A frequency threshold, applied over the whole corpus, can also be set by the user; we built our semantic spaces respectively on words that occur at least 100 times in the corpus, at least 50 times, at least 20 times, and finally with no frequency threshold at all (including everything down to *hapax legomena*). All spaces were filtered for stop-words, i.e. words that perform a primarily syntactic function, co-occurring with other words independently of their semantic properties, and are therefore considered irrelevant to semantic modelling. The stop-words are not used as either targets or contexts.[4] On the basis of their study of large English-language corpora, Bullinaria and

4. The list of stop-words, compiled by Alessandro Vatri based on the Perseus Hopper source, is available at `https://figshare.com/articles/Ancient_Greek_stop_words/9724613`.

Levy (2012) argue that stop-word filtering does not significantly improve or reduce the quality of the results, but it does reduce the dimensionality of the spaces, which is desirable for computational reasons.

The resulting combinations of parameters leads to 12 different semantic spaces to be assessed (3 window sizes x 4 frequency thresholds). It will be interesting, albeit tangential to the scope of the present study, to note how the behaviour of each of these spaces compares to literature on English-language corpora: in particular, Bullinaria and Levy (2007; 2012) argue that the best results are achieved with the smallest possible window size, i.e. 1, and without filtering for frequency. The authors, however, are working on minimally processed corpora (stripped of sentence boundaries and unlemmatised); this is a very different situation from the richly annotated Diorisis corpus.

2.3. *Semantic spaces*

The vector space models used in this article were built using the DISSECT toolkit (Dinu *et al.*, 2013). To use the terminology of Baroni *et al.* (2014), DISSECT is a count model, not a predictive model (neural network model). While predictive models such as word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) perform at least as well or better than count models on larger corpora (Baroni *et al.*, 2014), the very limited size of our ancient Greek corpus (10 million tokens vs. 2.8 billion tokens in Baroni *et al.* [2014], 6 billion tokens in Mikolov *et al.* [2013], 1 to 42 billion tokens in Pennington *et al.* [2014]) would create a problem for a predictive model, as an even smaller training corpus would need to be extracted, dramatically reducing the size of the available material.[5] Comparison with results obtained via word embeddings could be an interesting avenue for future work, but is not within the aims of the present study.

We used Positive Pointwise Mutual Information (PPMI) as our association measure (Evert, 2008), and applied Singular Value Decomposition (SVD) to reduce the resulting matrices to 300 latent dimensions. Dimensionality reduction is widely recognised to improve the accuracy of vector space models by reducing noise and highlighting the contribution of the most significant dimensions (Landauer and Dumais, 1997). While Bullinaria and Levy (2012) show that models with several thousands of dimensions appear to perform better on large, unprocessed corpora, there is no reason to think this should be the case for a small, information-rich corpus such as Diorisis.

One vector space model was generated for each combination of parameters as detailed in 2.2: `w1_t1`, `w1_t20`, `w1_t50`, `w1_t100`, `w5_t1`, `w5_t20`, `w5_t50`, `w5_t100`, `w10_t1`, `w10_t20`, `w10_t50`, `w10_t100`, with `w` referring to the size of the context window on each side of the target word and `t` to the frequency threshold.

5. Pennington *et al.* (2014) provide data on how GloVe performance scales up with corpus size.

3. Evaluation

As opposed to previous contributions on distributional semantics applied to ancient Greek (Rodda *et al.*, 2017; McGillivray *et al.*, 2019), we decided to assess the performance of the vector space models not by manually screening a sample of the results, but by comparing them to lists of synonyms obtained from three separate sources:[6] ancient scholarship (the *Onomasticon* by Julius Pollux, a Greek scholar of the 2nd century CE), modern lexicography (a 19th-century etymological dictionary of Greek), and a Natural Language Processing resource, the Open Ancient Greek WordNet (Boschetti *et al.*, 2016). Each resource is described in detail below.

Using pre-existing resources instead of individuals' *post-hoc* judgements allows us to provide a robust assessment based on independent data. The characteristics of each source limit the absolute rate of accuracy (see 3.2), but they still allow us to compare the relative accuracy of different Distributional Semantics Models. While work on English routinely uses independent benchmarks, such as the TOEFL performance data (Landauer and Dumais, 1997; Bullinaria and Levy, 2007; Bullinaria and Levy, 2012), handcraft thesaurus (Curran, 2004), and various other benchmark resources (Baroni *et al.*, 2014), to show how the ability of a DSM to identify synonyms compares to human performance on the same task, to the best of our knowledge this is the first time that such a comparison with pre-existing data has been attempted for ancient Greek. In particular, it is tricky to come up with existing lists of synonym pairs, as no ready-made resource of the kind used for English is available. Moreover, the lexicographical resources available for ancient Greek are particularly idiosyncratic, due to the lack of direct input from native speakers. This lack can be partially bridged by resorting to lexicographical works from the ancient world; one of these, Pollux' *Onomasticon*, was used in this article. Due to its unusual characteristics for a resource used in computational linguistics, it deserves more space than the other two resources used in this article, and will be discussed at length in section 3.1.1.

All of our sources for comparison have some level of diachronic depth. This provides another reason, in addition to size considerations, to avoid dividing the Diorisis corpus into periods (see above, 2.1). Pollux' *Onomasticon*, composed in the late second century CE, gives us a wealth of information on Greek vocabulary, with particular but by no means exclusive emphasis on words used by classical authors (see further section 3.1). Somewhat similarly, Schmidt's dictionary of synonyms (Schmidt, 1876-1886) privileges classical and Homeric usage but also takes Hellenistic sources into account, while the Open Ancient Greek WordNet includes data from dictionaries based on texts from a wide chronological span. While the level of diachronic depth varies from source to source, none of them provides a synchronic snapshot for any specific period, and there is no reason to try to achieve this in the corpus.

6. While synonymy and semantic relatedness are only partially overlapping concepts, as semantic relatedness has a broader scope (Levy and Goldberg, 2014), our work on lists of synonyms extracted from independent sources matches standard practice for English-language corpora.

3.1. *Gold-standard sets*

3.1.1. *Ancient lexicography*

For any semantic model to be worthwhile, it ought (as suggested already) to make some predictions that can be tested against something other than itself. For example, it might be expected to predict—to some degree of accuracy better than random guessing—whether native speakers will say that a particular pair of words is semantically similar or not. For a modern language, one way to evaluate a semantic space model is therefore to ask native speakers for some input: how well do the model's predictions stand up to testing against their intuitions? As mentioned in section 1.1, we cannot directly involve native speakers in a study of an ancient language. Yet numerous lexicographical works survive from the ancient Greek-speaking world, and it is worth considering what we can learn from these.

In an ideal world we would identify an ancient lexicographer whose goals were somewhat similar to ours: someone interested in telling us which words are semantically very close to one another, which words lie just slightly further away, and so on. A remarkable ancient work that in some respects does just this is the *Onomasticon* of Julius Pollux, who held the Chair of Rhetoric at Athens in the late second century CE (see Bethe [1917]; Dickey [2007, 96]; Vessella [2018, 24–5]).

Pollux announces in the Preface to Book 1 (*Onomasticon* 1.2) that his work will reveal which words are *sunṓnuma* (usually translated "synonyms"), and then adds by way of clarification that *sunṓnuma* are words that can be substituted for one another—a remarkably distributional way of thinking. While the work does much more than tell the reader which words are *sunṓnuma* (it is a work designed to be read for pleasure as well as instruction, full of learned discussions), it is indeed built around lists of *sunṓnuma*, arranged by topic. The lists show that words counting as *sunṓnuma* for Pollux do not necessarily denote the same thing as each other. While some of his lists collect expressions that do denote roughly the same thing (e.g. being bald, 2.25), others collect expressions denoting members of some category (e.g. parts of the body, 2.22–3). Once again Pollux' concept of semantic similarity is broader than traditional concepts of "synonymy" in our own culture. His concept is arguably closer to the one underlying Distributional Semantics, insofar as words denoting different members of some category might be expected to occur in similar contexts.

In the Preface to Book 1 (1.2), Pollux also announces the structure of the work: after starting with words for gods, he says, he will move on to other topics in the order in which they occur to him. The *Onomasticon* is a more carefully planned work than Pollux lets on here,[7] but it is indeed constructed so that one topic keeps leading to another via a chain of associations. Near the beginning of Book 2 (2.8–16), for example, Pollux gives a long list of expressions for male humans, starting with newborn babies and working up to old men. There follows a list of expressions for female

7. See König (2016, 301), and for an overview of the main themes dealt with in each of the ten books, see Bethe (1917, 776–7).

humans, working from young to old again and including some cross references back to the previous list—the expressions for "baby", for example, all turn out to be gender neutral (2.17). From here Pollux moves on to expressions for giving birth (2.19), for arriving at various ages (2.20), and for parts of the body (2.22–3); then words morphologically related to *thríks* "hair" (2.24), expressions for having various kinds of hair (2.25), expressions for being bald (2.25), and so on.

By arranging the work in this way, Pollux seems to suggest that Greek words (and phrases) are semantically linked via a web of closer and more distant associations, so that to follow them up in a linear order we have to choose an arbitrary path through this web. This thought too anticipates some of our own assumptions. At the same time, we should be wary of reading all our assumptions and goals too hastily into Pollux' work. Most importantly, it is not remotely his goal to sample the Greek lexicon in a way that avoids bias on his part. On the contrary, Pollux' work is designed to reflect his own value judgements, both in the topics and in the vocabulary he prioritises.

As regards vocabulary,[8] which is of particular interest to us, in the Preface to Book 4 Pollux explicitly tells his addressee (the emperor or future emperor Commodus) not to be too surprised if he notices that some word has been omitted. "For I might have omitted it even though I knew about it", Pollux says, "because I did not approve of it" (4.2; cf. Mauduit and Moretti [2010, 524]). Pollux was active at the height of the "atticistic movement"—an extraordinary obsession with reviving the Greek of classical Athens ("classical Attic Greek") for elegant writing and speech. A number of overtly prescriptive lexica from this period ("atticistic lexica"), listing approved and unapproved words, survive in their entirety or in part (see e.g. Dickey [2007, 9, 77, 96–9]; Vessella [2018, 12–26]). Pollux' work is often called an "atticistic lexicon" too, and it belongs in this cultural context, although it is more descriptive in its presentation than the others we have (see e.g. Mauduit and Moretti [2010, 523–4]; Tosi [2013, 144–5]; König [2016, 298–9]; Vessella [2018, 24–5]). Pollux generally lists words without explicitly attaching any value judgements, but occasionally he gives a list of words and then comments that certain further words are inelegant, or are suspected of being inelegant. For example, after giving various words for "delay" or "slowness", he proceeds to comment that the further word *hupérthesis* is "suspected of being cheap", and that *straggeíā* is "very bad" (9.137). By implication, the words he includes without any such comment are approved for use. The usage of classical authors of the fifth and fourth centuries BCE, writing in Attic Greek, looms large in his approved vocabulary, although he takes in vocabulary from works composed in other literary varieties of Greek too, and he is also more open to postclassical vocabulary than most atticists (see e.g. Bethe [1917, 778–9]; König [2016, 299, 304, 307–8]).

If we take Pollux' *Onomasticon* as a proxy for native speaker intuitions, then, we do so with a pinch of salt: Pollux lived five or six centuries after the authors whose vocabulary he tended to prioritise, and he drew on earlier lexicographical works as well as his own reading (see e.g. Bethe [1917, 777–8]; Mauduit and Moretti [2010,

8. On the topics Pollux prioritises, see especially König (2016).

532–6]). In addition, his work does not come down to us in its original form: the work we have is in essence an abbreviated version, although it shows signs of later additions as well as abbreviation (see Bethe [1900-1937, vol. 1, v-vii]; Bethe [1917, 776]).

All these caveats must be kept in mind, then, but the *Onomasticon* remains by far the most substantial surviving work of ancient Greek lexicography to be arranged by topic (the version that survives is about 120,000 words in length, and provides information on the semantics and/or morphology of about 16,000 words);[9] as such, we have chosen to explore its use as a source of independent judgements against which to test semantic space models.

In particular, we chose to find out what nouns (if any) Pollux considers most closely related to 32 target nouns, listed (as "headwords") at `https://github.com/alan-turing-institute/ancient-greek-semantic-space/blob/ancient-greek/pollux_lexicon.txt`. We first of all looked up each of these nouns in the physical *Index Glossarum* to Pollux (i.e. an index of words listed or mentioned, usually in lists of *sunṓnuma*) prepared by Gunnar Andersen and published in the third volume of Bethe's physical edition of the *Onomasticon* (Bethe, 1900–1937, iii. 14–128). From these instances of our words (including inflected forms) we discarded those in which one of our words occurs as part of a multi-word expression; an example is the occurrence of the accusative plural of our word *ómma* "eye" in the expression *epéskhe tà ómmata* "raised its eyes", as part of a list of expressions for what lightning does when it appears (1.117). We also discarded instances in which one of our words is listed together with other words with which it shares a feature of morphology rather than meaning; an example is an instance of our word *ómma* "eye" in a long list of derived nouns with the suffix *-ma* (6.181).[10]

Once we had arrived at a collection of instances of Pollux listing one of our words along with other semantically similar expressions, we then decided where exactly to draw the line between one of Pollux' lists and the next. In some instances this was a straightforward task, but in others some judgement was needed. For example, we

9. See e.g. Dickey (2007, 96); Tosi (2007, 3-6); König (2016, 298, 301-3). The total word count is a rounded version of the one given by the *Thesaurus Linguae Graecae* ("*TLG*", `http://www.tlg.uci.edu`). The estimated number of words forming the object of linguistic analysis is an estimate of the number of entries in the *Index Glossarum* to Bethe's edition of the *Onomasticon* (Bethe, 1900–1937, iii. 14–128); on this index see further below. Many words feature as the object of linguistic analysis more than once, in different connections; our figure of c. 16,000 counts such words only once each.

10. As a check on this method of data collection we re-collected our data from Pollux using simple word searches of the electronic version of Bethe's text of the *Onomasticon* in the *TLG*, for all case- and number-forms of our 32 target words. This process was far more cumbersome than the one based on the *Index Glossarum*, because *TLG* searches turned up many instances of Pollux simply using one of our target words rather than making it the object of his linguistic analysis. There was therefore a much larger number of irrelevant hits to be discarded through careful reading of the passages involved. The results of the two data collection methods were almost identical, but the *TLG* searches added one data point (an instance of the target word *ánemos* "wind", being linked semantically to *pnoḗ* "blowing", at 2.77).

decided to count the list of expressions for male humans (2.8–16) as one list rather than either (a) dividing it into sub-lists for male humans of various age categories or (b) considering it part of a longer list that also includes the expressions for female humans that follow. While this decision can be justified to some extent from Pollux' presentation (he begins the list of female humans by referring back to the beginning of the list of male humans, as if these two lists are parallel structural units) no hard and fast rules can be given, and different decisions could have been made. Here and in many other places Pollux gives us lists which can be divided up in more than one way, perhaps because he saw that reality too can be divided up in multiple ways.

After making working lists of all words (including words Pollux suggests are inelegant) that Pollux includes in the same immediate list (or lists) as one of our 32 target words, we discarded from these working lists all expressions consisting of more than one word, and all one-word expressions consisting of something other than a noun (for the rationale behind this decision, see section 3.1.2). For these purposes we defined "nouns" as words with an entry in Liddell *et al.* (1996) ("LSJ") treating them as basically nouns. Substantivised adjectives (including substantivised participles) were thus discarded, unless they are given an LSJ entry separate from that of the adjective or (in the case of participles) the verb. A substantivised adjective was defined as having its own LSJ entry if LSJ gives the noun its own lemma in bold type. We also turned words into the form in which they are in fact listed in LSJ; this was normally the nominative singular, but occasionally the nominative plural.

The resulting word lists are pale reflections of Pollux' work, and should not be confused with his work itself. What they give us is some judgements that various nouns are closely related in meaning to particular nouns among the 32 that we took as a starting point. These judgements are derived from Pollux' work, and are independent of the semantic space models that we would like to evaluate.

3.1.2. *Modern lexicography*

The modern lexicographical resource used for comparison is J.H. Schmidt's three-volume *Synonymik der griechischen Sprache* (Schmidt, 1876-1886). This dictionary contains lists of ancient Greek synonyms, organised into 150 lexical areas according to the editor's judgement. The material included, compiled with traditional philological methods, mostly reflects the usage of classical Greek authors (from the 5th and 4th century BCE) and of the Homeric poems. Each list of synonyms is followed by a detailed discussion, highlighting differences in usage and nuance.

Schmidt's sections are not organised by part of speech, nor do they exclusively contain synonyms in the stricter sense of words that can be used interchangeably in the same sentence context, but rather words that are closely related in meaning and/or derive from the same root. So, for instance, section 1 broadly gathers words for "speaking", ranging from verbs meaning "to speak" (with different nuances, from *légein* "to say" to *laleîn* "to chatter") to nouns for "word" and "voice". This organisation provides a fairly good match for the synsets included in Ancient Greek WordNet (see below).

Resource	Date	Lemmas	Creation	Synonymy defined as
Pollux	2nd c. CE	ca. 400	manual	substitution in context
Schmidt	1876–1886	ca. 1250	manual	similar meaning/etymology
AGWN	2016	ca. 22400	automatic	Princeton WordNet synset

Table 1. *Summary of the characteristics of the benchmark resources. For Pollux and Schmidt, the size reported is that of the sample used. The datasets are available at* `https://zenodo.org/badge/latestdoi/174973156`.

As was the case for Pollux's *Onomasticon*, the assessment that follows only considers nouns in each section of Schmidt's dictionary—not adjectives or verbs. This is to obtain a better match for the example analysis in section 4, where only nouns that act as fillers in Homeric formulae will be considered. Due to the purposes of this analysis, we are interested first and foremost in how accurately the DSM will be able to match existing resources in assessing the semantic similarity between nouns, not other parts of speech.

3.1.3. *Ancient Greek WordNet*

The third resource used for the benchmark comparison is Ancient Greek WordNet (Boschetti *et al.*, 2016) (`http://hdl.handle.net/20.500.11752/ILC-56/`), a lexico-semantic resource developed in collaboration between the Institute of Computational Linguistics "Antonio Zampolli" in Pisa, the Perseus Project in Boston, the Open Philology Project in Leipzig and the Alpheios Project in New York. AGWN is a WordNet built by automatically extracting Greek-English pairs from existing Greek-English dictionaries (LSJ, Middle Liddell, Autenrieth) and then linking the English word to its corresponding synset in the Princeton WordNet (Bizzoni *et al.*, 2014).

English, therefore, acts as an intermediate step in the construction of AGWN, which introduces a potentially significant amount of errors, including but not limited to erroneously grouping words that are translated by English homonyms under the same synset, misinterpreting the part of speech of a word (e.g. by considering expressions including the adjective "joint" as synonyms of the noun "joint" in its various possible meanings), etc. While some level of manual correction was performed by the developers to remove the most obvious mismatches (like the ones arising from the introduction of modern semantic areas such as aviation or telecommunications), AGWN still contains a high amount of noise (Bizzoni *et al.*, 2015).

The characteristics of the three gold-standard resources are summarised in table 1.

3.2. *Evaluation method*

In order to assess the way in which different parameters such as size of the context window and frequency threshold for the lemmas included in the semantic spaces

affected the spaces themselves, and in order to gain insights into the linguistic properties of the three lexical resources we relied on for this study, we compared different parameter configurations against the gold-standard sets.

The semantic spaces are defined based on corpus co-occurrence frequency counts, and therefore they offer a direct way to conceptualise geometric distances between lemmas in terms of their distributional features. For example, in the 300-dimensional semantic space obtained with the SVD-based dimensionality reduction on the corpus co-occurrence matrix defined by a context window of size 5 and a frequency threshold of 50, the top 10 neighbours of the lemma *hiketeía* "supplication", excluding *hiketeía* itself, are: *déēsis* "entreaty" (cosine similarity with *hiketeía*: 0.38), *oîktos* "pity" (0.43, *hiketeúō* "to beg" (0.44), *hikesía* "supplication" (0.44), *epikláō* "to bend, move to pity" (0.47), *epḗkoos* "listening" (0.48), *hupereîdon* "looked over" (past tense) (0.48), *aítēsis* "request" (0.49), *mneía* "mention" (0.49), and *liparḗs* "persisting" (0.50). On the other hand, as we have seen in section 3.1, the three lexical resources we considered as gold-standard differ to some extent in the way they group lemmas into groups of semantically related words. In the case of AGWN, these groups contain synonyms – and are therefore called synsets – which are defined based on linguistic and world knowledge rather than directly on corpus data. For example, AGWN records *hiketeía* in the synset 07187638-n glossed as "a humble request for help from someone in authority". This synset contains the following other lemmas, which are all synonyms of "prayer", "plea", and "supplication": *skêpsis*, *paraítēsis*, *próphasis*, *lipárēsis*, *goúnasma*, and *hikesía*. The aim of this evaluation was to find which semantic space model(s) most closely preserve(s) the semantic features displayed in the gold-standard resources, thus highlighting any linguistically-relevant differences.

3.2.1. *Precision and recall*

With the aim of measuring to what extent the corpus-driven distributional definition of a lemma's neighbours matched the definition of synonymy or semantic relatedness from the gold-standard resources, we focussed on the lemmas that appear in the resources and are also listed as top 10 corpus neighbours ("neighboursets") in the semantic spaces; we called these lemmas "shared lemmas". For each shared lemma, we compared its top 10 corpus neighbours[11] with its resource's synonyms,[12] and calculated precision and recall. Precision of a lemma l is defined in terms of the number of corpus neighbours for l which are also in the synset for l, divided by 10 (the number of corpus neighbours at our disposal); precision measures the proportion of

11. The selection of the top 10 corpus neighbours was driven by feasibility considerations. DISSECT offers the option of displaying the top x neighbours for each lemma. Given the considerable size of the output data returned by DISSECT and the high number of parameter combinations, we had to make the decision on x upfront. Choosing different values of x and measuring the effect of this variation could be the focus of further research, and beyond the scope of this study.

12. In the rest of this article, we will use the term "synonym" to broadly refer to any semantically related word that we included in the gold-standard lists, and the term "synset" for the list of such related words associated to a given lemma in the gold-standard resources.

corpus neighbours which are considered "correct" in the gold-standard. On the other hand, recall is the same number divided by the number of elements in the synset of l, and measures how many of the expected synonyms (according to the gold-standard) appear as corpus neighbours.

$$P(l) = \frac{|synset(l) \cap neighbourset(l)|}{|neighbourset(l)|} \quad [1]$$

$$R(l) = \frac{|synset(l) \cap neighbourset(l)|}{|synset(l)|} \quad [2]$$

Together, these complementary measures give us a complete picture of the extent to which the two sets of resources overlap in terms of their content, thus allowing us to gain linguistic insights into their similarities and differences.[13] In the example for *hiketeía* illustrated above, the two sets both contain the lemma *hikesía*; therefore precision is 1/10=0.1 and recall is 1/7 = 0.14.

The fourth and fifth columns of table 2 report the mean precision and mean recall across all shared lemmas by combination of parameters for the semantic spaces, namely size of context window, frequency threshold, and gold-standard resource.[14] The sixth column of table 2 shows the range (minimum and maximum) of values corresponding to the number of overlapping lemmas between corpus neighbour-sets and resources' synsets. The last column shows the number of shared lemmas between the semantic spaces and each of the resources.

For the majority of lemmas, there is no overlap between the corpus neighbour-sets and the resources' synsets. Consequently, mean precision and recall have low values, ranging between 0.02 and 0.10, and between 0.02 and 0.09, respectively. We interpret this as being due to the fundamental difference between corpus neighbours, which tend to be lemmas that behave in a distributionally similar way, and the synonyms recorded in the resources, which reflect the author's knowledge and intuition about the lemmas' semantic properties and usage.

The gold-standard resource whose synonyms most closely match the corpus neighbours according to our definitions of mean precision is Pollux (with the combination of context window 5 and frequency threshold 20), which achieves the highest value for mean precision of 0.10. The highest mean recall value, 0.09, is reached by Schmidt (context window 1 and frequency threshold 100). In general, we can say that the semantic spaces perform better when compared with both the resource from ancient lex-

13. These measures, however, only take into account whether synonyms appear in the top 10 neighbour lists rather than on their numeric distances in the semantic spaces. The latter will be the focus of the rank-based measures described in section 3.2.2.

14. Because the computational cost of performing the evaluation on AGWN for the semantic space `w1_t1` (context window 1 and frequency threshold 1) was excessively high due to the high number of AGWN synonym pairs, results for this space are not reported and are left to future research.

Res.	W	Freq	Avg P	Avg R	Range	Coverage
AGWN	1	20	0.02	0.02	[0,5]	6864
	1	50	0.03	0.02	[0,5]	4666
	1	100	0.03	0.02	[0,4]	3329
	5	20	0.03	0.02	[0,5]	6865
	5	50	0.03	0.03	[0,5]	4666
	5	100	0.03	0.02	[0,4]	3329
	10	20	0.03	0.02	[0,6]	6865
	10	50	0.03	0.02	[0,5]	4666
	10	100	0.03	0.02	[0,4]	3329
POLLUX	1	1	0.04	0.02	[0,3]	309
	1	20	0.07	0.04	[0,5]	236
	1	50	0.07	0.05	[0,4]	177
	1	100	0.08	0.05	[0,4]	146
	5	1	0.08	0.04	[0,6]	313
	5	20	**0.10**	0.05	[0,6]	236
	5	50	0.09	0.05	[0,4]	177
	5	100	0.09	0.05	[0,4]	146
	10	1	0.07	0.03	[0,6]	313
	10	20	0.08	0.04	[0,5]	236
	10	50	0.08	0.04	[0,5]	177
	10	100	0.08	0.04	[0,4]	146
SCHMIDT	1	1	0.03	0.03	[0,3]	1029
	1	20	0.06	0.07	[0,4]	701
	1	50	0.08	0.08	[0,4]	531
	1	100	0.08	**0.09**	[0,4]	423
	5	1	0.05	0.04	[0,4]	1046
	5	20	0.07	0.07	[0,4]	701
	5	50	0.08	0.08	[0,5]	531
	5	100	0.07	0.08	[0,4]	423
	10	1	0.04	0.04	[0,5]	1046
	10	20	0.06	0.06	[0,5]	701
	10	50	0.07	0.07	[0,5]	531
	10	100	0.06	0.06	[0,4]	423

Table 2. *Values of three evaluation metrics calculated between the semantic spaces with different parameter combinations and the three gold-standard resources. Column 1 contains the name of the resource under consideration: Ancient Greek WordNet (AGWN), Schmidt, and Pollux. Columns 2 and 3 show the size of the context window and the frequency threshold used to define the semantic spaces, respectively. Columns 4 and 5 show mean precision and mean recall across all shared lemmas, calculated based on the overlapping lemmas between corpus neighbour-sets and resources' synsets. Column 6 shows the range of values corresponding to the number of overlapping lemmas between corpus neighbour-sets and resources' synsets, and column 7 shows the number of shared lemmas between the semantic spaces and the resources. See text for a detailed explanation.*

icography (Pollux) and the one from modern lexicography (Schmidt) than when compared with the computational resource (AGWN). Moreover, to the extent that there is a difference, given a fixed context window, frequency threshold levels of 20 or 50 tend to yield higher mean precision and recall values than lower and higher levels.

3.2.1.1. Precision and recall by frequency and polysemy class

When selecting the content to analyse in Pollux, we focussed on 32 lemmas shared by the semantic space with a frequency threshold of 50, AGWN, and Schmidt. The full list of lemmas was provided in section 3.1. Three of these words are not present in the `t_100` semantic space. We categorised these 32 words into two frequency classes ("frequent" and "infrequent") and two polysemy classes ("polysemous" and "monosemous"). The former categorisation was based on the Diorisis corpus frequency counts by setting a threshold corresponding to the average frequency among words that occur at least 50 times, i.e. 548.86.[15] The latter categorisation was based on the number of sense labels in the TLG's online version of LSJ (`http://stephanus.tlg.uci.edu/lsj/`): words with more than 3 senses (as indicated by Roman numerals) and words with more than one sense marked with a Latin letter or with two separate dictionary entries (homonyms) were all marked as polysemous. The 32 words were chosen so that each came from a different synset in Schmidt. We extracted synsets with a random number generator, then chose one word per synset with the required combination of traits. This categorisation allowed us to analyse the possible effects of frequency and polysemy on the precision and recall metrics across the gold-standard resources and according to different semantic space parameters.

We collected a dataset containing, for each of the 32 categorised lemmas, its frequency and polysemy class, its precision and recall measures for each of the combination of parameters available for the semantic spaces (context window and frequency threshold) and for each of the three gold-standard resources. We then fitted a linear regression model that predicts recall based on the best subset of all possible predictors. We selected the best model according to the lowest Akaike Information Criterion (Akaike, 1974) using the Stepwise Algorithm (Hastie and Pregibon, 1992). We obtained the model with the following predictors:

– precision;

– polysemy: values "TRUE" or "FALSE";

– resource: values "AGWN", "Pollux", or "Schmidt".

The model's diagnostic checks make us confident that it fits the data reasonably well. The model's R^2, i.e. proportion of variation in the response by it, is 0.60 and its adjusted R^2 (i.e. corrected for the number of predictors included in the model) is also 0.60. The scatter plot of standardised predicted values versus standardised residuals

15. We used the average frequency rather than the median (141) as the latter, being very low, would have lead us to select words that do not appear in Schmidt.

Predictor	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.006976	0.004405	-1.584	0.1137 *
precision	0.843005	0.027876	30.242	< 2e-16***
resourcePOLLUX	-0.048343	0.010777	-4.486	8.52e-06 ***
resourceSCHMIDT	0.038958	0.004509	8.641	< 2e-16 ***
polysemousTRUE	-0.011308	0.004486	-2.521	0.0119 *

Table 3. *Summary of linear regression model predicting recall of a lemma based on its precision, its frequency and polysemy class, and the gold-standard resource. Significance codes: "***" means significant at the 0.001 level, "**" at the 0.01 level, and "*" significant at the 0.05 level.*

shows that the dataset meets the assumptions of homogeneity of variance and linearity, and the residuals are approximately normally distributed.

Table 3 reports the summary of the model in terms of the estimated coefficient for each predictor's value, the standard error of this estimate, the t value, and the associated t-statistic and p-values. The predictors that are significant at the 0.05 level are indicated by the presence of asterisks in the table. Higher precision values lead to higher recall values: we can interpret the coefficient 0.84 as the average effect on recall of a one-unit *increase* in precision, holding all other predictors fixed. Although this result is not completely surprising, because precision and recall are calculated based on the same numerator (i.e. the number of overlapping synonyms), precision was included in the list of predictors because this led to a model that fitted the data well, and therefore could be interpreted in a meaningful way. Moreover, this allows us to ascertain the relative effect that linguistically-interesting predictors such as "polysemy" and "resource", compared to precision, have on recall.

Regarding the effect of the gold-standard resources, compared to the reference level (AGWN),[16] comparing the corpus-driven semantic spaces to Schmidt *increases* recall by 0.04, while using Pollux has the effect of *decreasing* recall by 0.05. Coming to the features of the lemmas, we can see that the coefficient relative to polysemy is negative, which means that if a lemma is categorised as polysemous, this leads to a *decrease* in recall of approximately 0.01, which is expected given that polysemous words present more challenges in such semantic similarity tasks.

3.2.2. *Rank-based evaluation*

In addition to calculating precision and recall metrics, which compare synsets and neighbour-sets in terms of their content as discrete groups, we devised two other evaluation methods, which take into account a graded measure of semantic relatedness. The aim in this case is to assess to what extent the relationship of semantic related-

16. Because the reference level for the "resource" variable is AGWN, this does not appear in the list of table 3.

ness recorded in the gold-standard resources is reflected in the semantic spaces by considering the ranking of synonyms in the different resources.

For each resource we defined a square co-occurrence matrix based on the distribution of synonyms across the resource's synsets. The rows and columns of the matrix correspond to the lemmas shared between the resource and the semantic spaces, in each of their parameter configurations. For example, in Schmidt ((1876-1886)) the lemmas *epiméleia* "care, concern" and *mérimna* "care, thought" occur together in one synonym set, the 86th one. Therefore, the entry for *epiméleia* in the co-occurrence matrix for Schmidt has value 1 in the cell corresponding to the column for *mérimna*. Similarly, in AGWN these two lemmas occur together in the synset 00267522-n and therefore in the co-occurrence matrix for AGWN the entry for *epiméleia* has value 1 in the cell corresponding to the column for *mérimna*, too. Following this approach, we can define a vector for each shared lemma and therefore calculate cosine similarity measures between pairs of shared lemmas in the resources' spaces. The co-occurrence matrix defined this way allows us to measure the distance (which is 1 minus the similarity score) between two lemmas according to a metric based on such resource-specific semantic relatedness. For example, the cosine similarity between the vector for *epiméleia* and the vector for *mérimna* in the space defined by the semantic relatedness from Schmidt's resource is 1, as these two lemmas co-occur in exactly the same synsets.

For each pair of synonyms in the lexicons, we compared their resource-based distance to their corpus-based distance. In the example for *epiméleia* and *mérimna*, their cosine similarity in the semantic space with context window 5 and frequency threshold 50 is 1-0.67, which is a consequence of the fact that these lemmas co-occur with different sets of lemmas in the corpus, and therefore their distributions are not identical, as in the case of the space defined from Schmidt's resource.

For each parameter configuration for the semantic spaces and for each of the three lexical resources, we calculated the Average Inverse Rank (InvR), an information-retrieval measure which has been used in other relevant studies, such as Henestroza Anguiano and Denis (2011). For each lemma l, we considered its top 10 corpus neighbours[17] and ranked them by decreasing corpus-based distance. All neighbours appearing in a synset of l in the lexical resource were considered as relevant. InvR is then the average, among all shared lemmas, of the sum of inverse ranks of relevant neighbours:

$$InvR = \sum_{n} \frac{1}{rank(n)} \quad [3]$$

For example, for the lemma *stráteuma* "expedition; army", two corpus neighbours are also considered semantically related to it by Pollux: *stratópedon* "camp; army" and *stratiṓtēs* "soldier". Their positions in the ranking based on decreasing corpus

17. As explained previously, the choice of 10 as the number of corpus neighbours considered had the aim of keeping the calculations computationally manageable.

distance are 2 and 4, respectively, which leads to the inverse ranks of 0.5 and 0.25, respectively. Therefore, for *stráteuma*, the sum of inverse ranks is 0.50+0.25 = 0.75.

The fourth column of table 4 shows the average InvR for each configuration. This ranges from 0.36 and 0.54, which are comparable to the results on similar analyses for a modern language like French (Henestroza Anguiano and Denis, 2011). The maximum values are reached for Pollux (context window 1 and frequency threshold 100) and Schmidt (context 5 and frequency threshold 50).

We also ran a Spearman's correlation test on the following two distributions: the distribution of resource-based cosine similarity between pairs of shared synonyms and the distribution of corpus-based cosine similarity between the same pairs. The fifth column in table 4 reports the values of the coefficients of Spearman's correlation tests which returned significant results (α = 0.05). The coefficients indicate a very weak or weak positive relationship between the two distributions, ranging between 0.05 and 0.23. This is not completely unexpected, as this second rank-based measure imposes a stricter condition compared to InvR and the corpus-based distance is defined in a qualitatively different way from the resource-based distances. In spite of the weakness of this relationship, however, the Spearman's correlation coefficients are useful to identify the relative differences between the resources. The highest value is reached by the comparison between Pollux and the semantic space with context window 5 and frequency thresholds 50 and 100. As in the case of precision and recall measures, we notice that the semantic spaces perform better when compared with non-computational resources than when compared with AGWN according to both the rank-based measures.

4. Sketch of an application: the flexibility of Homeric formulae

This section will show the research potential of the distributional approach by applying it to a small-scale study of semantic flexibility in early Greek epic formulae. We will explore the behaviour of two verb phrase formulae denoting, respectively, holding and thinking: formulae of the type *(en/metà) khersìn ékhein*, "to hold [x] in one's hands" vs. formulae of the type *eû eidénai*, "to have [x] in mind". These two formulae were chosen for two reasons: (1) as they both contain a transitive verb, they involve an open slot that is essentially always filled; (2) they are the two most frequent formulae with characteristic (1) based on a formula search on the Chicago Homer (`http://homer.library.northwestern.edu/`).[18]

For both formulae, we considered their object fillers (the [x] variable in the above translation); we aim to highlight how the semantic coverage of possible objects of these formulae influences their behaviour. This will show the usefulness of the DSM to classicists, as well as provide further insights on its performance. While a sample size of 2 formulae obviously will not allow us to reach meaningful conclusions about

18. Excluding speech introduction formulae.

Resource	Window	Freq	Average IR	Spearman's corr.
AGWN	1	20	0.46	0.05
	1	50	0.48	0.06
	1	100	0.49	0.07
	5	20	0.45	0.06
	5	50	0.36	0.07
	5	100	0.46	0.08
	10	20	0.45	0.06
	10	50	0.46	0.07
	10	100	0.45	0.07
POLLUX	1	1	0.44	0.08
	1	20	0.43	0.14
	1	50	0.50	0.18
	1	100	**0.54**	0.18
	5	1	0.48	0.17
	5	20	0.49	0.20
	5	50	0.49	**0.23**
	5	100	0.46	**0.23**
	10	1	0.52	0.17
	10	20	0.48	0.19
	10	50	0.47	0.22
	10	100	0.44	0.21
SCHMIDT	1	1	0.49	0.05
	1	20	0.50	0.08
	1	50	0.52	0.09
	1	100	0.51	0.10
	5	1	0.44	0.07
	5	20	0.49	0.09
	5	50	**0.54**	0.09
	5	100	0.50	0.10
	10	1	0.42	0.07
	10	20	0.47	0.08
	10	50	0.47	0.09
	10	100	0.47	0.09

Table 4. *Values of rank-based evaluation metrics calculated between the semantic spaces with different parameter combinations and the three gold-standard resources. Column 1 contains the name of the resource under consideration: Ancient Greek WordNet (AGWN), Schmidt, and Pollux. Columns 2 and 3 show the size of the context window and the frequency threshold used to defined the semantic spaces, respectively. Column 4 shows the average inverse rank calculated based on the ranks of corpus neighbours which appear in the corresponding resource's synsets. Column 5 shows the correlation coefficient when a Spearman's correlation test run on the distributions of distances between pairs of shared synonyms (calculated in the spaces defined by the resources) and the distances between the same pairs of synonyms (calculated in the corpus semantic spaces) returned a significant result (significance threshold: 0.05).*

formulae in general, this is not the goal of the present use case; we merely aim to show how our distributional model can be applied in a "pocket-sized" version of what will be its ultimate application in further work.

The two formulae chosen for this study differ not only in their meaning (while both denoting a form of having/holding, literal or metaphorical), but also in their syntactic behaviour and usage. The formula for physically holding, "having in one's hands", only takes a direct object in the accusative, as opposed to the one for thinking/mentally holding, "having in mind", which can be construed with either the genitive, mostly used with nouns, or the accusative, mostly with pronouns such as *hóde* or *hoûtos*. (Cases in which the object is represented by a clause were excluded.) The latter rule, however, admits a few exceptions, where content words are construed in the accusative; this happens twice in the *Iliad* and four times in the *Odyssey* (*Il.* 13.665, 20.213; *Od.* 9.215, 11.442, 11.445, 14.365), a pattern that may hint at the accusative with content words being a more recent development. While pronouns like *hóde* or *hoûtos* have been filtered as stopwords, the above-mentioned cases of accusative with nouns and two cases with *pâs* and *hápas* "all, everything" are included. Therefore syntactic flexibility can be considered a relevant parameter.

Other notable differences concern frequency and attestation: the hands-formula occurs 59 times with an object, in both the Homeric poems and in poems from other branches of the tradition (3 times in Hesiod's *Theogony* and 6 in the *Homeric Hymns*), while the know-formula occurs 42 times and is limited to the Homeric epics. If we consider Hesiod and the Hymns to be later than the Homeric epics (Andersen and Haug, 2012), this may be a sign of chronological development; even if, however, we reject this chronological hypothesis,[19] we still know that the formula is productively used outside of Homer, i.e. in other branches of the epic tradition. The fact that the frequency of the two formulae is very similar in the Homeric epics (50 vs. 42 times) makes this unlikely to be due to sparsity of attestation of the know-formula.

Based on these characteristics, we could expect a correlation between semantic flexibility and either syntactic flexibility (hypothesis A, according to which the know-formula should be the most flexible of the two) or productivity of the formulae across traditions (hypothesis B, according to which we would expect higher flexibility in the hand-formula). While these are merely two possible explanatory factors, they serve the purpose of our example case, i.e. showing how quantitative information from the DSM can be used to assess hypotheses about semantic development. The experiment below allows us to test these hypotheses. This cannot, of course, lead to a claim of causality in a study conducted on a sample of two formulae; it can, however, provide a simple example of how the DSM model can be used to analyse formulaic behaviour.

19. For a detailed formulaic analysis that comes to the conclusion that the Hesiodic poems pre-date the Homeric ones, see Pavese and Venti (2000), Pavese and Boschetti (2003).

	Hand-construction	Know-construction
Min.	0.2640	0.3288
Median	0.4515	0.4085
Mean	0.4547	0.4044
Max.	0.6839	0.5404

Table 5. *Summary of the distribution of distances from the fillers of the hand- and know-constructions to their respective centroids in the* `w5_t50` *semantic space.*

4.1. *Choosing a DSM*

In light of the results of the benchmark comparison, the DSM chosen for our experiment is the `w5_t50` one (context window of 5 words on each side and frequency threshold of 50). The `w5` spaces consistently showed the best performance for both precision and recall among the three context windows that were tested (if we exclude one isolated result from the `w1_t100` space compared to Schmidt); as for frequency, while higher thresholds appear to improve accuracy, they also mean potentially being unable to draw conclusions for some of the fillers in our target constructions, as some of these words occur relatively rarely in the corpus. As it stands, out of 40 filler types for the hand-construction and 23 for the know-construction, respectively 10 and 5 are not present in the `t50` space, because of either stop-word filtering or the frequency threshold.

4.2. *Methodology and results*

Fillers for both constructions were extracted by hand through a lemma search on the Perseus Project's Scaife viewer (`https://scaife.perseus.org/`, last accessed March 2019). After extracting the filler words, a Python script (available at `https://zenodo.org/badge/latestdoi/174973156`) was used to compute the centroid of the coordinates of the fillers for each construction, and then measure the cosine similarity of each filler to this centroid. We can then use the radius of the distribution of the fillers as a proxy to its density, giving us a measure of how closely clustered (semantically similar) the fillers themselves are in the semantic space.

Finally, the distances for both constructions were loaded and analysed in R (R Core Team, 2019). Table 5 contains the summary of the two distributions, which were then compared using the Kolmogorov-Smirnoff test. Both the mean and the extremes of the distribution for the hand-construction are higher than for the know-construction; the difference is statistically significant (P = 0.003).

4.3. *Discussion*

As table 5 shows, the hand-formula shows higher semantic flexibility in its open slot. The objects of formulae of the type *(en/metà) khersìn ékhein*, "to hold [x] in one's hands", are more varied in their meaning than the objects of *eû eidénai*, "to have [x] in mind". If we return to our initial hypotheses as outlined before the experiment, we can now conclude that while the higher syntactic flexibility of the know-construction does not correlate with higher semantic flexibility, contrary to hypothesis A, the construction with higher semantic flexibility, i.e. the hand-construction, is the one that is productive outside of Homer, in accordance with hypothesis B.

This experiment, of course, needs to be extended to a much wider network of formulae in order to allow for significant conclusions to be drawn on the behaviour of Homeric formulae; however, this preliminary example shows how the DSM can be used to address the issue of formulaic behaviour from a quantitative standpoint.

5. Conclusions

As stated in section 1, our main objective for this study was to discuss how to build a resource that is actually useful for a DH approach, despite the inevitable limitations. These depend in part on the characteristics of the corpus, which is small in size and, by nature, cannot be expanded: this affects the accuracy of computational tools. They also have to do with the wider context of its use: we cannot gain direct access to native speakers' semantic judgements for ancient Greek, which means having to refer to comparison sources that are each in turn limited and idiosyncratic in their characteristics. The comparison between an ancient lexicographical work and the outcomes of modern computational semantic analysis is a topic that would warrant further study to understand how notions of synonymy and meaning in antiquity compare to modern distributional definitions. The gold-standard resources still allow for a reliable comparison between different DSMs, assessing their relative accuracy. Finally, the use case in section 4 showcases both the limitations and the strengths of the DSM. On the one hand, the proposed use of the semantic model also dictates its characteristics: frequency filtering improves the overall accuracy of the DSM but limits its applicability when dealing with rare words. On the other hand, our pocket-sized experiment shows how the distance measures extracted from the DSM can be used to assess hypotheses about semantic flexibility and formulaic usage from a quantitative perspective, something that has thus far never been attempted for ancient Greek epic.

Finally, this study wishes to highlight the strengths and the potential applications of Distributional Semantics as a resource for research in ancient studies. The most important advantage of this approach, and of computational models in general, is that they allow us to analyse semantic behaviour on a quantitative level, to make detailed and objective comparisons, and to advance quantifiable claims that can in turn be tested. These are all fundamental standards in quantitative historical linguistics (Jenset and McGillivray, 2017), and can all be reached with careful application and evaluation

of computational methods on ancient sources. Further developments in the accuracy of these methods can lead to valuable results for experimental research, an example of which we have outlined by discussing ancient Greek formulae. A more detailed study of this topic is in preparation, and we anticipate that more research on the subject will further strengthen the case for the use of distributional models in philological research.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. MAR's stay at The Alan Turing Institute was supported by the 2019/20 Enrichment scheme. All authors designed the study. MAR built the semantic spaces, compiled the lexicon from Schmidt (1876-1886), designed and carried out the analysis for the case study, and wrote all of the sections that are not otherwise credited. PP collected the data from Pollux and wrote section 3.1.1. BMcG designed and implemented the evaluation approach, advised on the creation of the semantic spaces, and wrote section 3.2. All authors gave final approval for publication.

6. References

Akaike H., "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, vol. 19, p. 716-723, 1974.

Andersen Ø., Haug D. T. (eds), *Relative chronology in early Greek epic poetry*, Cambridge University Press, Cambridge, 2012.

Baroni M., Dinu G., Denis P., "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 238-247, 2014.

Barðdal J., *Productivity: Evidence from case and argument structure in Icelandic*, John Benjamins, Amsterdam; Philadelphia, 2008.

Bethe E., *Pollucis Onomasticon*, Teubner, Stuttgart, 1900–1937.

Bethe E., "Iulius (398) Pollux", *in* A. Pauly, G. Wissowa, W. Kroll (eds), *Real-Encyclopädie der classischen Altertumswissenschaft, x.i*, Metzler, p. 773-779, 1917.

Bizzoni Y., Boschetti F., Diakoff H., Del Gratta R., Monachini M., Crane G., "The making of Ancient Greek WordNet", *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, European Language Resources Association (ELRA), Reykjavik, p. 1140-1147, 2014.

Bizzoni Y., Del Gratta R., Boschetti F., Reboul M., "Enhancing the accuracy of Ancient Greek WordNet by multilingual distributional semantics", *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, Accademia University Press, Trento, p. 47-50, 2015.

Boschetti F., *Copisti digitali e filologi computazionali*, CNR Edizioni, Roma, 2018.

Boschetti F., Del Gratta R., Diakoff H., "Open Ancient Greek WordNet 0.5", 2016.

Bullinaria J. A., Levy J. P., "Extracting semantic representations from word co-occurrence statistics: A computational study", *Behavior Research Methods*, vol. 39, p. 510-526, 2007.

Bullinaria J. A., Levy J. P., "Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD", *Behavior Research Methods*, vol. 44, p. 890-907, 2012.

Celano G., "Guidelines for the annotation of the Ancient Greek Dependency Treebank 2.0", 2014.

Curran J. R., *From Distributional to Semantic Similarity*, PhD Diss., Institute for Communicating and Collaborative Systems, School of Informatics, Edinburgh, 2004.

Dickey E., *Ancient Greek Scholarship: a guide to finding, reading, and understanding scholia, commentaries, lexica, and grammatical treatises, from their beginnings to the Byzantine period*, Oxford University Press, Oxford, 2007.

Dinu G., Pham N. T., Baroni M., "DISSECT - DIStributional SEmantics Composition Toolkit", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, p. 31-36, 2013.

Evert S., "Corpora and collocations", *in* A. Lüdeling, M. Kytö (eds), *Corpus linguistics. An international handbook*, Mouton de Gruyter, p. 1212-1248, 2008.

Fabre C., Lenci A., "Distributional Semantics today", *TAL – Traitement Automatique des Langues*, vol. 56, p. 7-20, 2015.

Foley J. M., *Homer's traditional art*, Pennsylvania State University Press, University Park (PA), 1999.

Grewcock R., *Computational semantics and the syntax of motion in Ancient Greek*, MPhil Thesis, University of Cambridge, Cambridge, 2018.

Harris Z. S., "Distributional structure", *Word*, vol. 10, p. 146-162, 1954.

Hastie T. J., Pregibon D., "Generalized linear models", *in* J. M. Chambers, T. J. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, chapter 6, 1992.

Haug D. T., Jøhndal M., "Creating a parallel treebank of the old Indo-European Bible translations", *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, p. 27-34, 2008.

Henestroza Anguiano E., Denis P., "FreDist: Automatic construction of distributional thesauri for French", *TALN - 18e conférence sur le traitement automatique des langues naturelles, Jun 2011*, Montpellier, France, p. 119-124, 2011.

Jenset G. B., McGillivray B., *Quantitative Historical Linguistics. A Corpus Framework*, Oxford University Press, Oxford, 2017.

Kahane A., "The complexity of epic diction", *Yearbook of Ancient Greek Epic Online*, vol. 2, p. 78-117, 2018.

König J., "Re-reading Pollux: encyclopaedic structure and athletic culture in *Onomasticon* Book 3", *Classical Quarterly*, vol. 66, p. 298-315, 2016.

Landauer T. K., Dumais S. T., "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge", *Psychological Review*, vol. 104, p. 211-240, 1997.

Levy O., Goldberg Y., "Dependency-based word embeddings", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Baltimore, Maryland, p. 302-308, 2014.

Liddell H., Scott R., Jones H., *et al.*, *A Greek-English Lexicon,* 9th edn, 1940, with revised supplement, 1996, Clarendon Press, Oxford, 1996.

Mauduit C., Moretti J.-C., "Pollux, un lexicographe au théâtre", *Revue des études grecques*, vol. 123, p. 521-541, 2010.

McGillivray B., Hengchen S., Lẗeenoja V., Palma M., Vatri A., "A computational approach to lexical polysemy in Ancient Greek", *Digital Scholarship in the Humanities*, 2019.

Mikolov T., Chen K., Corrado G., Dean J., "Efficient estimation of word representations in vector space", 2013.

Pavese C. O., Boschetti F., *A Complete Formular Analysis of the Homeric Poems*, Hakkert, Amsterdam, 2003.

Pavese C. O., Venti P., *A Complete Formular Analysis of the Hesiodic Poems*, Hakkert, Amsterdam, 2000.

Pennington J., Socher R., Manning C., "GloVe: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, p. 1532-1543, 2014.

Perek F., "Using distributional semantics to study syntactic productivity in diachrony: A case study", *Linguistics*, vol. 54, p. 149-188, 2016.

R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. 2019.

Rodda M. A., Senaldi M. S., Lenci A., "Panta Rei: Tracking semantic change with Distributional Semantics in ancient Greek", *Italian Journal of Computational Linguistics*, vol. 3, p. 11-24, 2017.

Schmid H., "Probabilistic Part-of-Speech tagging using decision trees", *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Schmidt J. H., *Synonymik der griechischen Sprache*, Teubner, Leipzig, 1876-1886.

Tanguy L., Sajous F., Hathout N., "Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques", *TAL – Traitement Automatique des Langues*, vol. 56, p. 103-127, 2015.

Tosi R., "Polluce: struttura onomastica e tradizione lessicografica", *in* C. Bearzot, F. Landucci, G. Zecchini (eds), *L'Onomasticon di Giulio Polluce: tra lessicografia e antiquaria*, Vita e Pensiero, p. 3-16, 2007.

Tosi R., "Onomastique et lexicographie: Pollux et Phrynichos", *in* C. Mauduit (ed.), *L'Onomasticon de Pollux: aspects culturels, rhétoriques et lexicographiques*, De Boccard, p. 141-146, 2013.

Vatri A., McGillivray B., "The Diorisis Ancient Greek Corpus", *Research Data Journal for the Humanities and Social Sciences*, 2018.

Vessella C., *Sophisticated speakers: Atticistic pronunciation in the Atticistic lexica*, De Gruyter, Berlin, 2018.