

# A computational approach to lexical polysemy in Ancient Greek

**Barbara McGillivray**

The Alan Turing Institute, United Kingdom

Theoretical and Applied Linguistics, Faculty of Modern and Medieval Languages, University of Cambridge, United Kingdom

**Simon Hengchen**

COMHIS, Department of Digital Humanities, University of Helsinki, Finland

**Viivi Lähteenoja**

Doctoral Programme for Philosophy, Arts, and Society, University of Helsinki, Finland

**Marco Palma**

Department of Statistics, University of Warwick, United Kingdom

**Alessandro Vatri**

Wolfson College, University of Oxford, United Kingdom

## 1. Introduction: lexical polysemy and semantic change

Language is a complex and dynamic system. If we consider word meaning, which is the scope of lexical semantics, we observe that some words have several meanings, thus displaying *lexical polysemy*. For example, *head* in English means the body part and a group leader. This situation may change over time, giving rise to *semantic change*, whereby a word acquires new meanings or loses existing ones. For example, *tweet* refers to the chirp of small birds, but in recent years it has been used also to refer to a post on Twitter (cf. e.g. entry ‘tweet, v.’ in the Oxford English Dictionary). Sometimes two meanings coexist before one takes over, as in the case of the French word *témoïn* ‘person giving testimony’, which originated from the Old French *tesmoin* ‘testimony/person giving testimony’, and derived from the Classical Latin word *testimonium* ‘testimony’ (Koch 2016: 25).

Semantic change phenomena are grounded in universal cognitive processes combined with cultural, linguistic, and historical factors (cf. Koptjevskaja-Tamm 2016). Its mechanisms have been studied by linguistics fields for some time, starting with diachronic linguistics in the 19th century and continuing with semiotics and cognitive semantics in the 20th century. Beyond linguistics research, understanding the way in which words change their meaning benefits a range of humanistic fields. As a consequence of semantic change, words used to refer to certain concepts can be different over time. The English adjective *gay* had the original meaning of ‘happy’ and ‘joyful’, but over the past century it has increasingly been used to refer to homosexual men (see the entry ‘gay, adj., adv., and n.’ in the Oxford English Dictionary). This has implications for historical studies of conceptual change that rely on textual sources. More generally, scholars searching on historical texts benefit from analyses of words’ meaning in their historical context, as this allows them to identify documents relevant to the topics of interest.

From a quantitative and computational point of view, semantic change can be successfully studied using probabilistic models, following the general paradigm put forward by Jensen and McGillivray (2017) and relative to linguistic change in general:

The spread of new linguistic behaviours among speakers through genres, linguistic environments, and social contexts, is a time-dependent phenomenon. The innovative form and the new form will coexist for a period of time [...], and there will be a more or less rapid adoption of the new form by the language communities. This can be described as a shift in probabilities, and it is clear that language spread should be dealt with in probabilistic terms. (Jensen and McGillivray (2017:44))

Here we explore the challenges of quantitatively modelling semantic change and polysemy based on the model by Perrone *et al.* (2019), and we discuss preliminary results of a statistical model

for semantic change in Ancient Greek. The focus on a historical language with a long textual tradition allows us to address important methodological challenges. First, in contrast with previous work in natural language processing, which has almost exclusively concentrated on relatively recent time periods (typically the last three centuries) and on earlier stages of extant languages (mostly English), the corpus at our disposal has a limited size and a specific composition, which can be traced back to the nature of the available texts. Consequently, the corpus is not balanced by genre, author, style, and register, and this needs to be accounted for in the computational analysis and interpretation of the results. Secondly, the presence of gaps and the lack of balance in the corpus, paired with the historical and cultural circumstances which characterize the development of Ancient Greek, condition the types of semantic change that can be observed based on the texts at our disposal, especially considering the literary nature of the corpus.

On the other hand, a long and rich scholarly tradition in Ancient Greek has provided us with valuable insights into the mechanisms of semantic change (cf. e.g. Leiwo 2012). These qualitative analyses can be leveraged to support and complement the computational modelling, and we will show an example of this in the present article.

One notable point concerns the interplay of polysemy and semantic change. We will pay particular attention to the crucial role played by genre in Ancient Greek lexical polysemy and semantic change. Let us take the example of the word *mus* which can mean ‘mussel’, ‘muscle’, ‘whale’, and ‘mouse’. Imagine that we have a late medical text which displays the use of *mus* as ‘muscle’. Can we say that *mus* is used with this meaning in that text because at that time the medical meaning had just emerged in the language, or rather that the ‘muscle’ meaning has always been used in medical contexts? If the only medical text we have is from a late era, we cannot be sure of our claims. On the other hand, if we had earlier medical texts, we could make some inferences, but the amount of available data would affect the reliability of our estimates.

The research questions we aim to answer are: can we quantify the role played by genre and time in polysemy? How can we computationally model polysemy in large-scale textual collections?

## 2. Previous Research

Semantic change has been the focus of different research fields, from historical linguistics to classics, history, and more recently, natural language processing. As written words are among the few traces of the past that researchers have access to, it is vital to most disciplines to correctly understand those fragments of the past from the point of view of their meaning. In this section, we will briefly review semantic change research in the humanities (including digital humanities), and in computational sciences.

In historical research, the study of conceptual change is mainly associated with Koselleck’s *Begriffsgeschichte*, the history of concepts. As Foxlee (2015) remarks, ‘the German word *Begriff* means both “concept” and “term”’. Consequently, concepts are mapped to linguistic units and the study of the latter gives insights into the study of the former. Koselleck divides the field into two tasks: synchronic and diachronic. This point is well illustrated in his study of the Ancient Greek term “crisis”, which he examines synchronically and diachronically (Koselleck 2006).

Indubitably, this approach can be seen as belonging to lexical semantics, where methods such as concordance analysis are used to look at the context in which words of particular interest appear. According to the distributional hypothesis (Harris, 1954), a word’s meaning can be defined by the context (i.e. the other words) with which it occurs. Distributional methods have proved to be very powerful, and corpora with temporal metadata can be harnessed to analyse the evolution of a particular word’s meaning diachronically. Ad-hoc methods, focussing on the specificity of a corpus, have also been devised to detect semantic change candidates: the focus shifts from the study of a particular (set of) word(s) to a corpus-driven approach where the texts define which words are of interest. To this end, other paradigms and methods have been put forward, for example in Hengchen (2017), including Trier (1931)’s semantic field theory and Blei *et al.* (2003)’s Latent Dirichlet Allocation.

Scholarship in classics also benefits from a better understanding of a word's meaning synchronically and diachronically. Traditionally, questions of lexical semantics have been addressed by examining all occurrences of a word in a corpus and interpreting them in context. Studies of this kind include, for instance, van Groningen's (1963) study of the noun *ekdosis* ('edition'), Allan's (1980) study of the verb *anagignôskô* ('to recognize/read'), Vatri's (2016) study of the meaning of *harmonia* in Aristotle, Pollitt's (1974) survey of the Greek and Latin terminology of ancient art criticism, and the studies of cultural meanings of individual lexical items in Christidis (2007:1116-53). Prime examples of the interest for computational methods on Ancient Greek texts are Boschetti's (2010) use of vector space distance to express semantic relations between words in the Thesaurus Linguae Graecae (TLG) corpus, and Rodda *et al.* (2017), presenting a data-driven method to explore semantic change between the pre-Christian and Christian eras, again using vector space models and the TLG corpus.

With the distributional hypothesis as a basis, computational linguists have devised algorithms that predict a word's primary sense given a list of possible senses (Word Sense Disambiguation). Tackling polysemy diachronically is a more recent advancement, cf. Kutuzov *et al.* (2018) and Tahmasebi *et al.* (2018). Hamilton *et al.* (2016) and Dubossarsky *et al.* (2015) propose "universal laws" of meaning change, which are later disproved (Dubossarsky *et al.*, 2017). Nonetheless, these studies usually focus on the change of the primary sense of a word, and do not capture the inter-sense dynamics of polysemy across time. To the best of our knowledge, aside from the model used in this paper, the only advanced technique that tackles such dynamics is SCAN (Fremann and Lapata, 2016) and its extension known as GASC (Perrone *et al.* 2019).

### 3. Corpus

To conduct this study we used the Diorisis Ancient Greek Corpus (Vatri and McGillivray, 2018). This corpus consists of 820 automatically lemmatized texts spanning between the beginnings of the Ancient Greek literary tradition (Homer) and the fifth century AD, and it counts approximately 10 million tokens. Each text is provided with metadata containing its approximate or exact (when known) date of composition and its literary genre and sub-genre. Table 1 displays the corpus' genres and sub-genres and Table 2 shows the distribution of words by century and genre. Due to the nature of the available Ancient Greek texts, the corpus has gaps and size unbalances regarding the chronological eras covered and the genres represented.

Genres	Subgenres
Poetry	Bucolic
	Didactic
	Epic
	Epigrams
	Erotic
	Choral
Comedy	Comedy
Tragedy	Tragedy
Philosophy	Philosophy
Essays	Essay
	e.g. works of Plutarch and Lucian
	Miscellanea
	The <i>Varia Historia</i> of Aelian and Athenaeus' <i>Deipnosophists</i>
Letters	Letters
Narrative	Biography
	Novel
	Mythology
	Church history
	History
Oratory	Oratory
Religion	Homily
	Hymns
	Pagan hymns
	Narrative
	Septuaginta and New Testament
	Protrepitics
	Christian protrepitics
	Psalms
	Theology
	Christian theology
Technical	Art history
	Geography
	Grammar
	Horsemanship
	Hunting
	Mathematics
	Medicine
	Military
	Natural history
	Politics
	Rhetoric, poetics, criticism
	Science
	Aristotelian treatises

Table 1: Categorization of genres and sub-genres in the Diorisis Ancient Greek corpus.

	8B C	7B C	5B C	4B C	3BC	2B C	1B C	1A D	2AD	3A D	4A D	5A D	Total
Comedy			79K	16K									95K
Essays				4K				475 K	263 K	361 K	24 K		1,127 K
Letters				10K	1K					10K	164		185K

											K		
Narrative			335K	209K	311K		661K	962K	483K	411K	98K		<b>3,470K</b>
Oratory			58K	529K				185K	296K	3K	56K		<b>1,127K</b>
Philosophy				895K					113K	213K			<b>1,221K</b>
Poetry	199K	16K	21K		81K	3K	7K		23K	18K	60K	127K	<b>555K</b>
Religion		16K			132K	463K		134K	45K		18K		<b>808K</b>
Technical			104K	327K	15K		386K	24K	394K	158K	4K		<b>1,412K</b>
Tragedy			207K										<b>207K</b>
<b>Total</b>	<b>199K</b>	<b>32K</b>	<b>804K</b>	<b>1,990K</b>	<b>540K</b>	<b>466K</b>	<b>1,054K</b>	<b>1,780K</b>	<b>1,617K</b>	<b>1,174K</b>	<b>424K</b>	<b>127K</b>	<b>10,207K</b>

Table 2: Number of words by century and genre in the texts of the Diorisis Ancient Greek corpus.

#### 4. The role of genre and time in lexical polysemy

The Diorisis Ancient Greek corpus provided the evidence for our study on semantic change. In order to better assess the complexity of the phenomenon and produce a handmade benchmark for the computational model, the two classicists in the team annotated a selection of sentences from the corpus (Vatri *et al.* 2019). We focussed on three frequent Ancient Greek polysemous words, which exhibited a range of clearly distinct senses, displayed an abstract sense and a concrete counterpart in the general vocabulary, and were attested in most of the time periods covered by the corpus and across different literary genres: *mus* ‘mouse/mussel/muscle/whale’, *harmonia* ‘fastening/agreement (harmony)/stringing (musical scale, melody)’, and *kosmos* ‘order/decoration/world’.

The range of meanings of these words was determined based on the standard scholarly Ancient Greek-English dictionary (Liddell *et al.*, 1996) and a philological study (Pollitt, 1974).

For the first two target words, all sentences from the corpus were annotated, whereas only the sentences from texts dated before 142 AD were annotated for *kosmos*, given the large amount of data in question. The annotation assigned the meaning of each context of use of these three words, and the type of information the annotators used to determine the meaning in context. Options for the latter task were:

- a. ‘collocates’: the meaning can be determined simply on the basis of the words occurring with the target word in the same sentence. E.g. Hippocrates, *Mochlicus* 23: *koinon touto arthroisi neuroisi musin enteroisin husterêisin alloisin* ‘this is common in joints, nerves, muscles, intestines, uteri, and other parts’. In this sentence the inflectional form *musin* (the dative plural of *mus*) occurs with other names of body parts, and has clearly the meaning of ‘muscle’. The computational model adopted in this study (see 4.2 below) looks precisely at collocates not exceeding sentence boundaries.
- b. ‘background knowledge’: the meaning cannot be established on the basis of the collocates occurring in the same sentence but can be determined based on knowledge of the topic and literary genre of the text or text section to which each target word belongs (types of information that could be extracted automatically or encoded as metadata in a new version of the model) or knowledge of the ancient Greek culture, society, history, literature, customs, and religion. E.g. Herodotus, *Histories* 4.192.3: *muôn de genea triksa autothi esti* ‘also there are in this place three kinds of mice’. In this sentence the form *muôn* refers to the meaning of ‘mouse’ because it occurs in a longer paragraph where animals living in Libya are described.

- c. ‘world knowledge’: the meaning can be determined based on context and knowledge of the world (but not specifically of the ancient Greek world) that can be reasonably assumed among today’s readers, with reference to the concepts of ‘common ground’ and ‘common knowledge’ by Clark and Brennan (1991). E.g. Aristophanes, *Wasps* 204: *isôs anôthen mus enebale soi pothen* ‘maybe a rat loosened it.’ In this sentence the word *mus* refers to a mouse that loosened a brick, because we can assume that, of all meanings of *mus*, this is the only one compatible with an animate being who can loosen a brick.
- d. ‘logic’: the meaning is selected because other meanings would be absurd, implausible, or illogical. E.g. Theocritus, 22.48–50: *en de mues stereois brakhiosin akron hup’ômon estasan êute petroi oloitrokhoi, ouste kulindeôn kheimarrous potamos megalais perixese dinais* ‘on the brawny arms, under the tip of the shoulders, muscles stood like rounded stones that a rushing river whirled and smoothed with its large eddies.’ In this sentence, the word *mus* could in principle mean ‘mouse’, but such an interpretation was deemed implausible by the annotators.
- e. ‘genre or register’: the meaning is determined by the conventions of or the topics covered by the literary genre or register of the text. Cf. Hippocrates, *Mochlicus* 4: *hoi gar mues houtoi toioutoi* ‘because the muscles are such [that they produce this effect]’. In this sentence the form *mues* (the nominative plural of the lemma *mus*) is overwhelmingly more likely to refer to the medical sense of ‘muscle’ as this is a medical text.
- f. ‘corpus processing error’: the sentence does not contain the target lemma and has been extracted due to corpus processing errors.

The reason for performing this second level of annotation, which to our knowledge has no precedents in the scholarly literature, is to provide an evidence base for comparing the computational models of semantic change with the manual analysis, so to better understand their potential and limits. As illustrated in section 2, such computational models (including the one described here) rely on the distributional hypothesis (captured by the category ‘collocates’ in the annotation).

Target word	Collocates	Backgro und knowled ge	World knowle dge	Logic	Genre or register	Corpus processing error	Annotated sentences
<i>mus</i>	114 (56%)	49 (24%)	21 (10%)	13 (6%)	1 (0%)	7 (3%)	205
<i>harmonia</i>	411 (69%)	176 (29%)	0 (0%)	7 (1%)	5 (1%)	0 (0%)	599
<i>kosmos</i>	1,102 (78%)	196 (14%)	7 (0%)	97 (7%)	3 (0%)	5 (0%)	1,411

Table 3: Distribution of the disambiguation categories in the annotated set.

Table 3 shows the distribution of the different categories recorded in the annotation for the three words. Although the distinction between categories b-e is sometimes debatable, it is easier to distinguish them from category a (‘collocates’), and non context-based disambiguation accounts for a considerable proportion (between 22% and 71%) of all annotated instances. We can conclude that any computational model relying on context information to infer word semantics will only be able to achieve the task at a less than perfect level of accuracy.

The process of manual annotation of the instances in the corpus was resource-intensive due to the amount of data and the complex nature of the semantics of the words. For example, in the case of *kosmos*, some instances were straightforward and contained references to women wearing *kosmos*, indicating that the relevant sense was ‘decoration’. Others were more complicated, especially cases where the senses identified were closely related, as when the context indicated moral connotations like a violation of *kosmos*, indicating a sense of either ‘order’, ‘propriety’, or ‘honour’.

#### 4.1 The role of genre in semantic change

The manual annotation of the senses of *mus*, *harmonia*, and *kosmos* allowed us to assess to what extent we should expect the time distribution of word senses for each lemma to follow the time distribution of that lemma across genres, and thus to gauge the weight of genre in lexical polysemy and semantic change in our Ancient Greek corpus.

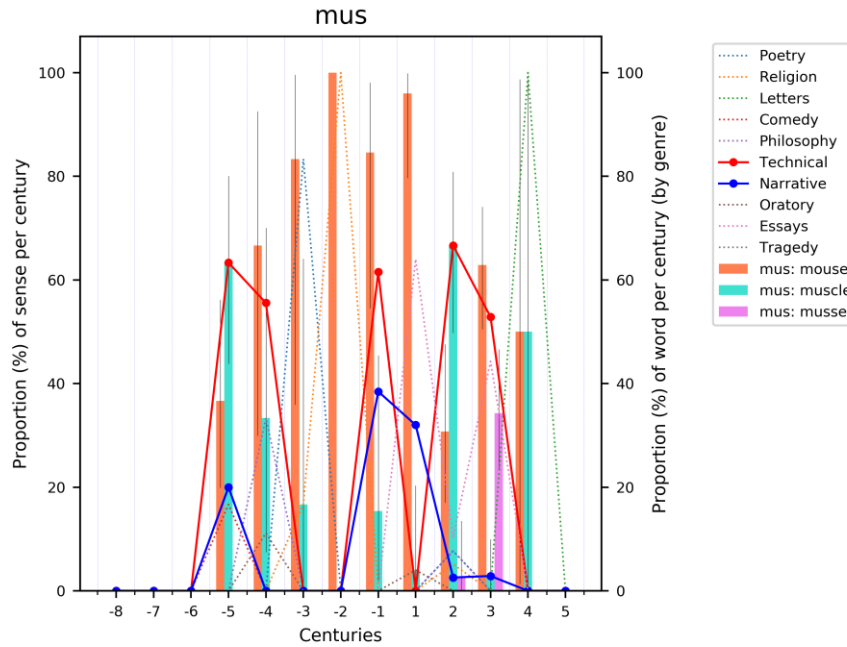


Fig. 1: Distribution of the occurrences of the lemma *mus* ‘mouse’/‘muscle’/‘mussel’ by genre over time and distribution of its senses over time. The coloured lines track the proportions of the occurrences of *mus* in each genre in each century, and the bars show the proportions of occurrences of *mus* with each sense in each century. The grey lines inside the bars show the confidence intervals. The data come from our manual annotation of the Diorisis Ancient Greek corpus.

The effect of genre was first estimated visually from graphs such as that in Fig. 1. If any line (representing the percentage of the occurrences of the target word in a literary genre across centuries) showed a similar trend to that of any set of bars (representing the percentage of the occurrences of a specific sense of the target word across centuries), we would be alerted that genre might play a more decisive role than diachrony in determining variation in the distribution of senses. In Fig. 1, for instance, a rough correspondence can be observed between the trend of the line representing the occurrences of *mus* in technical literature (red line) and that of the bars representing the occurrence of the sense ‘muscle’ (light blue bars). This suggests that ‘muscle’ would be associated with a technical genre such as medicine.

In order to measure the effect of genre, we calculated the Spearman’s Rank Correlation Coefficient ( $\rho$ ) for each word-sense  $s$  between the frequency  $f(s)$  of  $s$  across centuries and the frequency  $f(s,g)$  of  $s$  in each genre  $g$  across centuries. Significant correlation between  $f(s)$  and any  $f(s,g)$  would suggest that variation in the frequency of a word sense across centuries is not due to diachronic change, but to how frequently  $s$  is attested in  $g$  in each century (and, ultimately, to the amount of texts representing  $g$  in each century). Table 4 reports the results that reach statistical significance ( $p < 0.05$ ).

<i>harmonia</i>	‘agreement, harmony’	Technical ( $\rho = 0.888, p < 0.0001$ ) Narrative ( $\rho = 0.719, p = 0.006$ ) Essays ( $\rho = 0.561, p = 0.046$ )
-----------------	----------------------	---

	‘fastening’	Narrative ( $\rho = 0.663$ , $p = 0.013$ )
	‘stringing, musical scale’	Technical ( $\rho = 0.817$ , $p = 0.001$ ) Philosophy ( $\rho = 0.632$ , $p = 0.02$ ) Essays ( $\rho = 0.598$ , $p = 0.031$ )
<i>kosmos</i>	‘decoration’	Narrative ( $\rho = 0.887$ , $p = 0.001$ ) Technical ( $\rho = 0.705$ , $p = 0.023$ ) Oratory ( $\rho = 0.664$ , $p = 0.036$ )
	‘order’	Technical ( $\rho = 0.875$ , $p = 0.001$ ) Narrative ( $\rho = 0.862$ , $p = 0.001$ )
	‘world’	Technical ( $\rho = 0.792$ , $p = 0.006$ ) Oratory ( $\rho = 0.723$ , $p = 0.018$ )
<i>mus</i>	‘mouse’	Narrative ( $\rho = 0.813$ , $p = 0.001$ ) Essays ( $\rho = 0.743$ , $p = 0.004$ )
	‘muscle’	Technical ( $\rho = 0.766$ , $p = 0.002$ )
	‘mussel’	Narrative ( $\rho = 0.736$ , $p = 0.004$ ) Essays ( $\rho = 0.736$ , $p = 0.004$ ) Poetry ( $\rho = 0.613$ , $p = 0.026$ )

Table 4: Results of the correlation analysis between the distribution of the frequency of each of the three target lemmas (*harmonia*, *kosmos*, and *mus*) over time and the distribution of the frequency of each target lemma over time in each genre. The correlation analysis was carried out with Spearman’s Correlation test, and the values of the Spearman’s correlation coefficient  $\rho$  are shown. Only statistically significant ( $p < 0.05$ ) results are displayed.

This analysis indicates that each sense is significantly correlated with at least one genre, which supports the idea that literary Classical Greek is conservative when it comes to lexical semantics, and genre can often be said to play a larger role than time in polysemy.

The association of more than one meaning with one genre may be due to the coalescence of distinct sub-genres within the same high-level genre in the corpus. However, the size of the sub-corpora representing sub-genres is too small to conduct meaningful quantitative analyses and, as a consequence, we have opted for taking into account only coarse-grained high-level genre distinctions.

The manual annotation allowed us to perform a further test to identify the cases in which we observe semantic innovation independent from genre. If a sense  $s$  of a word  $w$  is first attested at time  $t$  in a genre  $g$ , and the word  $w$  is attested in genre  $g$  before time  $t$ , then the sense  $s$  can be called an innovation, since its appearance in our data at time  $t$  would not be due to the fact that  $w$  was not attested in  $g$  before  $t$ . In other words, if the sense  $s$  of a word  $w$  only appears in our data after a certain point in time  $t$ , this might conceal the fact that  $s$  existed in genres that are not represented in our data before  $t$  and would not allow us to conclude that  $s$  is a new sense of  $w$ . If  $w$  is attested with different senses in the same genre as the ‘new’ sense  $s$  before  $t$ , then we will be able to conclude that  $w$  acquired a new sense independently of the genre.

The results of this test are displayed below in Table 5.

<b>Word:sense</b> ( $w$ : $s$ )	<b>Time</b> ( $t$ )	<b>Genre</b> ( $g$ )	<b>Genres in which</b> $w$ is attested before $t$	<b>Genre-</b> <b>independent</b> <b>innovation</b>
<i>harmonia</i> : ‘agreement’	8BC	Poetry		No
<i>harmonia</i> :	8BC	Poetry		No



<b>‘fastening’</b>				
<i>harmonia</i> : <b>‘stringing’</b>	5BC	Poetry, Comedy	Poetry	Yes
<i>mus</i> : <b>‘mouse’</b>	5BC	Comedy, Narrative		No
<i>mus</i> : <b>‘muscle’</b>	5BC	Technical		No
<i>mus</i> : <b>‘mussel’</b>	2AD	Poetry	Comedy, Narrative, Poetry, Technical, Oratory, Essays, Philosophy, Religion	Yes
<i>kosmos</i> : <b>‘decoration’</b>	8BC	Poetry		No
<i>kosmos</i> : <b>‘order’</b>	8BC	Poetry		No
<i>kosmos</i> : <b>‘world’</b>	4BC	Philosophy, Technical, Oratory	Comedy, Narrative, Poetry, Technical, Oratory, Tragedy, Religion	Yes

Table 5: Results of the test performed to determine when semantic innovation independent from genre occurred in our annotated dataset. The first column displays the lemma of the target word and its sense *s*, the second column displays the first century in which that lemma is attested with that sense in our annotated dataset; the third column contains the genre of the text of first attestation; the fourth column contains the list of genres in which the lemma is attested in sense *s* before the century recorded in the second column, and the last column determines the cases in which we can talk about genre-independent semantic innovation.

Certain senses are attested at the beginning of our documentation (8<sup>th</sup> century BC); in some cases, however, it is possible to observe that senses arise in genres in which the target words were already attested, which suggests that some innovations occurred. Such innovations, however, are mostly located between pre-Classical (Homeric) Greek and Classical Greek. This also reminds us of the importance of considering philological information in the interpretation phase. The preliminary manual analysis of the corpus data suggests that the sense ‘mussel’ of *mus* is an innovation to be dated to the 2<sup>nd</sup> century AD. In the corpus, this sense is first attested in poetry, a genre in which *mus* appears before (*mus* appears in Oppian’s *Haliutica*, a didactic poem on fishing). Our data might not contain texts on topics in which the sense ‘mussel’ is likely to be attested, but which would have existed before that date (as a matter of fact, this sense is attested in fragments dating as far back as the 4<sup>th</sup> century BC, but which are not included in the corpus). Moreover, we are aware that this reasoning is only based on the texts of our corpus, and does not always generalize to Ancient Greek as a whole, considering the sparseness of our coverage of ancient Greek literature.

#### 4.2. A computational model for semantic change

The analyses illustrated so far show the complexities of semantic change phenomena in Ancient Greek. In the rest of this article, in order to answer our original research questions on large-scale analysis of polysemy, we discuss the challenges involved in creating the Genre-Aware Semantic Change (GASC; Perrone *et al.* 2019), a computational model aimed to allow scholars to scale up their manual analyses and identify general patterns in the semantic evolution of the Ancient Greek lexicon. Perrone *et al.* (2019) applied the SCAN model by Frermann and Lapata (2016), which extends a Bayesian topic model to the case of semantic change in English in the 16<sup>th</sup>-20<sup>th</sup> centuries. This enabled the analysis of the model’s estimated probability distribution of a word’s meaning over time

and raised interesting questions regarding the challenges of this computational research on Ancient Greek. Secondly, Perrone *et al.* (2019) combined the insight on the role played by genre in semantic change phenomena in historical languages within a new computational model. Originally, GASC infers word meanings not only from the context of words, but also from the genre of the texts, making it possible to estimate the probability distribution of a word's meaning over time *and* by genre. In the next section we will discuss the results of SCAN and GASC.

## 5. Results

We applied the genre-independent model of Frermann and Lapata (2016), SCAN, to the Ancient Greek corpus on the three target lemmas *harmonia*, *kosmos*, and *mus* for which we had manually-annotated data to compare. Given the properties of the corpus, we defined the time intervals to correspond to centuries, as this gave a sufficient amount of data. Since SCAN is a topic model, it requires as input a parameter K specifying the number of topics (in this case the number of macro-senses) of the words. We experimented with K=4 and K=12 and observed that the former value gave more interpretable results, so we will report on those here. The value K=3 could also have been considered, but due to the amount of cases for which no expert sense assignment is reported we decided to exclude it.

SCAN generates two forms of output:

- a) K lists of 10 lemmas, each describing a sense that is identified as distinct; the lemmas are listed in decreasing order of probability, so the first words are more strongly associated with the sense.
- b) A distribution of senses over time, which can be visualized as a bar chart like Figure 2.

We will describe the results for *kosmos*, which has the highest frequency among the three words we annotated and offers an opportunity to discuss challenges and insights from the process.

The four lists for the model-assigned senses of *kosmos* are shown at the bottom of Figure 2. Interpreting the sense of *kosmos* that the lists described was challenging, as it is commonly the case with topic labelling. First, the lists contain common verbs (e.g. becoming/being, having, and speaking). Second, the most telling words such as 'city' or 'citizenship', implying a sense of 'government', and 'gold(en)' or 'horse', implying a sense of decoration, were often found in the same lists and at times in multiple lists. In Figure 2 the words 'golden/*chruseos*' and 'gold/*chrusos*' appear in two different lists (first and second) and 'city/*polis*' occurs together with 'gold' in the second list (and in the fourth list). It is however possible to interpret the lists in light of the senses that we know *kosmos* has. The first sense (green) can be interpreted as 'order' thanks to the presence of words like *tithēmi* 'to set' and *kosmeō* 'to order', which are both associated with placing something in order, as well as *sōma* 'body', with its association with being a kind of ordered whole. The second sense (dark orange) is more difficult to interpret because it is associated to generic words (*poieō* 'to make', *anthrōpos* 'human', *chōreō* 'to give way', and *polus* 'many'). The third sense (blue) can be interpreted as 'world' thanks to words such as *theos* 'god', *khōra* 'space', *mesos* 'middle', *gē* 'earth', and *ouranos* 'sky'. The fourth sense (pink) can be interpreted as 'government' thanks to *polis* 'city', a unit with its own government, and *naus* 'ship', often associated with the military or merchant navy of a city-state (admittedly, *theos* also appears here, but with a lower probability ranking than in the third list).

Concerning the time distribution of senses for *kosmos*, because this word has not undergone dramatic semantic changes over time, the bar chart in Figure 2 reflects the distribution of senses in the texts making up the corpus (which is not balanced), more than diachronic trends, and it is important to keep this in mind in the interpretation.

Fig. 2: Output of the SCAN model for *kosmos*. The bar chart shows the distribution of senses over the time periods covered by the corpus; senses are defined by the lists of words at the bottom. *Kosmos* happens to be a collocate of itself, hence its presence in the green sense.

Given the importance of genre in lexical polysemy of Ancient Greek words, we also decided to apply the GASC model to the words in the corpus. GASC provides the following output:

- K (4) lists of 10 lemmas, each describing a sense that is identified as distinct; the lemmas are listed in decreasing order of probability;
- For each genre, a distribution of senses over time, which can be visualized as a bar chart, such as that in Figure 2.

Considering each individual genre in the corpus would mean that the model would work with very sparse data, which can lead to unreliable statistical estimates. Therefore, we grouped senses into two categories. For *kosmos*, we focussed on the Narrative genre and contrasted it to the other genres. Figure 3 visualizes the output of the GASC for *kosmos*.

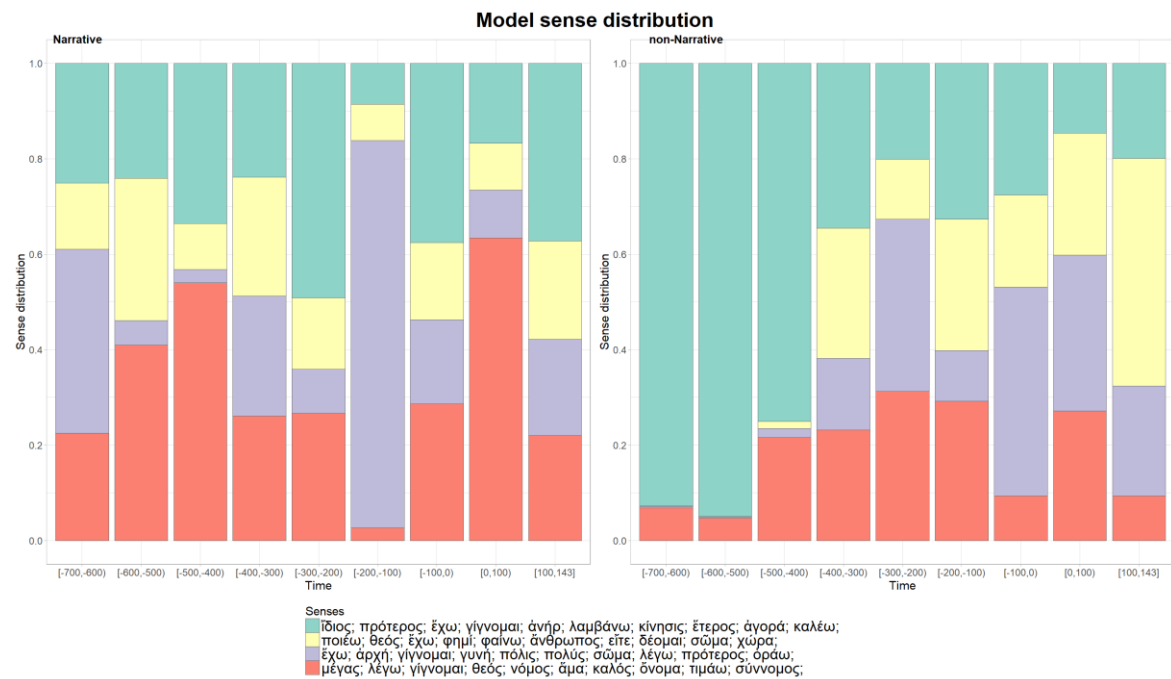


Fig. 3: Output of the GASC model for the *kosmos*. For each macro-genre (Narrative on the left and Non-Narrative on the right, which includes all genres other than Narrative), the bar charts show the time distribution of senses, which are defined by the lists of words at the bottom.

Assigning sense labels to the word lists generated by the model was again challenging because many of the words are generic enough to be applicable to most of the senses of *kosmos*. However, if we focus on non-generic words and those words that are distinctive of one sense over the others, we can interpret the senses as follows. The first sense (light blue) can be interpreted as ‘order’, thanks to the presence of words like *idios* ‘private’, *anēr* ‘man’, *agora* ‘marketplace’, and *kaleō* ‘call’. The second sense (yellow) can be interpreted as ‘world’ (cf. *theos* ‘god’, *phainō* ‘appear’, and *khōra* ‘space’). The third sense (violet) can be interpreted as ‘decoration’ (cf. *gunē* ‘woman’, *soma* ‘body’). Finally, the fourth sense (red) can be interpreted again as ‘order’ (cf. *nomos* ‘law’, *ama* ‘together’, *timaō* ‘to honour’, and *sunnomos* ‘lawful/regular’) and its distribution over time may be summed to that of the first sense.

Comparing Figure 3 with Figure 4 (derived from the expert annotation), we see some similarities in trends, and expecting a perfect match would be unrealistic given the fact that the

models provide probability estimates. In both figures the sense for ‘world’ (yellow and pink bars, respectively) increases in the later time periods in Non-Narrative texts.

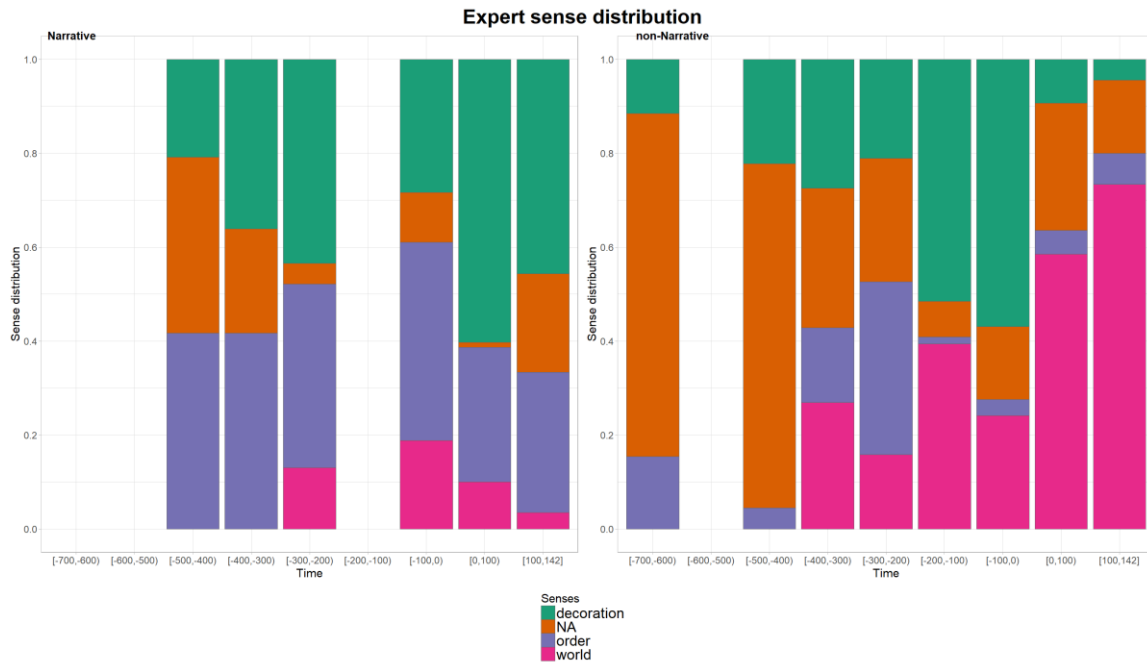


Fig. 4: Distribution of the expert-annotated senses of *kosmos* in Narrative (left) and all other texts (right). The label for ‘NA’ is for senses which were annotated based on criteria other than just context.

Finally, we tested SCAN on words for which diachronic semantic change can be presupposed based on existing philological research. We selected the words *paradeisos*, a Persian loan-word whose meaning expanded from ‘garden’ to the Jewish-Christian ‘paradise’ in the Greek translation of the Old Testament (the Septuaginta, to be dated to the 3<sup>rd</sup> and 2<sup>nd</sup> centuries BC) and in the New Testament (Kyrtatas 2007), and *parabolê*, whose original meaning ‘comparison’ expanded to ‘parable’. Figure 5 displays the results for *paradeisos*.

## SCAN sense distribution

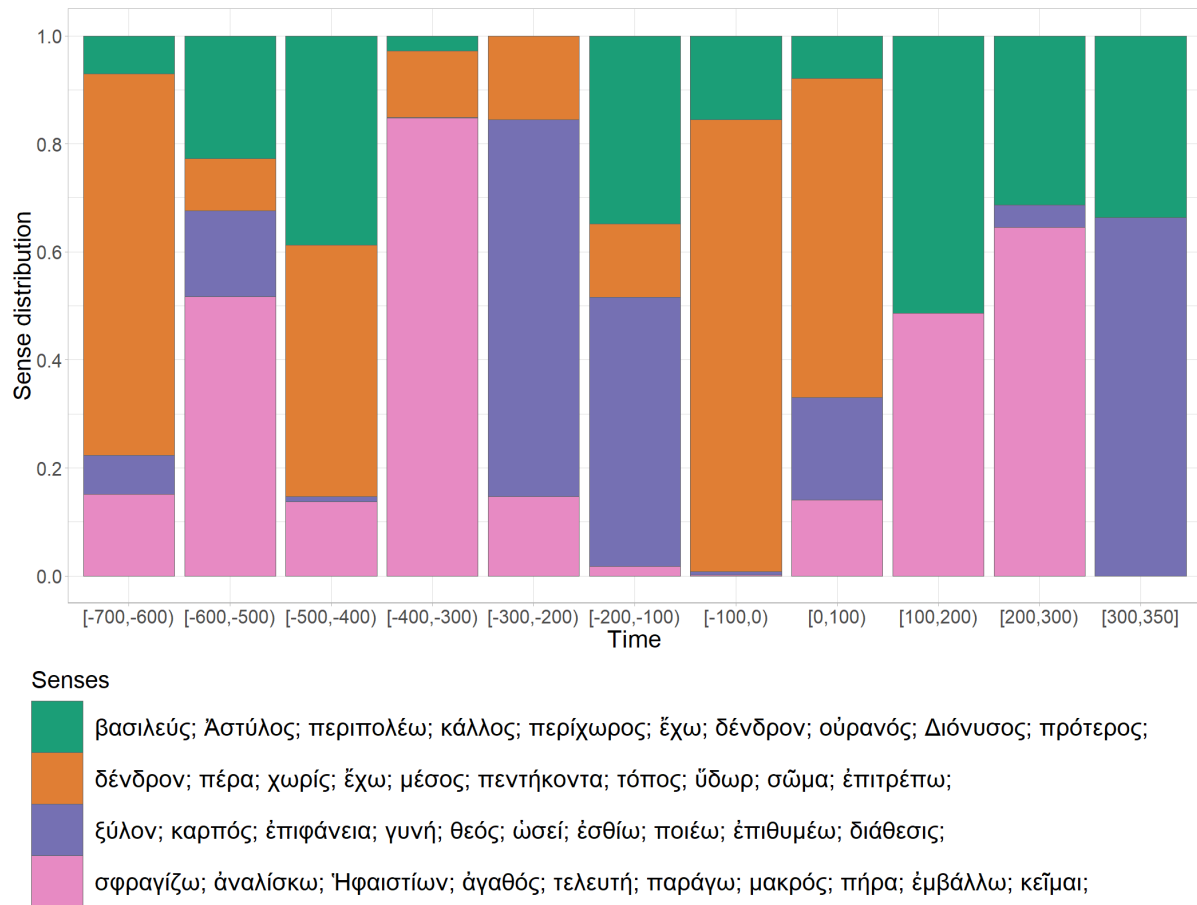
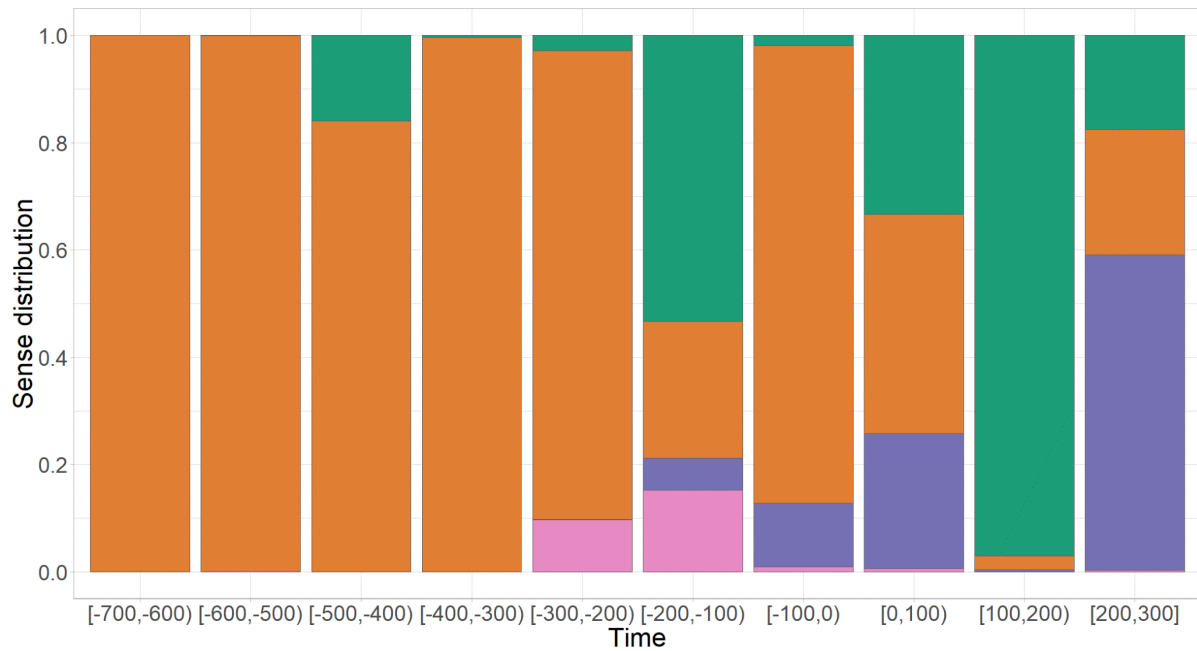


Fig. 5: Output of the SCAN for the word *paradeisos*.

Most words in the lists point to the original meaning ‘garden’ (e.g. *dendron* ‘tree’, *topos* ‘place’, *hudôr* ‘water’), but some in the third list point to the Biblical account of the garden of Eden and the fall of Adam and Eve (*karpos* ‘fruit’, *gunê* ‘woman’, *theos* ‘god’, *esthiô* ‘eat’). If the third list corresponds to the Jewish-Christian meaning of ‘paradise’, it is remarkable that its probability peaks right around the time to which the translation of the Old Testament is dated (3<sup>rd</sup> and 2<sup>nd</sup> centuries BC) and later on in the Christian era.

Figure 6 shows the results for *parabolê*. The lists contain a high number of verbs of saying (*legô*, *eipon* ‘say’, *laleô* ‘chat’, *apantaô* ‘answer’) and speech-related nouns (*logos* ‘speech’, *paroimia* ‘proverb’). Only the first list contains words pointing to religious discourse, namely *oneidismos* ‘reproach’ (which occurs in the letters of Paul) and *prophêtês* ‘prophet’. If this association is correct, it might explain the probability peaks for the first list in the Christian era, but not much so for that in the 2<sup>nd</sup> century BC.

## SCAN sense distribution



### Senses

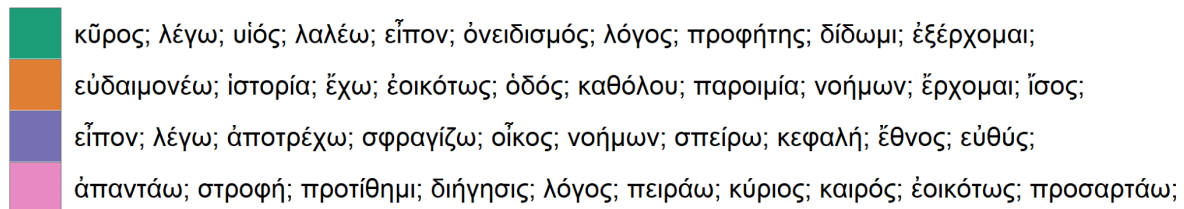


Fig. 6: Output of SCAN for the word *parabolê*.

## 6. Discussion and Conclusion

The ability to use computers to scale up manual analyses of word semantics over time is going to open up exciting opportunities in humanistic research. In this article we report on research that takes the first steps in this direction for Ancient Greek. This offers us the opportunity to reflect on the challenges and the future research directions in this area, as well as more generally on the value of computational modelling in Humanities research.

We know that computational modelling of polysemy and semantic change in natural languages is a difficult task, as proven by the fact that this is still an open problem in computational linguistics. For historical and Ancient languages specifically, the task is made even harder by a combination of factors, as shown by the analyses we reported on in this article.

When SCAN (Fremann and Lapata 2016), a statistical model of semantic change developed for relatively recent stages of a modern language (English), is applied to the Ancient Greek corpus, several adjustments had to be made. The features of the corpus meant that genre balance could not be assumed, leading us to experiment with the GASC model (Perrone *et al.* 2019) that considers the texts' genre characteristics and their effect on polysemy. As McGillivray (2014) noted for computational processing of Latin data, another question concerned data sparsity, although the Diorisis corpus is one of the largest annotated corpora of Ancient Greek available. This was addressed by a periodization of the corpus by century, so that the model was provided with enough data to draw its inferences from. Despite these adjustments, a close study of *kosmos* shows that data sparsity is still a challenge, as shown by the fact that the target words appear several times in their contextual list of words.

Further linguistic considerations are also in order. Semantic polysemy is a complex phenomenon, which encompasses cultural and historical events, literary and stylistic conventions, but also different linguistic patterns. For example, one meaning can emerge as an extension of another, leading to fuzzy boundaries. The example of *paradeisos* illustrates this, as the religious sense is closely related (also distributionally) to the original meaning of ‘garden’. This has important implications for computational implementations, which can partially capture even such subtle changes. On the other hand, as for *parabolê*, when the emerging meaning is not distinguishable from the original one based on the context of the word, but based on other factors, the task becomes harder for distributional systems.

Finally, from a methodological viewpoint, the question of how to evaluate computational models of semantics is still open. We have shown the challenges of comparing computational output with expert annotation, and of setting correct expectations regarding the acceptable level of misalignment between expert analysis and automatic analysis. In spite of these limits, we have shown the potential of such statistical models to allow researchers to quantify complex semantic phenomena, thus going beyond the traditional account of historical semantics in terms of examples of first attestations of meanings (as reported in historical dictionaries, for instance). Such quantitative accounts make it possible, among other things, to gain insights into the dynamics between different meanings of a word, and ultimately, to place this in its wider historical and causal context.

### Acknowledgements

This work was supported by The Alan Turing Institute [EPSRC grant EP/N510129/1, SF0042 to BMcG].

### Contributions

BMcG designed and oversaw the study, designed the genre analysis described and the evaluation approach, contributed to the design of the computational model, drafted sections 1, 4, 5, and 6, reviewed all other sections, and acquired the funding; SH drafted sections 1, 2 and 6, contributed to the design of the computational model, and reviewed all other sections. VL annotated the sample for *kosmos* and interpreted its results, and drafted section 5. MP contributed to the design and implementation of the computational model, drafted sections 4.2 and 5, and prepared figures 2, 3, 4, 5, and 6. AV collected and processed the corpus data, annotated the sample for *mus* and *harmonia* and interpreted their results, contributed to the design of the computational model, conducted the genre analysis, and drafted sections 3 and 5. All authors gave final approval for publication.

### References

**Allan, D.J.** (1980). Ἀναγγινώσκω and Some Cognate Words. *The Classical Quarterly* n. s. 30: pp. 244–51.

**Blei, D.M., Ng, A.Y. and Jordan, M.I.** (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan), pp. 993-1022.

**Boschetti, F.** (2010). A corpus-based approach to philological issues. Ph.D. thesis, University of Trento.

**Christidis, A.-P.** (ed.) (2007). *A history of Ancient Greek: from the beginnings to Late Antiquity*. Cambridge University Press, Cambridge.

**Clark, H. H. and Brennan, S. E.** (1991). Grounding in communication. In Resnick, L. B.; Levine, J. M., (eds.), Perspectives on socially shared cognition, American Psychological Association.

**Clarke, M.** (2010). Semantics and Vocabulary. In Bakker, E. J. (ed.). A Companion to the Ancient Greek Language, pp. 120–33. Wiley-Blackwell, London.

**Dubossarsky, H., Tsvetkov, Y., Dyer, C. and Grossman, E.** (2015). A bottom up approach to category mapping and meaning change. In NetWordS (pp. 66-70)

**Dubossarsky, H., Weinshall, D. and Grossman, E.** (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing pp. 1136-1145.

**Foxlee, N.** (2015). From Analogue to Digital: Conventional and Computational Approaches to Studying Conceptual Change. Retrieved online at [https://www.academia.edu/20046437/From\\_Analogue\\_to\\_Digital\\_Conventional\\_and\\_Computational\\_Approaches\\_to\\_Studying\\_Conceptual\\_Change\\_2015\\_](https://www.academia.edu/20046437/From_Analogue_to_Digital_Conventional_and_Computational_Approaches_to_Studying_Conceptual_Change_2015_) on September 12<sup>th</sup>, 2018.

**Frermann, L. and Lapata, M.** (2016). A Bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics, 4, pp.31-45

**Hamilton, W.L., Leskovec, J. and Jurafsky, D.** (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.0909

**Harris, Z.S.** (1954). Distributional structure. Word, 10(2-3), pp.146-162

**Hengchen, S.** (2017). When does it mean? Detecting semantic change in historical texts. Ph.D. thesis, Université libre de Bruxelles.

**Jenset, G. B. and McGillivray, B.** (2017). Quantitative Historical Linguistics. A corpus framework. Oxford University Press, Oxford.

**Koch, Peter** (2016). Meaning change and semantic shifts. In Juvonen, Päivi & Koptjevskaja-Tamm, Maria (eds). The Lexical Typology of Semantic Shifts. De Gruyter, Berlin.

**Koptjevskaja-Tamm, Maria** (2016). “The lexical typology of semantic shifts. An introduction”. In Juvonen, Päivi & Koptjevskaja-Tamm, Maria (eds). The Lexical Typology of Semantic Shifts. De Gruyter, Berlin.

**Koselleck, R.** (2006). “Crisis”. Trans. Michaela W. Richter. Journal of the History of Ideas, 67(2): 357-400.

**Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E.** (2018). Diachronic word embeddings and semantic shifts: a survey. Arxiv: <http://arxiv.org/abs/1806.03537>

**Kyrtatas, D. J.** (2007). Παράδεισος. In Christidis, A.-P. (ed.). A history of ancient Greek: from the beginnings to Late Antiquity, 1137–40. Cambridge University Press, Cambridge.



**Leiwo, M.** (2012). Introduction: Variation with Multiple Faces. In Leiwo, M., Halla-aho, H., and Vierros, M. (eds.). *Variation and change in Greek and Latin*, 1–11. Suomen Ateenan-instituutin säätiö, Helsinki.

**Liddell, H.G. and Scott, R.** (1996). *A Greek English Lexicon*. 9<sup>th</sup> edition with revised supplement. Oxford University Press, Oxford.

**McGillivray, B.** (2014). *Methods in Latin Computational Linguistics*. Brill, Leiden.

**Nerlich, B.** (1992). *Semantic Theories in Europe, 1830-1930: From etymology to contextuality*, volume 59. John Benjamins Publishing, Amsterdam.

**Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q., McGillivray, B.** (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. Arxiv: <https://arxiv.org/abs/1903.05587>

**Pollitt, J.J.** (1974). *Ancient View of Greek Art* (pp. 247-54). New Haven: Yale University Press.

**Reisig, K.** (1839). *Vorlesungen über lateinische Sprachwissenschaft*. Lehnhold, Leipzig.

**Rodda, M.A., Senaldi, M.S. and Lenci, A.** (2017). *Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek*. In CLiC-it/EVALITA.

**Setatos, M.** (2007). *Semantic Change*. In Christidis, A.-F. (ed.). *A History of Ancient Greek: from the Beginnings to Late Antiquity*, 667–76. Cambridge University Press, Cambridge.

**Schmid, H.** (1995a). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

**Schmid, H.** (1995b). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin.

**Tahmasebi, N., Borin, L., Jatowt, A.** (2018). *Survey of Computational Approaches to Lexical Semantic Change*. arXiv preprint arXiv:1811.06278.

**Trier, J.** (1931). *Der Deutsche Wortschatz im Sinnbezirk des Verstandes: von den Anfängen bis zum Beginn des 13. Jahrhunderts* (Vol. 31). Winter.

**van Groningen, B.A.** (1963). Ἐκδοσις. *Mnemosyne* 16: 1–17.

**Vatri, A.** (2016). Between Song and Prose: The Meaning(s) of *Harmonia* in Aristotle's *Rhetoric* and *Poetics*. *Rhetorica* 34: 372–92.

**Vatri, A. and McGillivray, B.** (2018). The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*. [https://brill.com/view/journals/rdj/aop/article-10.](https://brill.com/view/journals/rdj/aop/article-10.1163-24523666-01000013.xml)

1163-24523666-01000013.xml

**Vatri, A., Lähteenoja, V. and McGillivray, B.** (2019). *Ancient Greek Semantic Change - Annotated Datasets and Code*. figshare. doi:10.6084/m9.figshare.c.4445420.